WHEN DOES COVARIANCE MINIMIZE THE EXPECTED MAXIMUM EDIT RESPONSE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Covariance-based knowledge editors choose an update direction by minimizing its mean squared response on unrelated inputs. A small mean square can hide a large response among the probes that evaluate the edit. We ask when the same direction also minimizes the expected largest absolute response over a finite probe budget. In a fixed feasible subspace, covariance placement is optimal for every nonzero target and every number of iid probes exactly when the symmetrized whitened activation law is spherical. Centered elliptical laws with finite covariance meet this condition, including laws with polynomial tails. Bounded laws with the same covariance can need different directions at different budgets. A dimension-free converse turns measured directional anisotropy into a lower bound on the worst-target risk ratio. The remaining gain depends on the target. In 32-dimensional target-conditioned slices from five architectures, the median held-out risk ratio of covariance placement to a fitted alternative at 128 probes is 2.20 at searched worst targets and 1.048 at real edit targets. On ROME's edits in two architectures, budget-fitted placement replaces the quadratic objective. It has lower geometric-mean held-out risk than ROME and a matched quadratic refit at every evaluated budget at or above its fitting budget, with rewrite accuracy unchanged. For response magnitude, fitting an empirical Weibull shape parameter raises the held-out coverage of the fitted sub-Weibull predictor from 90.1% to 97.9% at 5,000 probes.

1 INTRODUCTION

Covariance-based knowledge editing chooses the key direction of an update by minimizing its mean squared response on unrelated inputs. A small mean square can hide a large response among the probes that evaluate the edit. We study the expected largest absolute response over M probes and ask when minimizing it selects the mean-square direction and when the probe budget changes that choice.

ROME (Meng et al., 2022) places a rank-one update with the activation second moment, MEMIT (Meng et al., 2023) extends second-moment fitting to batches and layers, and AlphaEdit (Fang et al., 2025) builds a preservation subspace from the same statistic. The question applies to all three. We minimize the expected maximum at one layer over directions that meet a prescribed target gain in a fixed feasible subspace.

Whitening makes the quadratic objective Euclidean, so covariance placement becomes the shortest feasible direction. Two directions with equal quadratic cost can have different projection tails and different expected maxima (Figure 1). Covariance placement is optimal at every target and every iid budget exactly when the symmetrized whitened activation law is spherical. Centered elliptical laws with finite covariance qualify, including polynomial-tailed ones. Bounded laws with equal covariance can need different directions at different budgets.

Real activations are anisotropic. The gain over covariance placement depends on the edited target and on the subspace the edit may use. On ROME's edits, budget-fitted placement lowers held-out risk below ROME at every evaluated budget at or above its fitting budget.

- **An exact condition and a quantitative converse.** In a fixed feasible subspace, covariance placement is optimal at every target and iid budget exactly when the symmetrized whitened

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

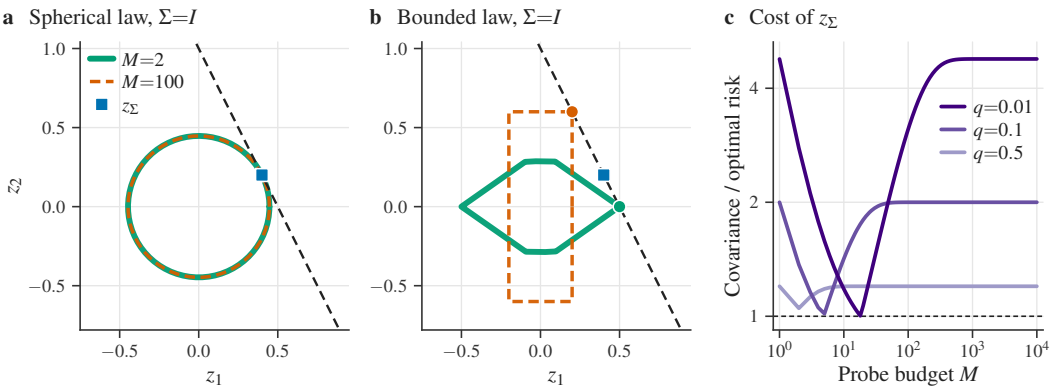

Figure 1: Equal covariance can yield different budget-optimal directions. In **a** and **b**, level sets of the expected maximum R_M pass through the constrained optimum; the black dashed line is the gain constraint $2z_1 + z_2 = 1$. **a** The covariance direction z_Σ is optimal at every budget. **b** The bounded law puts mass q=0.1 on $\pm e_1/\sqrt{q}$ and $1-q$ on $\pm e_2/\sqrt{1-q}$; its optimum (dots) moves with the budget. **c** z_Σ risk over the optimum, by q.

law is spherical. Measured anisotropy forces a dimension-free floor on the worst-target ratio.

- **A target-dependent gap on real activations.** In 32-dimensional slices over 329 facts in five architectures, the median held-out risk ratio of covariance placement to a fitted alternative at 128 probes is 2.20 at searched worst targets and 1.048 at real edit targets.
- **Budget-fitted placement for ROME.** On 150 requests per setting in two architectures, budget-fitted placement (LP) has lower geometric-mean held-out risk than ROME and a matched quadratic refit at every evaluated budget at or above its fitting budget. Rewrite accuracy is unchanged.
- **Sizing the response.** Fitting an empirical Weibull shape parameter raises the held-out coverage of the fitted sub-Weibull predictor. At matched coverage, an Orlicz scale gives tighter median thresholds than root-mean-square scaling.

2 Problem Setup

Fix a layer-ℓ MLP down-projection $W_\ell \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, a candidate update ΔW_ℓ, and a unit readout $r \in \mathbb{R}^{d_{\text{out}}}$. For an unrelated activation $k \sim \mathcal{D}_\ell$ and an update of any rank, define

$$s_\ell(k) := r^\top \Delta W_\ell k = w^\top k, \qquad w := \Delta W_\ell^\top r.$$

We compare directions by the expected maximum absolute response $\mathbb{E}\max_{i \le M} |s_\ell(k_i)|$ over a pre-specified budget of M probes from $\mathcal{D}_\ell$; for a fixed edit, probabilities and expectations are over the probes. For a rank-one update $\Delta W_\ell = \alpha u v^\top$ with unit u and v, $s_\ell(k) = \alpha(r^\top u)(v^\top k)$ separates magnitude, readout alignment and key direction; placement comparisons fix the target gain. In Section 5, r is the unit-normalized unembedding row of the new object's first token. We analyze one layer update at a time.

Activation second moments. Placement methods weight directions by the quadratic form of the activation second-moment matrix $\Sigma_\psi := \mathbb{E}[kk^\top]$, estimated on 10^5 English Wikipedia activations (Wikimedia Foundation, 2022); it equals the covariance for centered activations (conventions in Section E).

3 When Covariance Placement is Optimal

ROME's placement (Meng et al., 2022) is a linearly constrained minimum-variance program (Capon, 1969); adding preservation constraints (Frost, 1972) gives

$$\min_{v \in \mathcal{R}} v^\top \Sigma_\psi v \quad \text{subject to} \quad v^\top k_\star = \gamma, \qquad K_{\text{pres}}^\top v = 0,$$

for target key $k_\star$, preservation keys K_{pres}, retained subspace $\mathcal{R}$ and feasible gain γ, with a closed-form solution (Section C).

We now minimize the expected maximum response. In the feasible subspace $\mathcal{S} = \mathcal{R} \cap \ker K_{\text{pres}}^\top$ with orthonormal basis B, population covariance $\Sigma = \Sigma_\psi$ of centered probes, $H = B^\top \Sigma B \succ 0$ and $L = BH^{-1/2}$, put $X = L^\top k$, $a = L^\top k_\star$ and $v = Lz$. Then $v^\top \Sigma v = \|z\|_2^2$ and $v^\top k_\star = a^\top z$, so the unit-gain covariance direction is $z_\Sigma = a/\|a\|_2^2$, and the criterion is $R_M(z) = \mathbb{E} \max_{i \le M} |z^\top X_i|$ over iid copies of X.

Theorem 1 (Covariance sufficiency across targets and budgets)**.** *In the preceding population-whitened space, z_Σ minimizes R_M subject to $a^\top z = 1$ for every nonzero target a and every integer $M \ge 1$ if and only if εX is spherically distributed, where ε is an independent uniform sign. If X is centrally symmetric, the condition is sphericity of X itself.*

Centered elliptical laws with finite covariance satisfy this condition, including laws with polynomial tails. Bounded laws with identity covariance can require budget-dependent placement (Table 23, Figure 1). The proof has two steps. At a fixed budget, optimality on every target hyperplane forces R_M to be constant on the unit sphere (Precupanu, 2013; Koldobsky, 1993; Blanco & Turnšek, 2006). Equality of these directional risks at every budget then forces all unit absolute projections to share one law (Hoeffding, 1953; Molchanov & Turin, 2021), which gives sphericity after sign symmetrization (Section J). The condition plays the role that elliptical symmetry plays for mean–variance sufficiency (Chamberlain, 1983; Benveniste et al., 2024), with the expected maximum (Sun & Yao, 2018) as the risk.

3.1 How far covariance placement can fall short

Outside the spherical case, two ratios measure what a departure costs: one compares unit directions, the other compares covariance placement with the best gain-feasible direction at the worst target. For fixed M let the directional variation be $C_M = \max_{\|z\|_2=1} R_M(z) / \min_{\|z\|_2=1} R_M(z)$, and let $\kappa_M = \sup_{a \ne 0} \kappa_M(a)$ be the worst-target risk ratio of the covariance direction, where $\kappa_M(a) = R_M(z_\Sigma) / \min_{a^\top z = 1} R_M(z)$ at target a.

Theorem 2 (Quantitative converse)**.** *Let R_M be continuous, even, positively homogeneous and positive off the origin; every norm and every empirical R_M on probes that span the space qualifies. Write $\lambda = \log \kappa_M$ and $\Phi(\lambda) = \inf_{N \in \mathbb{N},\, N \ge 2} N\big(\lambda - \log \cos \frac{\pi}{2N}\big)$. Then*

$$\kappa_M \le C_M \le \exp \Phi(\lambda) \le \exp\big(\pi\sqrt{\lambda} + 2\lambda\big), \qquad \Phi(\lambda) = \tfrac{\pi}{\sqrt{2}}\sqrt{\lambda}\,(1 + o(1)) \ \text{ as } \lambda \to 0.$$

Read in reverse, a measured C_M forces $\kappa_M \ge \exp[\Phi^{-1}(\log C_M)]$, so any anisotropy that a search exhibits on a fitting bank certifies a floor on that bank's worst-target ratio. The bound is dimension free, applies to the polyhedral risks that finite probe banks induce, and gives a floor above one for every $C_M > 1$. The constant-factor bound $C_M \le 12\kappa_M + 3$ for general norms (Khurana, 2024, arXiv v4, Lem. 2.6 and Eq. (2.1)) forces nothing below $C_M = 15$, above every value measured here. Sections J.3 and J.4 give the proof, its order-sharp square-root dependence, and the budget-dependent counterexamples.

4 Covariance placement on real activations

We study directional geometry in 17 architecture–layer cells of five architectures (2.8B–14B) (Wang & Komatsuzaki, 2021; Biderman et al., 2023; Grattafiori et al., 2024; Yang et al., 2025; Abdin et al., 2024), with 40 CounterFact targets per cell (329 distinct facts) (Meng et al., 2022). The probe bank has one token from each of 20,000 articles, split so the halves share no article (Table 9).

For M draws with replacement from a finite fitting bank, the empirical R_M is an ordered-weighted-ℓ_1 norm, so the minimum in $\kappa_M(a)$ is a linear program. Feasible subspaces (slices) are 32-dimensional, spanned by the target and the leading principal directions (target-conditioned slices), and whitened by the fit-half second moment. The fitting-bank minimizer $\hat{z}_{\text{fit}}$ is compared with covariance placement on the evaluation bank through $H_M(a) = R_M^{\text{eval}}(z_\Sigma) / R_M^{\text{eval}}(\hat{z}_{\text{fit}})$; values above one favor the fitted alternative. The worst target found is the top-ranked target of a search over the fitting bank; each candidate's linear program is solved to a certified gap. The null-normalized D_M

Table 1: Fitted alternatives gain more at the worst targets found than at real edit targets. Excess is $100(D_M - 1)$. Probes and target are projected off the preservation keys unless a row names another projection. A probe-row target is a fit-half probe; the ambient direction is random in activation space.

Measure / target / projection	M=32	M=128	M=512
Raw real-target median risk ratios, real / null			
held out	**1.011** / 0.994	**1.048** / 0.978	**1.098** / 0.932
in sample	1.016 / 1.004	1.067 / 1.018	1.161 / 1.073
Null-normalized excess (%), held out / in sample			
Worst target found	+51.4 / +46.8	+137.4 / +113.3	+252.1 / +175.5
Random target	+8.8 / +8.8	+21.9 / +20.2	+36.9 / +26.9
Probe-row target	+3.2 / +2.2	+13.4 / +8.2	+27.9 / +8.6
Real target, no projection	+3.1 / +2.4	+10.4 / +7.8	+22.7 / +13.2
100 random directions	+3.0 / +2.3	+10.3 / +7.7	+22.2 / +13.0
Mean preservation key	+2.5 / +1.9	+10.1 / +7.5	+22.9 / +12.7
Preservation keys	+1.7 / +1.2	+7.2 / +4.8	+17.8 / +8.2
100 principal directions	+0.7 / +0.3	+3.4 / +1.3	+10.8 / +3.4
Ambient direction	+0.5 / +0.2	+1.5 / −0.3	+4.5 / −2.7
Fit-half geometry, real / null			
directional variation $\widehat{C}_M$	2.28 / 1.16	3.94 / 1.40	6.39 / 1.86
worst-target floor from $\widehat{C}_M$	1.140 / 1.004	1.404 / 1.023	1.787 / 1.079

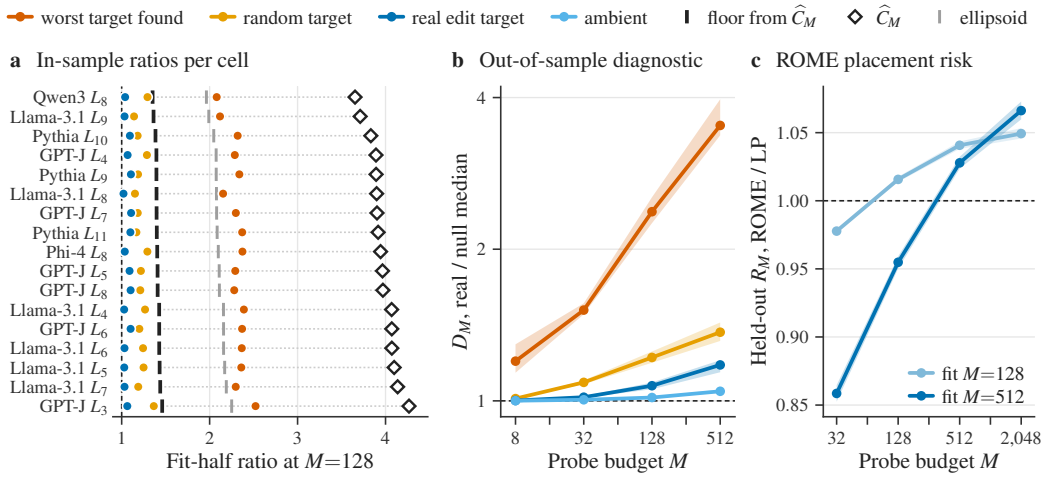

Figure 2: Directional placement gaps and budget-fitted ROME edits. In **a** and **b**, projection and target classes are as in Table 1. **a** Median fitting-bank risk ratio $\kappa_M(a)$ per target class in each architecture–layer cell at M=128, with $\widehat{C}_M$ (diamonds), the worst-target floor that Theorem 2 guarantees from it, and the worst-target ratio of an inner-product norm with the same directional variation (ellipsoid). **b** Held-out D_M with 95% cell-bootstrap bands. **c** Geometric-mean held-out directional risk ratio ROME / LP at unit target gain over 150 GPT-J L_5 requests, with 95% request-bootstrap bands; values above one favor LP.

divides the pooled median of H_M over a target class by the same median under a spherical reference that keeps a real slice's radii and draws uniform directions (Section H).

Searched gaps exceed the certified floor and persist on held-out probes. At M=128 the median directional variation $\widehat{C}_M$ on the fit half is 3.94 (spherical reference: 1.40), so Theorem 2 certifies a median fitting-bank worst-target floor of 1.404. The target search reaches 2.32 (Figure 2(a)), and each of the 680 slices clears its own floor (Figure 12). On held-out probes the searched targets keep a median ratio of 2.20, and the searched target exceeds its slice's floor in all 680 slices at every budget; in the spherical reference, no slice does so at $M \geq 32$ (Section H). Per target, the fit-half ratio carries over with Spearman correlation 0.93 at real and 0.94 at random targets (Figure 14).

Table 2: Budget-fitted placement (LP) lowers held-out risk at every evaluated budget at or above its fitting budget. Entries are geometric means of per-request ROME / LP risk ratios; values above one favor LP. The first block compares directions at unit averaged-key gain, the second includes the realized output multiplier before the precision cast, and the third scores the matrices written to the model. Behavior is LP minus ROME in percentage points.

Cell and fit budget	M=32	128	512	2,048
Unit-gain directional risk				
GPT-J L_5, fit 128	0.978	1.016	1.041	1.049
GPT-J L_5, fit 512	0.859	0.955	1.028	1.066
Llama-3.1 L_5, fit 128	0.983	1.011	1.033	1.043
Factor reconstruction: layer-output norm risk				
GPT-J L_5, fit 128	0.976	1.014	1.039	1.048
GPT-J L_5, fit 512	0.857	0.953	1.026	1.064
Llama-3.1 L_5, fit 128	0.984	1.012	1.034	1.043
Matrices written to the model				
GPT-J L_5, fit 128	0.977	1.014	1.039	1.048
GPT-J L_5, fit 512	0.857	0.953	1.026	1.065
Llama-3.1 L_5, fit 128	0.984	1.012	1.033	1.041

Cell and fit budget	Rewrite	Paraphrase	Neighborhood
GPT-J L_5, fit 128	0.00	−0.33	+0.47
GPT-J L_5, fit 512	0.00	0.00	+0.73
Llama-3.1 L_5, fit 128	0.00	+0.33	+0.20

Fitted gains transfer when the fitting bank spans many articles. On a contiguous cache of 10,000 tokens from about twenty articles, fitted gains survive a split by token and fall below the spherical reference once the halves are split at an article boundary (Figure 3(c), Table 15). Directions fitted on a few articles describe those articles, so every result in this section fits on 10,000 articles and scores on 10,000 others.

The gap depends on the target and on the feasible subspace. At M=128 the median held-out ratio is 2.20 at searched targets and 1.048 at real edit targets. In each of the 17 cells, the median rises from real to random to searched targets (Figure 13). Real and searched targets are nearly orthogonal in these whitened slices. Their median $|\cos|$ is 0.014 (spherical reference: 0.127) and stays lower at every width (Figure 3(d)). Wider slices find larger searched gaps (Figure 3(a,b)). Searched and real-target gaps grow with the budget, to 3.06 and 1.098 at M=512 (Figure 11). Slices spanned by principal directions without the target raise the real-target median in 16 of the 17 cells at M=128, by a geometric-mean factor of 1.109 over cells (Table 18).

The preservation constraint is one such subspace choice. At M=128, projecting probes and target off the preservation keys lowers the null-normalized excess at real targets from 10.4% to 7.2%. Projecting off 100 random directions leaves it at 10.3% (Table 1).

Budget-fitted placement lowers the risk of ROME's edits. Budget-fitted placement (LP) minimizes the fit-half R_M at a fitting budget $M_{\text{fit}} \in \{128, 512\}$ over an affine family that moves ROME's direction within a 32-dimensional slice. Probes are uncentered, as in ROME's second moment. On 150 requests per setting, LP replaces ROME's direction inside ROME's edit (Figure 2(c)). At every evaluated $M \geq M_{\text{fit}}$, all three geometric-mean scorings in Table 2 favor LP in every setting, with ratios up to 1.066 at M=2,048. The article-resampling interval for the written matrices lies above one at each such budget (Table 22). Rewrite accuracy is unchanged in every setting.

The two objectives select different directions. A quadratic refit over LP's fitting probes, affine family and gain constraint isolates the objective (Table 3). It improves on ROME by 0.4 to 1.2%. Its expected-maximum risk exceeds LP's at every evaluated $M \geq M_{\text{fit}}$, by 2.9% and 3.8% at M=512 and 2,048 for the GPT-J 128 fit and by 2.3% and 3.2% on Llama-3.1 L_5. No paired interval covers one. Like the optimum of Figure 1(b), the LP direction is specific to its fitting budget (Section H).

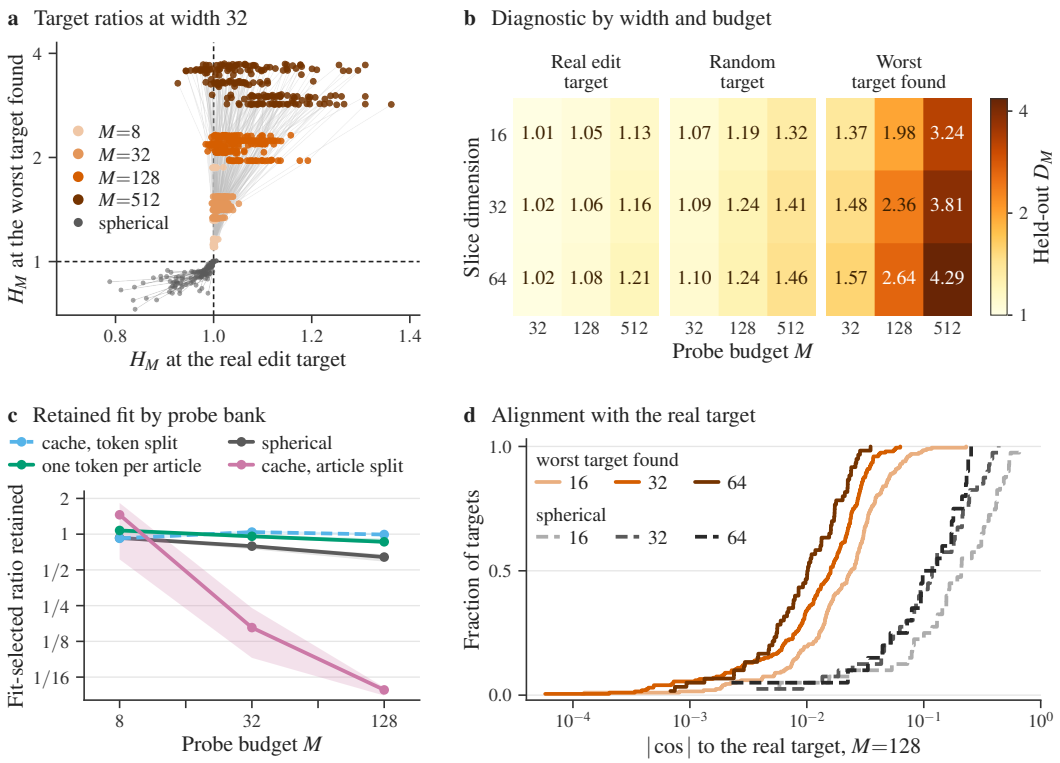

Figure 3: Searched gaps grow with slice width; searched targets are nearly orthogonal to the real edit target. Panels **a**, **b** and **d** use the five cells measured at every slice width. **a** One point per target and budget, joined across budgets. **b** Held-out D_M. **c** Held-out over fit-half value of the ratio between the two extreme directions on GPT-J L_5 and Llama-3.1 L_6: median with 10th–90th percentile band (interquartile for the spherical reference). **d** Cumulative distribution of $|\cos|$ between the worst target found and the real edit target, by slice width, on real slices (solid) and in the spherical reference (dashed).

Table 3: At every evaluated $M \geq M_{\text{fit}}$, LP beats the matched quadratic refit (quad), and every such interval excludes one. Blocks not marked Llama are GPT-J L_5. Entries are geometric-mean held-out risk ratios with 95% request-bootstrap intervals, from runs separate from those of Table 2; LP below quad is the share of requests on which LP has the lower risk.

Fit budget	Eval. M	ROME / LP	ROME / quad	quad / LP [95%]	LP below quad
M_{fit}=128	32	0.9781	1.0039	0.9743 [0.9726, 0.9760]	0.0%
	128	1.0159	1.0084	1.0075 [1.0059, 1.0090]	76.7%
	512	1.0411	1.0116	1.0292 [1.0272, 1.0313]	98.0%
	2,048	1.0495	1.0115	1.0375 [1.0344, 1.0405]	98.0%
M_{fit}=512	32	0.8582	1.0040	0.8547 [0.8485, 0.8608]	0.0%
	128	0.9548	1.0086	0.9467 [0.9419, 0.9512]	1.3%
	512	1.0276	1.0118	1.0157 [1.0116, 1.0198]	76.0%
	2,048	1.0663	1.0118	1.0538 [1.0480, 1.0595]	92.7%
Llama, M_{fit}=128	32	0.9849	1.0054	0.9797 [0.9783, 0.9811]	0.0%
	128	1.0131	1.0088	1.0042 [1.0029, 1.0056]	67.3%
	512	1.0345	1.0113	1.0230 [1.0205, 1.0255]	97.3%
	2,048	1.0433	1.0113	1.0316 [1.0281, 1.0355]	98.0%

5 Predicting the size of the response

Theorem 1 ties the optimal direction to directional symmetry. Along a fixed direction, the tail of the response sets how fast its maximum grows with the probe budget. We predict that magnitude for fixed edits, $s_\ell(k) = w^\top k$, on archives spanning five architectures, fifteen editor configurations and

six factual-editing datasets (Table 9). The fitted sub-Weibull predictor has the form of a standard Orlicz maximum bound and is scored on held-out probes.

Response scales. For a real random variable Y, we call Y *sub-Weibull-θ* when the threshold-2 Orlicz functional is finite:

$$\|Y\|_{\psi_\theta} := \inf\left\{c > 0 : \mathbb{E}\, e^{(|Y|/c)^\theta} \le 2\right\}, \tag{1}$$

following Kuchibhotla & Chakrabortty (2022). Writing $K = \|Y\|_{\psi_\theta} > 0$ gives $\Pr(|Y| > t) \le 2\exp(-(t/K)^\theta)$ for $t > 0$. Here θ=2 is sub-Gaussian (Wainwright, 2019) and θ=1 sub-exponential (Vershynin, 2018).

Probe assumptions. The maximal inequality below uses three conditions. **(A1) Fixed edit and evaluation law:** the edit is fixed independently of the evaluation probes, whose marginals are $\mathcal{D}_\ell$. **(A2) Population projection-tail control:** along the fixed key direction, $\|v^\top k\|_{\psi_\theta} \le K < \infty$ for a population bound K. **(A3) Fixed probe protocol:** collection, token filtering and centering are fixed and define the probe population.

Theorem 3 (Sub-Weibull-θ tail bound)**.** *For a rank-one update $\Delta W_\ell = \alpha u v^\top$ under (A1)–(A3), let $\theta > 0$, $M \ge 1$ and $\delta \in (0,1)$; the probes $k_1, \dots, k_M$ may be dependent. With probability at least $1-\delta$,*

$$\max_{i \le M} |s_\ell(k_i)| \;\le\; \big(\log(2M/\delta)\big)^{1/\theta}\, |\alpha|\, |r^\top u|\, K.$$

The proof (Section I) combines the classical Markov–Orlicz inequality (Kuchibhotla & Chakrabortty, 2022; Vladimirova et al., 2020) with a union bound. It covers an update of any rank with $K_w = \|w^\top k\|_{\psi_\theta}$ in place of $|\alpha||r^\top u|K$. With $\delta = 2/M$ the threshold is $(2\log M)^{1/\theta} K_w$, the form of the fitted predictor below. A two-point law on the Orlicz ball, chosen for each $M \ge 2$, attains the same M-dependence with probability at least $1/2$ (Theorem 6). At that matched confidence, the upper and lower thresholds have ratio $(1 + \log 4/\log M)^{1/\theta} \to 1$, so the rate is sharp on the ball (Section G.1).

5.1 From a population bound to a fitted predictor

For the responses $x_i = |w^\top k_i|$ on fitting probes, we fit an empirical Weibull shape parameter $\widehat{\theta}$ to the top 5% and solve $n^{-1}\sum_i \exp\{(x_i/\widehat{K}_w)^{\widehat{\theta}}\} = 2$ for the empirical scale, giving the fitted predictor $\widehat{B}_M = \widehat{K}_w(2\log M)^{1/\widehat{\theta}}$. On laws with known tail index, $\widehat{\theta}$ orders the laws by tail heaviness (Section F.2). Table 4 lists the inputs that the population bound, the fitted predictor and a conformally calibrated threshold $\rho\widehat{B}_M$ each need.

Fitting the shape parameter raises held-out coverage at every budget, from 90.1% to 97.9% at M=5,000 (Table 5). Across architectures, the second-moment reference is exceeded 14.7 to 72.4 times as often as the fitted predictor (Figure 4).

Table 4: Three thresholds for the same local response, with the inputs each needs and what it delivers. The inflation factor ρ is set from calibration scores (Section D.3).

Object	Required input	What it delivers
Population bound B_{pop}	Valid population (K_w, θ); fixed edit and probe law	At fixed budget M, coverage over fresh probes is at least $1-\delta$ for $K_w[\log(2M/\delta)]^{1/\theta}$.
Fitted predictor $\widehat{B}_M$	Empirical $(\widehat{K}_w, \widehat{\theta})$ fitted on probes	Predicts the local maximum; coverage and threshold size are measured on held-out probes.
Calibrated threshold $\rho\widehat{B}_M$	Exchangeable calibration and new-edit scores; fixed M and complete selection policy	Covers a new edit with probability at least $1-\delta$, averaged over the calibration sample and the new edit's probes.

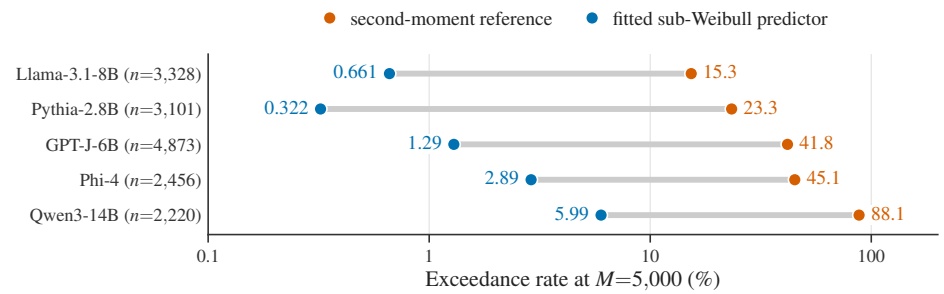

Figure 4: On every architecture, the second-moment reference is exceeded more often than the fitted sub-Weibull predictor. Rows are architectures (Table 9), with n edits each. The reference $2|\alpha||r^\top u|\sqrt{v^\top \Sigma_\psi v \log M}$ uses only the second-moment estimate and the update's top singular triple (α, u, v), and is scored on the first 5,000 probes. The fitted predictor $\widehat{B}_M$ is estimated on those probes and scored on the held-out 5,000. Table 5 scores both on the same held-out probes.

Table 5: Fitting the shape parameter raises coverage at every budget, most at the largest. Coverage in percent over 11,439 distinct nonzero response vectors from 119 cells. Row (2) is $\widehat{B}_M$, row (1) the same scale fit at θ=2, and row (0) the second-moment reference of Figure 4.

	evaluation-split coverage at probe budget M			
Predictor	500	1,000	2,000	5,000
(0) Second-moment reference	81.2	77.8	70.4	64.1
(1) Per-edit fit, θ=2	98.6	97.4	93.8	90.1
(2) Per-edit fit, $\hat{\theta}$	99.8	99.3	98.3	97.9

5.2 Tighter thresholds at matched coverage

Inflation raises any predictor's coverage, so we compare thresholds at equal coverage. For record i let $Z_i = \log(B_i/O_i)$, with B_i a predictor's threshold and O_i the observed held-out maximum. One global inflation per predictor and budget, chosen on the evaluation records, brings observed coverage to c; the matched-coverage log looseness $\mathrm{EFF}_q(c) = Q_q(Z) - Q_{1-c}(Z)$ is the q-quantile of Z after it. We compare seven coverage targets $c \in [0.90, 0.999]$ and four budgets, 28 operating points in all.

An Orlicz scale is tighter than RMS at the median at every operating point. The 41,208 records score $|v_1^\top k|$ on the one-token-per-article bank, where v_1 is the update's leading right singular vector (Table 11). At the median record, the fixed-θ=2 Orlicz scale is tighter than the per-record root mean square (RMS) of the fitting responses at all 28 operating points, by 1.16 to 1.53× (Figure 5, Table 12), and tighter than the variance proxy $v_1^\top \Sigma_\psi v_1$ at all 28 (Table 13). At M=5,000 and coverage 0.98 it is tighter than RMS by 1.362× (group-bootstrap 95% CI [1.265, 1.470]). At the 90th-percentile record, $\hat{\theta}$ has lower looseness than fixed $\theta = 2$ at all 28 points, by 1.06 to 1.25× (Table 12).

6 Related work

Liu et al. (2026d) study covariance-related instability in same-subject editing, where edited facts are not recalled under user instructions. The present work concerns placement for the expected maximum response on unrelated probes. Spectral, hyperspherical-energy and embedding-alignment constraints (Ma et al., 2025; Liu et al., 2026b; Dai et al., 2026) regularize the update, and behavioral locality (Hoelscher-Obermaier et al., 2023; Liu et al., 2026c) evaluates downstream model outputs. Here the response is measured at the edited layer. Our criterion applies the expected-maximum functional studied as MAXVAR (Sun & Yao, 2018) to absolute activation projections, and Theorem 2 relates to the stability theory of approximately orthogonality-preserving maps (Mojškerc & Turnšek, 2010; Zhang et al., 2014; Khurana, 2024).

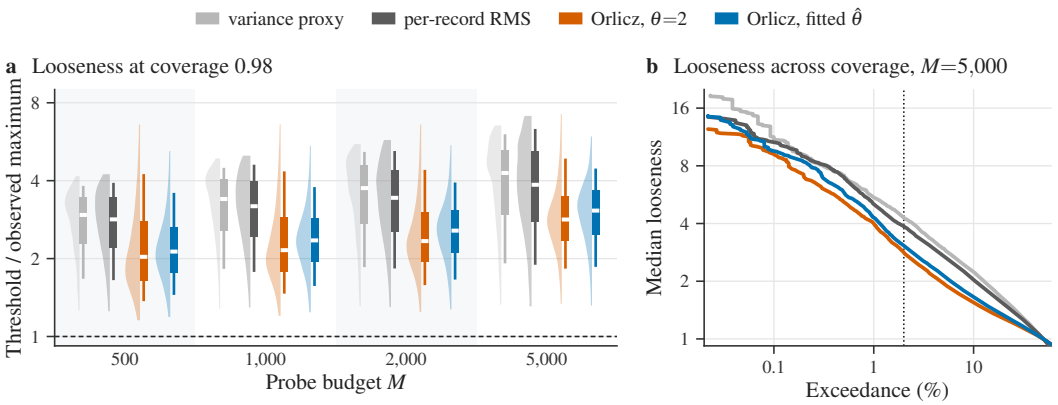

Figure 5: At coverage 0.98 and every budget, both Orlicz scales are tighter than RMS at the median. **a** Per-record looseness after one global inflation per predictor brings its coverage to 0.98. Boxes span the quartiles, whiskers the 10th to 90th percentiles, and the white rule is the median; half-violins show the full distribution. The variance-proxy series covers the 13,016 records that have it; Table 13 compares the scales on that subset. **b** Each curve traces one predictor's median looseness as its inflation varies; the dotted line is the coverage of **a**.

7 Conclusion

In a fixed feasible subspace, covariance placement minimizes the expected maximum response at every target and every iid budget exactly when the symmetrized whitened activation law is spherical. Centered elliptical laws with finite covariance qualify, including laws with polynomial tails. Without sphericity, measured anisotropy forces a dimension-free lower bound on the worst-target risk ratio. On real activations the bound certifies a median fitting-bank floor of 1.404 at 128 probes. The gap at real edit targets is small and depends on the feasible subspace. On ROME's edits, a budget-fitted direction lowers held-out expected maxima below ROME's direction and a matched quadratic refit at every evaluated budget at or above its fitting budget. The probe budget belongs in the placement objective whenever the expected maximum response is the quantity being minimized. Along a fixed direction, fitting an empirical Weibull shape parameter raises the held-out coverage of the fitted sub-Weibull predictor from 90.1% to 97.9% at 5,000 probes.

The use of large language models

A large language model assistant was used to edit prose, develop and revise analysis and figure-generation code, cross-check reported numbers, examine derivations, and assist with result interpretation. The authors take responsibility for the final manuscript, including its claims, proofs, analyses and references.

Reproducibility statement

The assumptions and proofs of the characterization and the quantitative converse are in Section J. Section H specifies the directional-risk optimization, target search, placement protocol and evaluation. Sections E and F document the response-scale analyses and cohort definitions, and Section D.5 the provenance audit. Table 10 maps each main-text figure and table to its source cohort and probe bank. The supplementary material provides the placement implementation and scripts that regenerate every figure, every main-text table and the principal appendix tables from the archived per-record results; it is also at `https://anonymous.4open.science/r/ICLR27-53362`.

References

Marah Abdin et al. Phi-4 technical report. *arXiv preprint*, 2024. arXiv:2412.08905.

Elie Benveniste, Petter N. Kolm, and Gordon Ritter. Untangling universality and dispelling myths in mean-variance optimization. SSRN working paper 4747461, 2024.

Stella Biderman et al. Pythia: A suite for analyzing large language models across training and scaling. *arXiv preprint*, 2023. arXiv:2304.01373.

Antonio Blanco and Aleksej Turnšek. On maps that preserve orthogonality in normed spaces. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, 136(4):709–716, 2006. doi: 10.1017/S0308210500004674.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Ding Cao, Yuchen Cai, Yuqing Huang, Xuesong He, Rongxi Guo, Guiquan Liu, and Guangzhong Sun. On the superimposed noise accumulation problem in sequential knowledge editing of large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(36):30139–30147, 2026. doi: 10.1609/aaai.v40i36.40263.

Jack Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418, 1969. doi: 10.1109/PROC.1969.7278.

Gary Chamberlain. A characterization of the distributions that imply mean–variance utility functions. *Journal of Economic Theory*, 29(1):185–201, 1983. doi: 10.1016/0022-0531(83)90129-1.

Qizhou Chen, Dakan Wang, Taolin Zhang, Zaoming Yan, Chengsong You, Chengyu Wang, and Xiaofeng He. UniEdit: A unified knowledge editing benchmark for large language models. In *NeurIPS Datasets and Benchmarks*, 2025. arXiv:2505.12345.

Qizhou Chen, Chengyu Wang, Taolin Zhang, and Xiaofeng He. An information-theoretic framework for robust large language model editing. *arXiv:2512.16227*, 2026.

Jacek Chmieliński and Rafał Stypka. Additive operators approximately preserving Birkhoff–James orthogonality. *Aequationes Mathematicae*, 99(6):2847–2854, 2025. doi: 10.1007/s00010-025-01210-4.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics (TACL)*, 2024.

Yanbo Dai, Zhenlan Ji, Zongjie Li, and Shuai Wang. EAMET: Robust massive model editing via embedding alignment optimization. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2505.11876.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, 2022. arXiv:2208.07339.

Bethan Evans and Jared Tanner. Theory of minimal weight perturbations in deep networks and its applications for low-rank activated backdoor attacks. *arXiv:2601.16880*, 2026.

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 *ICLR (Outstanding Paper)*, 2025. arXiv:2410.02355.

Otis Lamont Frost. An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, 60(8):926–935, 1972. doi: 10.1109/PROC.1972.8817.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint*, 2024. arXiv:2407.21783.

Bruce M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.

Wassily Hoeffding. On the distribution of the expected values of the order statistics. *The Annals of Mathematical Statistics*, 24(1):93–100, 1953. doi: 10.1214/aoms/1177729086.

Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. Detecting edit failures in large language models: An improved specificity benchmark. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11548–11559, 2023.

Houcheng Jiang, Junfeng Fang, Tianyu Zhang, Baolong Bi, An Zhang, Ruipeng Wang, Tao Liang, and Xiang Wang. Neuron-level sequential editing for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. arXiv:2410.04045; code: github.com/jianghoucheng/NSE.

Pranav Kumar Kaliaperumal. Activation outliers in transformer quantization: Reproduction, statistical analysis, and deployment tradeoffs. *arXiv:2603.04308*, 2026.

Divya Khurana. On approximately orthogonality preserving and reversing operators. *arXiv preprint arXiv:2409.12546*, 2024. doi: 10.48550/arXiv.2409.12546.

Alexander Koldobsky. Operators preserving orthogonality are isometries. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, 123(5):835–837, 1993. doi: 10.1017/S0308210500029528.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. BERT busters: Outlier dimensions that disrupt transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, 2021. arXiv:2105.06990.

Arun K. Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11(4):1389–1456, 2022. doi: 10.1093/imaiai/iaac012.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018. doi:10.1080/01621459.2017.1307116.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In *CoNLL*, 2017.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. PMET: Precise model editing in a transformer. In *AAAI Conference on Artificial Intelligence*, 2024. arXiv:2308.08742.

Bingqing Liu, Wei Liu, and Yuhua Li. BetaEdit: Null-space constrained sequential model editing. *arXiv:2605.09285*, 2026a.

Qingyuan Liu, Jia-Chen Gu, Yunzhi Yao, Hong Wang, and Nanyun Peng. Energy-regularized sequential model editing on hyperspheres (SPHERE). In *International Conference on Learning Representations (ICLR)*, 2026b. arXiv:2510.01172.

Wei Liu, Haomei Xu, Hongkai Liu, Zhiying Deng, Ruixuan Li, Heng Huang, Yee Whye Teh, and Wee Sun Lee. Are we evaluating the edit locality of LLM model editing properly? *arXiv:2601.17343*, 2026c.

Xiyu Liu, Qingyi Si, Zhengxiao Liu, Chenxu Yang, Naibin Gu, and Zheng Lin. Beyond the covariance trap: Unlocking generalization in same-subject knowledge editing for large language models (RoSE). *arXiv:2603.15518*, 2026d.

Sicheng Lyu, Yu Gu, Xinyu Wang, Jerry Huang, Sitao Luan, Yufei Cui, Xiao-Wen Chang, and Peng Lu. EvoEdit: Evolving null-space alignment for robust and efficient knowledge editing. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1520–1540, 2026. doi: 10.18653/v1/2026.findings-acl.75. arXiv:2510.13851.

Jun-Yu Ma, Hong Wang, Hao-Xiang Xu, Zhen-Hua Ling, and Jia-Chen Gu. Perturbation-restrained sequential model editing (PRUNE). In *ICLR*, 2025. arXiv:2405.16821.

Xin Ma, Wei Chen, Qi Liu, Derong Xu, Zhi Zheng, Tong Xu, and Enhong Chen. More edits, more stable: Understanding the lifelong normalization in sequential model editing. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, 2026.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *ICLR*, 2023.

Blaž Mojškerc and Aleksej Turnšek. Mappings approximately preserving orthogonality in normed spaces. *Nonlinear Analysis: Theory, Methods & Applications*, 73(12):3821–3831, 2010. doi: 10.1016/j.na.2010.08.007.

Ilya Molchanov and Riccardo Turin. Convex bodies generated by sublinear expectations of random vectors. *Advances in Applied Mathematics*, 131:102251, 2021. doi: 10.1016/j.aam.2021.102251.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, volume 2430 of *Lecture Notes in Computer Science*, pp. 345–356, 2002. doi:10.1007/3-540-36755-1_29.

Haewon Park, Sangwoo Kim, and Yohan Jo. SUIT: Knowledge editing with subspace-aware key-value mappings. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.24502.

Teodor Precupanu. Duality mapping and Birkhoff orthogonality. *Analele Ştiinţifice ale Universităţii Al. I. Cuza din Iaşi. Matematică*, 59(1):103–112, 2013. doi: 10.2478/v10157-012-0027-6.

Jie Sun and Qiang Yao. On coherency and other properties of MAXVAR. *Vietnam Journal of Mathematics*, 46:87–94, 2018. doi: 10.1007/s10013-017-0262-y.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *Conference on Language Modeling*, 2024. arXiv:2402.17762.

Lukas Thede, Karsten Roth, Matthias Bethge, Zeynep Akata, and Tom Hartvigsen. WikiBigEdit: Understanding the limits of lifelong knowledge editing in LLMs. In *ICML*, 2025. arXiv:2503.05683.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020.

Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier tailed distributions. *Stat*, 9 (1):e318, 2020. doi: 10.1002/sta4.318.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, Cham, 2nd edition, 2022. doi:10.1007/978-3-031-06649-8.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 billion parameter autoregressive language model, 2021. https://github.com/kingoflolz/mesh-transformer-jax.

Peng Wang, Biyu Zhou, Xuehai Tang, Jizhong Han, and Songlin Hu. LyapLock: Bounded knowledge preservation in sequential large language model editing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://aclanthology.org/2025.emnlp-main.327/.

Xinyu Wang, Sicheng Lyu, Yu Gu, Jerry Huang, Peng Lu, Yufei Cui, and Xiao-Wen Chang. Beyond hard writes and rigid preservation: Soft recursive least-squares for lifelong LLM editing (RLSEdit). In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-ECAI)*, 2026. arXiv:2601.15686.

Wikimedia Foundation. Wikipedia 20220301.en HuggingFace dataset, 2022. https://huggingface.co/datasets/wikipedia.

Paweł Wójcik. On mappings approximately transferring relations in finite-dimensional normed spaces. *Linear Algebra and its Applications*, 460:125–135, 2014. doi: 10.1016/j.laa.2014.07.046.

Jianhao Yan, Futing Wang, Yun Luo, Yafu Li, and Yue Zhang. Keys to robust edits: From theoretical insights to practical advances. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. arXiv:2410.09338.

An Yang et al. Qwen3 technical report. *arXiv preprint*, 2025. arXiv:2505.09388.

Ye Zhang, Yanni Chen, Don Hadwin, and Liang Kong. A note on linear approximately orthogonality preserving mappings. *arXiv preprint arXiv:1409.7781*, 2014. doi: 10.48550/arXiv.1409.7781.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *EMNLP*, 2023.

Guide to the appendices. Section A collects the notation and Section B extends the related work. Section J proves Theorems 1 and 2. Sections C and G cover constrained placement, its proofs and the matching lower bound on the sub-Weibull maximum. Section H details the geometry and placement experiments. Sections D to F cover the response-scale analyses, cohorts and provenance, and Section I proves the tail bound.

A Notation

Table 6 collects the symbols of the main text and the appendices. Its first block defines the directional risk and the quantities built from it. The second block covers edits, probes and tails.

B Extended related work

Second-moment placement. ROME places along $\Sigma_\psi^{-1} k_\star$ (Meng et al., 2022), and MEMIT carries the same second-moment objective to batched multi-layer updates (Meng et al., 2023). AlphaEdit uses second moments to build a preservation subspace, and a regularized solve writes the update inside it (Fang et al., 2025). Proposition 4 obtains the covariance and Euclidean constrained solutions from one closed form by the choice of quadratic metric.

Preservation constraints in sequential editing. Null-space preservation has been extended to sequential editing with history-aware updates (Liu et al., 2026a) and replaced by a soft recursive least-squares update (Wang et al., 2026). Other work studies the stability of long edit sequences (Cao et al., 2026; Ma et al., 2026).

Robustness and specificity. Error bounds for the semantic keys that localize an edit show that robustness to paraphrased or long-context queries trades off against specificity (Yan et al., 2025). An information-bottleneck framework targets both properties (Chen et al., 2026). Minimal-norm weight changes that produce a prescribed output change have been derived for deep networks (Evans & Tanner, 2026).

Tails of activations. Transformer analysis and quantization studies report activation outliers and measure them per coordinate (Dettmers et al., 2022; Kovaleva et al., 2021; Sun et al., 2024; Kaliaperumal, 2026). The Orlicz scale here describes the response along one edit direction, and the empirical Weibull shape parameter $\hat{\theta}$ is fitted per edit.

C Constrained inverses under different quadratic metrics

The covariance-weighted and Euclidean constrained solutions minimize a quadratic cost $v^\top G v$ for two different metrics G. Each minimum is taken over the preservation subspace at the required gain $\gamma = v^\top k_\star$. One closed form covers every G.

Table 6: Notation used throughout. In the first block, a parenthesis names where a quantity is computed: population, fit half or held-out half.

Symbol	Meaning
Directional risk and the quantities built from it	
$R_M(z)$, $R_M^{\mathrm{eval}}(z)$	expected maximum absolute response along z over M probes: population, or empirical on the fit half; R_M^{eval} is the empirical value on the held-out half
a, z_Σ	whitened target $L^\top k_\star$ and unit-gain covariance direction $a/\|a\|_2^2$ in the whitened feasible space; the experiments whiten by the fit-half second moment
$\kappa_M(a)$, κ_M	risk ratio $R_M(z_\Sigma)/\min_{a^\top z=1} R_M(z)$ at target a, and the worst-target risk ratio $\kappa_M = \sup_{a\neq 0} \kappa_M(a)$ (population; fit half in the experiments)
C_M	directional variation: ratio of the maximum to the minimum of R_M on the unit sphere (population)
$\widehat{C}_M$, $\widehat{C}_M^{\mathrm{eval}}$	largest such ratio that the direction search attains on the fit half, a lower bound on the fit-half C_M; $\widehat{C}_M^{\mathrm{eval}}$ is the ratio of the same two directions on the held-out half
$H_M(a)$	held-out risk ratio $R_M^{\mathrm{eval}}(z_\Sigma)/R_M^{\mathrm{eval}}(\hat{z}_{\mathrm{fit}})$, where $\hat{z}_{\mathrm{fit}}$ minimizes the fit-half R_M subject to $a^\top z = 1$ (held-out half)
D_M	null-normalized diagnostic: pooled median of H_M over a target class divided by the same median in the spherical reference, a ratio of medians; its excess is $100(D_M - 1)$
Edits, probes and tails	
ℓ	edited layer (MLP down-projection)
ΔW_ℓ; $\alpha\, uv^\top$	captured weight update at layer ℓ; $\alpha uv^\top$ is a rank-one update or the top singular component of ΔW_ℓ
α	edit scale
u, v	output (write) direction and key direction
r	unit readout direction; in the response-scale analyses, the normalized unembedding row of the requested new object's first token, fixed by the edit request before any evaluation probe is drawn
k, k_i	probe activations drawn from $\mathcal{D}_\ell$
$k_\star$	target-fact key
$\mathcal{D}_\ell$	layer-ℓ activation distribution
M, M_{fit}	probe budget: number of probes in the maximum; budget-fitted placement (LP) minimizes the fit-half R_M at M_{fit}
$s_\ell(k) = r^\top \Delta W_\ell k = w^\top k$	per-probe response, with $w = \Delta W_\ell^\top r$
θ, $\hat{\theta}$	sub-Weibull tail index (2 sub-Gaussian, 1 sub-exponential); $\hat{\theta}$ is the empirical Weibull shape parameter fitted per edit
ψ_θ, $\|\cdot\|_{\psi_\theta}$	Orlicz function $\exp(x^\theta) - 1$ and threshold-2 Orlicz functional (Section I)
K, K_w	per-edit Orlicz scales $\|v^\top k\|_{\psi_\theta}$ of the key projection and $\|w^\top k\|_{\psi_\theta}$ of the full response; $\widehat{K}_w$ is fitted on the fit half, and $\hat{K} = \widehat{K}_w/\|\alpha(r^\top u)\|$ estimates K for a rank-one update
$\widehat{B}_M$	fitted sub-Weibull predictor $\widehat{K}_w(2\log M)^{1/\hat{\theta}}$, equal to $(2\log M)^{1/\hat{\theta}}\|\alpha\|\|r^\top u\|\hat{K}$ for a rank-one update
$\tilde{u}$	fixed output vector of the rank-one update $\tilde{u}v^\top$ in Section J.1
n_e	edits applied per cell
$\Sigma_\psi = \mathbb{E}[kk^\top]$	activation second-moment matrix
τ, λ_1	spectral truncation cutoff and top eigenvalue
$\Sigma_\psi^{+,\tau}$; Q	truncated covariance inverse (A in Theorem 5); Q is an editor's placement operator, $\Sigma_\psi^{+,\tau}$ or $Q_{K_{\mathrm{pres}}}$ of eq. (3)
G, Q_G	quadratic cost metric and its constrained inverse (Proposition 4); $G{=}\Sigma_\psi$ is the covariance metric, and $G{=}I$ gives AlphaEdit's null-space projector $P_\perp$
K_{pres}	matrix whose columns are the preservation keys
ρ	predictor inflation factor (Section D.3)
ρ_{S}, r_{P}	Spearman and Pearson correlation coefficients
d_{in}, n	MLP input dimension and number of fitting probes in the empirical Orlicz equation

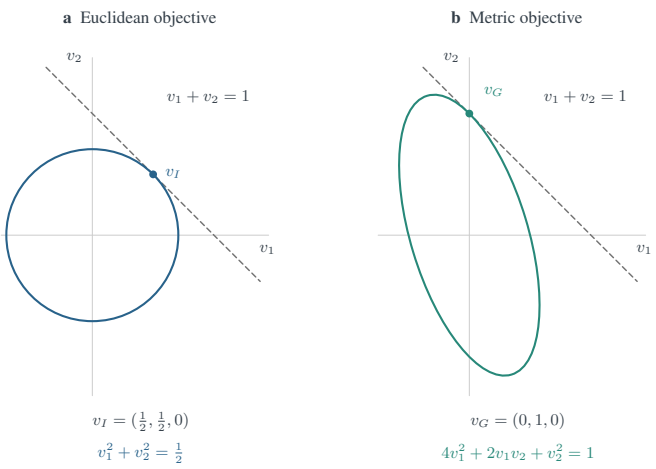

Figure 6: Two metrics select different directions within the same preservation subspace. Exact example with $K_{\text{pres}} = e_3$ and $k_\star = (1, 1, 0)^\top$, drawn in the plane $v_3 = 0$ that is the range of both operators. Each panel shows the level set of its objective through its minimizer, tangent to the dashed gain constraint. **a** Euclidean objective $v^\top v$. **b** Metric objective $v^\top G v$ with $G = [\begin{smallmatrix} 4 & 1 \\ 1 & 1 \end{smallmatrix}]$ on that plane. Under G, the cost is $7/4$ at the Euclidean minimizer v_I and 1 at v_G.

Proposition 4 (Constrained inverse for a general metric G)**.** *Let $G \succ 0$, let $K := K_{\text{pres}}$ have full column rank, and assume $k_\star \notin \text{range}(K)$. For any prescribed gain γ, the unique minimizer of $v^\top G v$ subject to $K^\top v = 0$ and $v^\top k_\star = \gamma$ is*

$$v^\star = \gamma \frac{Q_G k_\star}{k_\star^\top Q_G k_\star}, \qquad Q_G = G^{-1} - G^{-1}K(K^\top G^{-1} K)^{-1} K^\top G^{-1}.$$

If N is an orthonormal basis of $\ker K^\top$, then $Q_G = N(N^\top G N)^{-1} N^\top$. Thus Q_G is symmetric positive semidefinite with range $\ker K^\top$. At $G = I$ it equals the Euclidean null-space projector $P_\perp = I - K(K^\top K)^{-1} K^\top = NN^\top$ (Fang et al., 2025). The normalized updates $Q_G k/(k^\top Q_G k)$ and $P_\perp k/(k^\top P_\perp k)$ agree for every feasible key k if and only if $N^\top G N = cI$ for some $c > 0$. In that case $Q_G = c^{-1} P_\perp$. The equality $Q_G = P_\perp$ requires $N^\top G N = I$.

The Euclidean and covariance metrics share the preservation subspace $\ker K^\top$ and weight directions within it differently (Figure 6). Setting $G = \Sigma_\psi \succ 0$ gives the untruncated covariance solution. The deployed truncated inverse $\Sigma_\psi^{+,\tau}$ of Theorem 5 restricts the optimization to its range. On that range the same construction gives the operator $Q_{K_{\text{pres}}}$ of eq. (3), whose range is $\text{range}(\Sigma_\psi^{+,\tau}) \cap \ker K^\top$.

C.1 The covariance-metric quadratic direction

Without preservation keys, covariance placement minimizes the quadratic cost $v^\top \Sigma_\psi v$ over a retained eigenspace at a prescribed gain.

Theorem 5 (Covariance-metric quadratic optimizer)**.** *Let R be a retained eigenspace of Σ_ψ with strictly positive eigenvalues, and let A equal its inverse on R and zero on $R^\perp$. If $k_\star^\top A k_\star > 0$, the unique minimizer of $v^\top \Sigma_\psi v$ over $v \in R$ with prescribed gain $v^\top k_\star = \gamma$ is $v^\star = \gamma A k_\star / (k_\star^\top A k_\star)$. Deployment uses $A = \Sigma_\psi^{+,\tau}$ with $\tau = 10^{-3}\lambda_1$, restricting the feasible set to the retained modes.*

The optimization in Theorem 5 is the classical minimum-variance quadratic program (Capon, 1969). Equation (3) adds linear preservation constraints (Frost, 1972).

C.2 Adding an explicit preservation constraint

The preservation keys form the columns of K_{pres}. Imposing $K_{\text{pres}}^\top v = 0$ makes the layer update vanish at each of those keys:

$$\min_{v \in R} \; v^\top \Sigma_\psi v \quad \text{s.t.} \quad K_{\text{pres}}^\top v = 0, \;\; v^\top k_\star = \gamma. \tag{2}$$

Here R is the retained space of Theorem 5, and γ is a prescribed feasible gain. The solution is $v = \gamma Q_{K_{\mathrm{pres}}} k_\star / (k_\star^\top Q_{K_{\mathrm{pres}}} k_\star)$ when the denominator is positive, with

$$Q_{K_{\mathrm{pres}}} := \Sigma_\psi^{+,\tau} - \Sigma_\psi^{+,\tau} K_{\mathrm{pres}} \big(K_{\mathrm{pres}}^\top \Sigma_\psi^{+,\tau} K_{\mathrm{pres}}\big)^\dagger K_{\mathrm{pres}}^\top \Sigma_\psi^{+,\tau}. \tag{3}$$

The Gram pseudoinverse keeps eq. (3) defined when the projected constraints are linearly dependent. Whenever the gain is feasible, the solution satisfies $K_{\mathrm{pres}}^\top v = 0$ exactly.

D Coverage, calibration, and reproducibility

D.1 Models and edits

We edit GPT-J-6B (Wang & Komatsuzaki, 2021), Pythia-2.8B (Biderman et al., 2023), Llama-3.1-8B (Grattafiori et al., 2024), Qwen3-14B (Yang et al., 2025) and Phi-4 (Abdin et al., 2024). The 176 cells of the canonical corpus cover every architecture on six datasets: CounterFact (Meng et al., 2022), ZsRE (Levy et al., 2017), RippleEdits (Cohen et al., 2024), MQuAKE (Zhong et al., 2023), UniEdit (Chen et al., 2025) and WikiBigEdit (Thede et al., 2025). The placement intervention writes rank-one ROME-family edits at MLP down-projections. The canonical corpus and the stored-direction frontier span the editors of Section E. The canonical corpus scores the full response $|r^\top \Delta W_\ell k|$ over Wikipedia tokens k, which measures the tail along the edit direction w. The frontier scores $|v_1^\top k|$ for each archived leading right singular vector v_1.

D.2 Predicted against observed maxima

Figure 7 plots predicted against observed maxima at $M{=}5{,}000$ for the fitted predictor and the second-moment reference. On every benchmark, the fitted predictor covers more than 97.1% of edits on held-out probes (Table 7).

Table 7: Coverage by benchmark at $M{=}5{,}000$. The fitted predictor is estimated on the first $5{,}000$ probes and scored on the held-out $5{,}000$. The second-moment reference is scored on the first $5{,}000$. The rows partition the canonical corpus.

Benchmark	Edits	**Fitted predictor**	Second-moment reference
UniEdit (Chen et al., 2025)	1,208	$\mathbf{99.92}\%$	79.9%
RippleEdits (Cohen et al., 2024)	1,820	$\mathbf{98.57}\%$	52.6%
CounterFact	8,150	$\mathbf{98.26}\%$	62.4%
ZsRE	1,240	$\mathbf{97.66}\%$	38.5%
WikiBigEdit (Thede et al., 2025)	1,780	$\mathbf{97.19}\%$	56.8%
MQuAKE (Zhong et al., 2023)	1,780	$\mathbf{97.13}\%$	64.5%

D.3 Marginal calibration of the inflation factor

Fix the score construction in advance: an independent fitting dataset of unrelated probes and a probe protocol that includes the evaluation budget M. The fitting dataset needs no locality labels. Each candidate update is computed on a copy of the model before its evaluation probes are drawn. For each edit define $S = O/\widehat{B}$, where O is the observed maximum over those probes and $0 < \widehat{B} < \infty$ is the edit's fitted predictor. Compute the n_{cal} calibration scores and one fresh score under the same complete selection policy, including any gating or layer selection. Suppose these $n_{\mathrm{cal}} + 1$ scores are jointly exchangeable conditional on the fitting dataset. For a prespecified $\delta \in (0, 1)$, let $q = \lceil (1-\delta)(n_{\mathrm{cal}} + 1) \rceil$ and set

$$\rho = \begin{cases} S_{(q)}, & q \le n_{\mathrm{cal}}, \\ +\infty, & q = n_{\mathrm{cal}} + 1, \end{cases}$$

where $S_{(q)}$ is the qth smallest calibration score. Exchangeability of the rank of the fresh score, treating ties conservatively, yields the split-conformal guarantee (Papadopoulos et al., 2002; Lei et al., 2018; Vovk et al., 2022)

$$\Pr\Big[O_{\mathrm{new}} \le \rho \widehat{B}_{\mathrm{new}} \mid \text{fitting data}\Big] \ge \frac{q}{n_{\mathrm{cal}} + 1} \ge 1 - \delta.$$

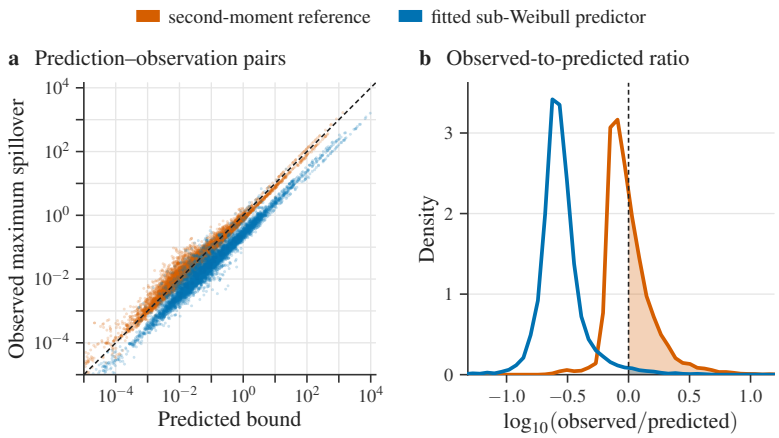

Figure 7: Archived prediction–observation pairs at M=5,000. The second-moment reference is scored on the first 5,000 probes (16,318 pairs). The fitted predictor is estimated on those probes and scored on the held-out 5,000 (16,298 pairs). Some cells enter twice through duplicate archive copies, which the canonical corpus counts once. **a** Observed maximum against predicted bound, a random subsample of at most 8,000 pairs per series. **b** Density of $\log_{10}(\text{observed}/\text{predicted})$ over all pairs, central 99.9% shown; shading marks exceedances, 40.18% of reference pairs and 1.85% of fitted-predictor pairs. Dashed lines mark equality.

The probability is marginal over the calibration sample, the fresh edit and its probes. Table 8 reports held-out exceedances on the canonical corpus for a fixed grid of ρ.

Table 8: Held-out exceedances of $\rho\widehat{B}_M$ on a fixed grid of inflation factors. M=5,000, over the 15,978 canonical edits of Table 9.

ρ	1	1.2	1.5	2	3	5
Exceedances	299	197	128	66	25	2
Coverage (%)	98.13	98.77	99.20	99.59	99.84	99.99

D.4 Broader impact

This work measures the side effects of knowledge edits on unrelated inputs. The same editors can also write false or harmful facts into a model. The diagnostics measure a layer-level response on a specified probe population. A deployment report that uses them should state that population and the endpoint.

D.5 Reproducibility

Provenance. The archived per-cell records supply the coverage and benchmark results. For 125 of the 176 canonical cells (12,605 of 15,978 records), the raw responses are reconciled with the canonical summaries by record identity and every recorded evaluation-prefix maximum. These cells include all 33 Llama-3.1-8B, 27 Pythia-2.8B, 23 Qwen3-14B and 18 Phi-4 cells, and 1,780 MQuAKE records whose identifier mapping was reconstructed. The other 51 cells, all GPT-J-6B with 3,373 records, rest on their archived summaries alone. The headroom decomposition, the placement results and the stored-direction frontier use separately collected banks of 20,000 activations, one token per article. Table 10 gives the source cohort of each main-text result.

Code and data. The supplementary material contains the method package, per-record values extracted from the archives, and a script for each regenerated figure and table that reads those values. Four drivers recompute the archived results from their inputs. The geometry driver reads edit-key files and probe banks. The placement driver reads model weights, probe banks and the AlphaEdit code base. The frontier driver reads probe banks and the archived update directions. The coverage driver recomputes Table 5 from the per-edit response arrays of the estimator-calibration cohort.

Seeds and software. Canonical single-edit summaries carry seed 0, and the supplementary scripts fix every resampling seed. Model revisions are recorded as Hugging Face commit hashes with the archive, and the supplementary README lists the library versions of the scripts and drivers.

Probe collection. The contiguous caches come from English Wikipedia (Wikimedia Foundation, 2022). Collection starts at article offset 200, excludes padding and the first two token positions, and collects activations in article order. The stride can equal one, so neighboring probes may share an article and context.

Compute. Edits, second-moment estimates and probe caches were computed on NVIDIA V100 GPUs. Each model layer's second-moment matrix and probe cache are computed once and reused across editors.

AlphaEdit reproduction. On Fang et al. (2025)'s GPT-J/CounterFact setting at n=2,000, our AlphaEdit implementation gives efficacy 0.9975, paraphrase 0.968 and neighborhood 0.7524, a harmonic score of 0.8916. Fang et al. (2025) report 0.9975, 0.9638 and 0.7548 in this setting. These values give a harmonic score of 0.8916.

E PROTOCOL, COHORTS, AND TAIL-SCALE CONTROLS

Edits, probes, and tail-fitting protocol. The 176 cells span $n_e \in \{20, 50, 200\}$, six datasets (CounterFact, ZsRE, RippleEdits, MQuAKE, UniEdit, WikiBigEdit) and five architectures (GPT-J-6B, Pythia-2.8B, Llama-3.1-8B, Qwen3-14B (dense), Phi-4). The archives span fifteen editor configurations: Σ_ψ-Proj (Q=$\Sigma_\psi^{+,\tau}$, the operator of Theorem 5), Σ_ψ-Proj+K_{pres} (eq. (3), the retained-space form of Proposition 4), Euclid+K_{pres}, AlphaEdit (Fang et al., 2025), EAMET (Dai et al., 2026), PMET (Li et al., 2024), SPHERE (Liu et al., 2026b), ROME (Meng et al., 2022), LyapLock (Wang et al., 2025), NSE (Jiang et al., 2025), EvoEdit (Lyu et al., 2026), SUIT (Park et al., 2026), AlphaEdit+PRUNE and ROME+PRUNE (Ma et al., 2025), and AlphaEdit-Aligned, the AlphaEdit variant distributed with the EAMET codebase for its aligned evaluation protocol. For each edit we record the top singular triple (α, u, v) of the captured update $\Delta W_\ell = W_\ell^{\mathrm{post}} - W_\ell^{\mathrm{pre}}$. The probe response $|r^\top \Delta W_\ell k_i|$ uses the full update. The fitted predictor $\widehat{B}_M$ and the per-edit $\theta = 2$ control are both fitted to this full response on the same fitting half. Because $\hat{\theta}$ is scale-free and the empirical Orlicz scale is homogeneous, dividing the response by a nonzero factor $|\alpha(r^\top u)|$ divides $\widehat{B}_M$ by the same factor. The second-moment reference of Figure 4 is computed from the top singular triple and Σ_ψ. The captured update ΔW_ℓ is zero in 80 records, which come from the four L_{12} cells of AlphaEdit and Σ_ψ-Proj+K_{pres} on Phi-4 and Qwen3-14B. Their archived responses are zero. These records carry no tail fit and count as covered. Coverage is 98.13% with them and 98.12% without them.

Probes are up to 10,000 English Wikipedia (Wikimedia Foundation, 2022) activations from each cell's contiguous cache, collected as in Section D.5. The first two token positions are excluded (Sun et al., 2024). They are centered with the mean of all 10,000 probes and split into a fitting half (the first 5,000) and a held-out half (the last 5,000). Per edit, $\hat{\theta}$ is the maximum-likelihood Weibull shape (Virtanen et al., 2020) of the top 5% of the fitting-half responses, treated as a complete sample. Over the 15,898 edits with a finite fit, $\hat{\theta}$ has median 1.09 and interquartile range $[1.02, 1.15]$. On the 11,835 edits of the estimator-calibration re-analysis, the per-edit excess kurtosis of the response has median 2.78. On the same edits, $\hat{\theta}$ has Pearson correlation $r_{\mathrm{P}} = -0.701$ with the Hill estimate of the extreme-value index (Hill, 1975), which grows with tail heaviness.

Centering. For 2,349 sampled GPT-J and Pythia edits scored on their contiguous caches, we recenter the probes with the fitting-half mean. At every budget from 500 to 5,000, the held-out expected maximum changes by a median of at most 0.07% over edits (99th percentile 3.2%, largest 6.3%).

Cohort definitions. Table 9 lists the cohort behind each reported count. Table 10 gives the cohort, probe bank and split behind each main-text figure and table.

Table 9: Cohorts behind the reported counts. Indented rows are subsets of the nearest unindented row above. The architecture rows and the two preservation rows each partition the canonical corpus.

Cohort	Edits	Filter
Canonical corpus (176 cells)	15,978	seed-0 strict subset at M=5,000
by architecture	4,873	GPT-J-6B
	3,328	Llama-3.1-8B
	3,101	Pythia-2.8B
	2,456	Phi-4
	2,220	Qwen3-14B
K_{pres}-protected (158 cells)	15,228	editors carrying a preservation constraint
no-preservation (18 cells)	750	unconstrained baselines
finite $\hat{\theta}$ fit	15,898	$\Delta W_\ell \neq 0$; behind the $\hat{\theta}$ summaries
covered by $\widehat{B}_M$ (98.13%)	15,679	held-out maximum below $\widehat{B}_M$, of all 15,978
Estimator-calibration re-analysis (119 cells)	11,835	cells retaining raw probe records
distinct nonzero response vectors	11,439	each counted once; Table 5
Stored-direction frontier (147 groups)	41,208	leading right singular vectors, deduplicated within each architecture–layer bank and rescored on the one-token-per-article bank
variance proxy available	13,016	direction records carrying $v_1^\top \Sigma_\psi v_1$
Headroom decomposition (17 cells)	680	target measurements over 329 CounterFact facts, Section H
ROME placement	3×150	requests per cell–fit setting behind Tables 2 and 21

E.1 Quantile, maximum and fixed-exponent scales on the frontier cohort

Paired protocol. The 41,208 stored-direction records of Section 5.2 are scored on the one-token-per-article banks, one bank of 20,000 activations per architecture–layer. Each bank is centered with the mean of all its 20,000 activations and split in dump order into a fitting half (the first 10,000 articles) and an evaluation half (the last 10,000). Let $s_i = v_1^\top k_i$ for the $n = 10{,}000$ fitting probes. We add five scales to those of Section 5.2: the maximum of $|s_i|$; the empirical 0.95, 0.99 and 0.999 quantiles of $|s_i|$, interpolated linearly at index $(n-1)p$; and a fixed-$\theta = 1$ Orlicz scale, which solves $n^{-1}\sum_i \exp(|s_i|/K_1) = 2$ and predicts $2K_1 \log M$. All scales share the records, the two halves and the archived sampled maxima. The maximum and the quantiles enter unscaled, because $\mathrm{EFF}_q(c)$ is unchanged when every record's score at one budget is multiplied by the same positive constant. The comparison measures per-record differences in scale.

Results. At the 90th percentile of looseness, the 99.9% quantile and the fitted exponent are tighter than fixed $\theta = 2$ at all 28 points (Table 12). At $M = 5{,}000$ and $c = 0.98$ their ratios are 0.954 $[0.882, 1.032]$ and 0.915 $[0.876, 0.960]$. The interval contains one for the 99.9% quantile and excludes one for the fitted exponent. Fixed $\theta = 1$ is tighter at 22 points and the 99% quantile at 16.

F Sub-Weibull bound: constants and estimation

For $M \geq 3$, the choice $\delta = 2/M < 1$ in Theorem 3 gives the factor $(2\log M)^{1/\theta}$ of $\widehat{B}_M$. With probability at least $1 - 2/M$, the maximum response is then at most $|\alpha||r^\top u|\, K(2\log M)^{1/\theta}$, where $K = \|v^\top k\|_{\psi_\theta}$. Under Vershynin's threshold-2 definition (Vershynin, 2018, Def. 2.5.6), a standard Gaussian $Z \sim \mathcal{N}(0,1)$ has $\|Z\|_{\psi_2}^2 = 8/3$, so $K = \sigma\sqrt{8/3}$ when $v^\top k \sim \mathcal{N}(0, \sigma^2)$. At $\theta = 2$ the directional term is then $K\sqrt{2\log M} = (4/\sqrt{3})\,\sigma\sqrt{\log M} \approx 2.31\,\sigma\sqrt{\log M}$, which is $2/\sqrt{3} \approx 1.15$ times the Gaussian Chernoff threshold $2\sigma\sqrt{\log M}$ (Boucheron et al., 2013).

Table 10: The cohort and probe bank behind each main-text float. The contiguous cache holds tokens from about twenty articles. The diverse bank is the one-token-per-article bank of Section 4, drawn from 20,000 articles.

Float	Cohort	Probe bank and split
Figure 1	constructed laws	none; the panels are computed from the stated distributions
Table 1, Figure 2(a,b)	headroom decomposition, 680 target measurements in 17 cells over 329 facts; principal-direction and probe-row rows from a separate run over 341 facts	diverse bank, article-disjoint halves
Figure 2(c)	the 150 ROME requests of Table 2	diverse bank for that cell, article-disjoint halves
Figure 3(a,b,d)	the five cells measured at every width, 200, 200 and 60 measurements at widths 16, 32 and 64	diverse bank, article-disjoint halves
Figure 3(c)	two cells at 12 targets each on the contiguous cache and 40 on the diverse bank	the contiguous cache split by token, the same cache split at the article boundary, and the diverse bank
Table 2	150 ROME requests per cell; written-matrix rows from the runs of Table 22	diverse bank for that architecture–layer, article-disjoint halves
Table 3	the same 150 requests, in runs separate from Table 2	the same bank, affine family and gain constraint for both arms
Table 4	definitional	none
Table 5	estimator-calibration re-analysis, 11,439 distinct nonzero response vectors in 119 cells	archived per-cell probe records, fit half and held-out half
Figure 4	canonical corpus by architecture, 4,873/3,328/3,101/2,456/2,220 edits	each cell's archived contiguous cache at M=5,000; the fitted predictor on the held-out half, the second-moment reference on the first half
Figure 5	stored-direction frontier, 41,208 records in 147 groups, with the variance proxy on 13,016	diverse bank, first and last 10,000 articles in dump order

Table 11: The median-normalized expected maxima of the stored leading direction and of the full update agree within 5% for at least 97% of the sampled edits. Per-edit ratio of the full update's median-normalized expected maximum to that of the stored direction. Dividing by the median cancels the rank-one factor $|\alpha(r^\top u)|$. The 2,342 edits are sampled (at most 40 per archive) from the GPT-J and Pythia archives and scored on their original contiguous caches. Deviations are measured from one.

M	Median ratio	5th pct	95th pct	Largest deviation	Within 1% / 5%
128	0.9998	0.9868	1.0038	0.534	93.0% / 98.1%
512	0.9998	0.9850	1.0041	0.698	91.7% / 97.7%
2,048	0.9998	0.9814	1.0044	0.829	90.7% / 97.4%
5,000	0.9998	0.9809	1.0049	0.871	90.2% / 97.3%

The empirical Orlicz scale. The Weibull fit supplies the exponent, and the equation below supplies the scale. For an exponent θ, the scale $\widehat{K}_w(\theta)$ is the smallest $K > 0$ satisfying

$$n^{-1} \sum_{i=1}^{n} \exp\{(x_i/K)^\theta\} = 2, \qquad x_i = |r^\top \Delta W_\ell k_i|,$$

over the n fitting probes, found by a bracketed root search. The fitted predictor uses $\theta = \hat{\theta}$, and row (1) of Table 5 uses $\theta = 2$. For $\theta = \hat{\theta}$ the equation forces $\max_i x_i \le \widehat{K}_w(\log(2n))^{1/\hat{\theta}}$ on the

Table 12: At the 90th percentile of looseness, the 99.9% quantile and the fitted exponent are the two scales tighter than fixed $\theta = 2$ at all 28 points. All 41,208 stored-direction records. Looseness is the ratio of the threshold, inflated to coverage c, to the observed maximum. Ratios divide each scale's median or 90th-percentile looseness by that of fixed $\theta = 2$ (2.831 and 4.867) at $M = 5{,}000$ and $c = 0.98$. Intervals come from 2,000 paired resamples of the 147 architecture–layer–archive groups (an archive is one editor and dataset), with articles and directions held fixed. Tighter / 28 counts the budget–coverage points at which a scale's looseness is below that of fixed $\theta = 2$.

	Looseness ratio [95% interval]		Tighter / 28	
Scale	Median	90th percentile	Median	90th
RMS	1.362 [1.265, 1.470]	1.302 [1.154, 1.464]	0	10
Fit quantile 95%	1.430 [1.329, 1.551]	1.371 [1.219, 1.551]	0	5
Fit quantile 99%	1.383 [1.301, 1.488]	1.176 [1.066, 1.312]	0	16
Fit quantile 99.9%	1.264 [1.201, 1.329]	0.954 [0.882, 1.032]	1	28
Fit maximum	0.975 [0.969, 0.993]	1.030 [1.024, 1.046]	16	0
Orlicz, $\theta = 1$	1.131 [1.086, 1.191]	1.017 [0.950, 1.095]	0	22
Orlicz, $\theta = 2$	1.000 (reference)	1.000 (reference)	—	—
Orlicz, fitted $\hat{\theta}$	1.083 [1.049, 1.120]	0.915 [0.876, 0.960]	1	28

Table 13: In median looseness, fixed $\theta = 2$ is at least as tight as the variance proxy at all 28 points. The 13,016 records that carry a variance proxy, in 93 of the groups. All entries, including the fixed-$\theta = 2$ reference, are computed on this subset as in Table 12. The proxy's scale comes from the archived second moment. The other scales come from the fitting-half projections.

	Looseness ratio [95% interval]		Tighter / 28	
Scale	Median	90th percentile	Median	90th
RMS	1.834 [1.677, 1.966]	1.755 [1.586, 1.894]	0	2
Fit quantile 95%	1.968 [1.770, 2.126]	1.874 [1.669, 2.031]	0	1
Fit quantile 99%	1.647 [1.513, 1.775]	1.514 [1.357, 1.641]	0	5
Fit quantile 99.9%	1.288 [1.220, 1.359]	1.122 [1.040, 1.194]	0	15
Fit maximum	0.977 [0.955, 0.996]	1.047 [1.024, 1.062]	15	0
Orlicz, $\theta = 1$	1.357 [1.266, 1.418]	1.276 [1.175, 1.346]	0	9
Orlicz, $\theta = 2$	1.000 (reference)	1.000 (reference)	—	—
Orlicz, fitted $\hat{\theta}$	1.115 [1.081, 1.154]	1.027 [0.981, 1.070]	0	21
Variance proxy	2.020 [1.838, 2.177]	2.003 [1.804, 2.169]	0	1

fitting probes, so in-sample coverage of $\widehat{B}_M$ holds by construction when $n \leq M^2/2$. Coverage of $\widehat{B}_M$ is measured on held-out probes.

On the 11,835 edits of the estimator-calibration re-analysis, let $\hat{K}_2 = \widehat{K}_w(2)$ and let $\hat{\sigma}$ be the root mean square of the fitting responses. The ratio $\hat{K}_2/\hat{\sigma}$ has median 1.90, standard deviation 1.13 and range $[1.53, 15.73]$. For 4.8% of edits, the ratio lies within 1% of the Gaussian plug-in value $\sqrt{8/3}$ (Figure 8).

F.1 Coverage with the cell as the unit of analysis

Edits within one cell share a probe set, so the interval below resamples whole cells. Over the 176 canonical cells, the mean held-out coverage of $\widehat{B}_M$ is 97.88% (95% interval $[97.39, 98.38]$). Pooled over edits, coverage is 98.13%. The median cell is at 100%. In 91 cells (51.7%), $\widehat{B}_M$ covers every edit, and 161 cells (91.5%) have coverage of at least 95%. The fitted predictor covers at least as many edits as the second-moment reference in every cell, and more in 160 (Figure 9).

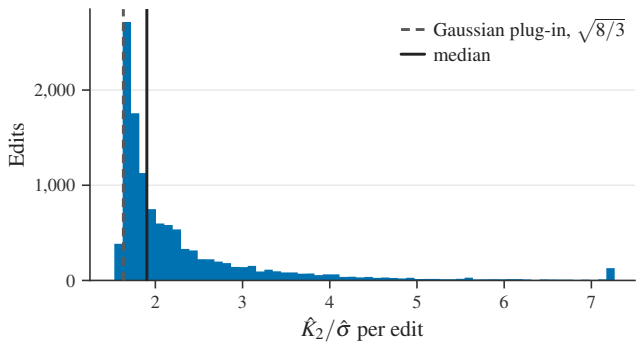

Figure 8: The fitted θ=2 scale of row (1) in Table 5 exceeds the Gaussian plug-in for 96.1% of edits. Histogram of $\hat{K}_2/\hat{\sigma}$ over the 11,835 edits of the estimator-calibration re-analysis, with $\hat{\sigma}$ the root mean square of the fitting responses. The last bar collects the edits above the 99th percentile.

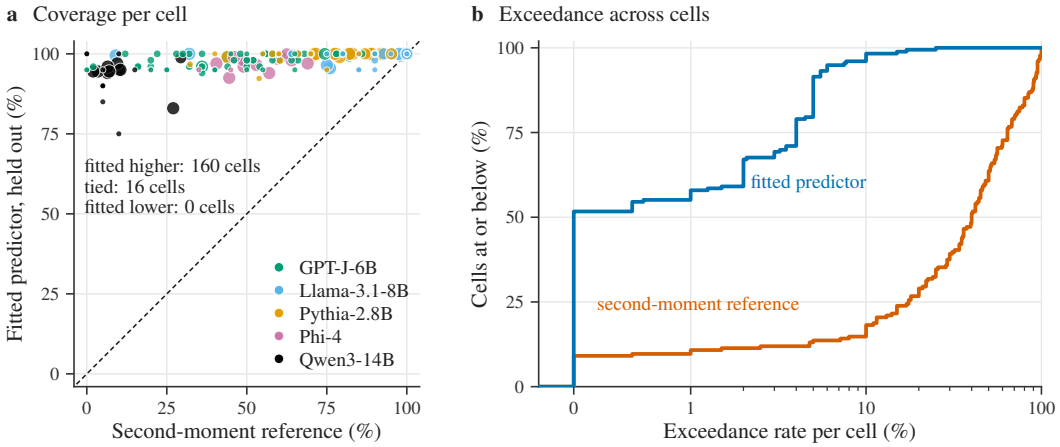

Figure 9: The fitted predictor covers at least as well as the second-moment reference in every canonical cell. Canonical corpus at M=5,000. As in Figure 4, the reference is scored on the first 5,000 probes and $\widehat{B}_M$ on the held-out 5,000. **a** Per-cell coverage, one point per cell with marker area growing with its edit count; tied cells have full coverage under both predictors. **b** Cumulative distribution over cells of the exceedance rate, on a symmetric-log axis so that cells without exceedance sit at zero.

F.2 Shape estimators on laws with known tail index

We draw n=5,000 samples (30 replicates) from six sub-Weibull laws with θ from 0.5 to 2 and from a Pareto control whose Orlicz norm is infinite for every θ. Each draw is fit twice: by the deployed estimator (the top-5% fit of Section 5.1) and by a conditional peaks-over-threshold (POT) maximum-likelihood fit, which conditions on exceeding the 95th percentile. The deployed estimator orders the seven laws by tail heaviness, with values in $[0.87, 1.19]$ (Table 14). We therefore read $\hat{\theta}$ as an empirical shape parameter.

Estimator choice. On the 11,835 edits of the estimator-calibration re-analysis, the median $\hat{\theta}$ is 1.06 for conditional POT and 1.10 for the deployed estimator. With $\widehat{K}_w$ refitted at each exponent, $\widehat{B}_M$ has held-out coverage 97.98% with the POT exponent and 97.89% with the deployed one.

Per-architecture medians of $\hat{\theta}$ range from 1.15 on Llama-3.1-8B to 0.99 on Qwen3-14B (Figure 10). Across the five architectures, the coverage of the second-moment reference tracks $\hat{\theta}$ (Spearman ρ_S=0.90) and spans more than seventy points. The fitted predictor uses a per-edit $\hat{\theta}$, and its held-out exceedance rate is at most 6.0% in every architecture (Figure 4).

Table 14: Conditional POT tracks the true index, and the deployed estimator orders the laws with values from 0.87 to 1.19. The conditional-POT interquartile range contains the true θ for every law with a finite Orlicz norm, and the fit falls to 0.11 on the Pareto control. The last column is the deployed estimator of Section 5.1 on the same draws.

Law	True θ	Conditional-POT $\hat{\theta}$ Median [IQR]	Deployed $\hat{\theta}$
Gaussian	2.0	1.67 [1.50, 2.23]	1.19
Weibull(1.5)	1.5	1.51 [1.23, 1.82]	1.17
Weibull(1.1)	1.1	1.27 [1.02, 1.46]	1.12
Exponential	1.0	1.02 [0.82, 1.20]	1.08
Weibull(0.7)	0.7	0.72 [0.66, 0.86]	0.99
Weibull(0.5)	0.5	0.52 [0.45, 0.63]	0.91
Pareto(3.5)	– (no finite Orlicz norm)	0.11 [0.10, 0.11]	0.87

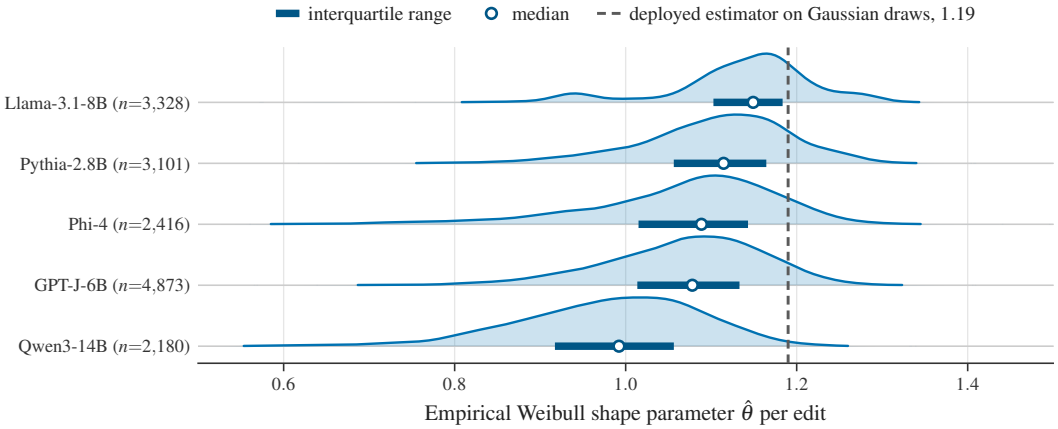

Figure 10: The empirical Weibull shape parameter $\hat{\theta}$ varies by architecture, and every architecture's median lies below the deployed estimator's value on Gaussian draws. Kernel density of per-edit $\hat{\theta}$ over the 15,898 canonical edits with a finite fit (Table 9), one row per architecture ordered by median; bars span the interquartile range and circles mark the median. The dashed line is the deployed estimator on Gaussian draws (Table 14).

G Proofs for constrained placement and the matching lower bound

Proof of Theorem 5. Let U be an orthonormal basis of the retained space R, and set $H = U^\top \Sigma_\psi U \succ 0$, $a = U^\top k_\star$. Writing $v = Uz$ reduces the problem to minimizing $z^\top H z$ subject to $a^\top z = \gamma$. Stationarity gives $Hz = \lambda a$, and the constraint yields $\lambda = \gamma/(a^\top H^{-1} a)$. Since $A = UH^{-1}U^\top$, this is the stated formula; strict convexity on R proves uniqueness. □

Derivation of Proposition 4. Writing $v = Nz$ for an orthonormal basis of $\ker K^\top$ reduces the cost to $z^\top (N^\top G N) z$. The preceding quadratic argument gives $Q_G = N(N^\top G N)^{-1} N^\top$. Eliminating the constraint multiplier gives the displayed inverse-minus-correction form of the same operator. The range of Q_G is $\ker K^\top$. Equality with $NN^\top$ requires $(N^\top G N)^{-1} = I$. Equality of all normalized updates requires every nonzero coordinate vector $N^\top k$ to be an eigenvector of $(N^\top G N)^{-1}$, which is equivalent to $N^\top G N = cI$. The converse follows by cancellation of c^{-1} in numerator and denominator. □

G.1 A MATCHING LOWER BOUND ON THE SUB-WEIBULL MAXIMUM

Fix the Orlicz ball $\{Y : \|Y\|_{\psi_\theta} \le K\}$. A threshold that bounds the maximum of M iid copies of Y with probability above $1/2$ for every Y in this ball is at least $K(\log M)^{1/\theta}$. For each M, a two-point law with one rare large atom reaches this threshold with probability $1/2$.

Theorem 6 (Worst-case lower bound on the Orlicz ball)**.** *For any $\theta > 0$, $K > 0$, and integer $M \ge 2$, there exists a random variable Y with $\|Y\|_{\psi_\theta} \le K$ (eq. (1)) such that iid samples $Y_1, \dots, Y_M \sim Y$ satisfy $\Pr[\max_{i \le M} |Y_i| \ge K(\log M)^{1/\theta}] \ge \frac{1}{2}$.*

For $M \ge 3$ and $\delta = 2/M$, the upper threshold $K(2 \log M)^{1/\theta}$ of Lemma 9 is $2^{1/\theta}$ times this lower bound. The remark after the proof compares the two at the matched confidence $1/2$.

Proof of Theorem 6. Let $q := 1 - 2^{-1/M} \in (0, 1)$ and $a := K\left(\log(1 + 1/q)\right)^{1/\theta}$. Define $Y = a \cdot \varepsilon$ with $\varepsilon \sim \mathrm{Bernoulli}(q)$. At scale K, the defining expectation of the Orlicz functional is

$$\mathbb{E}\, e^{(|Y|/K)^\theta} \;=\; (1 - q) + q\, e^{(a/K)^\theta} \;=\; (1 - q) + q\,(1 + 1/q) \;=\; 2,$$

so $\|Y\|_{\psi_\theta} = K$. For the max, $\Pr[\max_i |Y_i| < a] = (1 - q)^M = 2^{-1}$, hence $\Pr[\max_i |Y_i| \ge a] \ge 1/2$. It remains to check $a \ge K(\log M)^{1/\theta}$. Since $q = 1 - 2^{-1/M} \le (\log 2)/M$ by convexity of $x \mapsto e^x$, we have $1 + 1/q \ge 1 + M/\log 2 \ge M$, so $\log(1 + 1/q) \ge \log M$, giving $a \ge K(\log M)^{1/\theta}$. □

Remark (the two-point law is extremal). The Markov–Orlicz bound $(1 - p) + p \cdot e^{(a/K)^\theta} \le \mathbb{E}\, e^{(|Y|/K)^\theta} \le 2$ for $p := \Pr[|Y| \ge a]$ gives $a \le K(\log(1 + 1/p))^{1/\theta}$, with equality on $\mathrm{Bern}(q)$ at scale a. Setting $p = q$ (the smallest p with $(1 - p)^M \le 1/2$) yields the sharpest admissible $a = K(\log(1 + 1/q))^{1/\theta} = K(\log M)^{1/\theta}(1 + o(1))$. At failure probability $1/2$, the union bound of Lemma 9 gives $\max_i |Y_i| \le K(\log M + \log 4)^{1/\theta}$. The ratio of this upper threshold to the lower bound $K(\log M)^{1/\theta}$ is $(1 + \log 4/\log M)^{1/\theta}$, which tends to 1, so the rate $K(\log M)^{1/\theta}$ is sharp on the Orlicz ball.

Application to the edit response. If $v^\top k$ has the two-point law above and $Y_i = v^\top k_i$, then $|s(k_i)| = |\alpha(r^\top u)|\,|Y_i|$. Multiplying the lower bound on $\max_i |Y_i|$ by $|\alpha(r^\top u)|$ therefore bounds the maximum rank-one response $\max_{i \le M} |s(k_i)|$ from below.

G.2 CONSTRAINT PRESERVATION UNDER THE REGULARIZED SOLVE

The Σ_ψ-Proj+K_{pres} editor of Section E runs AlphaEdit's update with $Q = Q_{K_{\mathrm{pres}}}$ of eq. (3) in place of the null-space projector $P_\perp$. That update solves a regularized linear system. Lemma 7 shows that its solution satisfies the preservation constraints for every regularization strength $\lambda > 0$.

Lemma 7 (Preservation under the regularized solve)**.** *Write $\Delta = (\Delta W)^\top \in \mathbb{R}^{d_{\mathrm{in}} \times d_{\mathrm{out}}}$, so that Q and A act on the input dimension. For every $\lambda > 0$, the solution Δ of $(QA + \lambda I)\Delta = Qb$ preserves the constraint $K_{\mathrm{pres}}^\top \Delta = 0$, where $A = KK^\top$ and $b = KE^\top$ are AlphaEdit's standard objects for the residual E, and $Q = Q_{K_{\mathrm{pres}}}$.*

Proof of Lemma 7. Let $A_0 = \Sigma_\psi^{+,\tau}$, $C = A_0^{1/2} K_{\mathrm{pres}}$, and $P_C = C(C^\top C)^\dagger C^\top$. Then $Q = A_0^{1/2}(I - P_C)A_0^{1/2}$ is positive semidefinite and $K_{\mathrm{pres}}^\top Q = C^\top (I - P_C) A_0^{1/2} = 0$. For $A = KK^\top \succeq 0$, the nonzero eigenvalues of QA equal those of $Q^{1/2} A Q^{1/2} \succeq 0$, so $QA + \lambda I$ is invertible for $\lambda > 0$. Its solution satisfies $\lambda\Delta = Q(b - A\Delta)$ and hence lies in $\mathrm{range}(Q) \subseteq \ker K_{\mathrm{pres}}^\top$. □

H HEADROOM DECOMPOSITION: PROTOCOL AND FULL RESULTS

Probe banks. For every architecture–layer cell we collect a bank of 20,000 activations at the editor's key site, one token per article. Articles are spread evenly over the Wikipedia dump. Each is truncated to its first 128 tokens, the first two positions are dropped, and one remaining position is drawn uniformly. For the geometry measurements, the bank is centered by the mean of all 20,000 activations and then split at random into fit and evaluation halves of 10,000 articles each; the halves

are article-disjoint and share this centering. The archived contiguous cache holds tokens from about twenty consecutive articles, so any split of it either shares articles or leaves about ten per half. On GPT-J L_5 and Llama-3.1 L_6, the directional structure measured on that cache transfers to held-out probes when the halves share articles. When the cache is split at the article boundary, the median held-out ratios at the worst target found and between the two extreme directions fall below one. On the one-token-per-article bank the structure transfers across article-disjoint halves (Table 15).

Table 15: Directional structure transfers across article-disjoint halves of the one-token-per-article bank. With article-disjoint halves of the contiguous cache, the worst-target ratio and $\widehat{C}_M^{\text{eval}}$ fall below one. Medians at $M{=}128$, preservation-projected arm, 32-dimensional slices. The contiguous cache uses 12 targets per cell and is split either by rows at random, so the halves share articles, or at the article boundary. The one-token-per-article bank uses 40 targets per cell. The first two ratio columns give the held-out ratio H_M at the worst target found, on real slices and in the spherical reference. The last two give the ratio between the two extreme directions selected on the fit half, on held-out probes ($\widehat{C}_M^{\text{eval}}$) and on the fit half ($\widehat{C}_M$).

Cell	Probe split	Worst target	Spherical reference	$\widehat{C}_M^{\text{eval}}$	$\widehat{C}_M$
GPT-J L_5	rows at random	2.418	0.857	4.083	4.015
	article boundary	0.570	0.850	0.214	4.237
	one token per article	1.959	0.933	3.300	3.966
Llama-3.1 L_6	rows at random	2.429	0.897	4.211	4.246
	article boundary	0.553	0.907	0.219	4.890
	one token per article	2.185	0.934	3.721	4.072

Table 16: Held-out diagnostic excess grows with slice width at the real target and at the worst target found. The raw held-out real-target ratio changes by at most 0.006 across widths. Preservation-projected arm on the five cells measured at every width (GPT-J-6B L_5, Llama-3.1-8B L_6, Phi-4 L_8, Pythia-2.8B L_{10}, Qwen3-14B L_8), with 200, 200 and 60 cell–target measurements at widths 16, 32 and 64. Width 64 uses 12 targets per cell, three random targets per slice, four spherical-reference slices per arm and no ascent in the worst-target search. Raw ratios are real-target medians at $M{=}128$. Excess columns give the held-out diagnostic excess $100(D_M - 1)$, in percent, for the real target and the worst target found.

	Real, raw ratio		Real, excess (%)		Worst, excess (%)	
Width	held out	in sample	$M{=}128$	$M{=}512$	$M{=}128$	$M{=}512$
16	1.039	1.045	+5.1	+13.0	+97.5	+223.7
32	1.039	1.055	+6.2	+16.5	+135.6	+280.6
64	1.033	1.065	+8.0	+20.8	+164.1	+329.4

Table 17: Target-independent slices give larger median real-target ratios. Median held-out ratio H_M over 40 targets per cell, preservation-projected arm, 32-dimensional slices. The target-conditioned slice spans the target and the 31 leading principal directions of the fit half. The target-independent slice spans the 32 leading principal directions, and its run draws its own 40 targets from the same 400 CounterFact records. Table 18 pools all 17 cells.

		Real target		Worst target	
Cell	M	target-cond.	target-indep.	target-cond.	target-indep.
GPT-J L_3	8	1.0014	1.0084	1.320	1.319
	32	1.0074	1.0966	1.547	1.540
	128	1.0414	1.2368	2.585	2.550
	512	1.1080	1.3220	3.933	3.968
GPT-J L_5	8	1.0010	1.0065	1.163	1.188
	32	1.0194	1.0573	1.442	1.491
	128	1.0765	1.1375	1.959	1.913
	512	1.1274	1.2113	2.850	2.685

Slices and arms. For each target the 32-dimensional slice is spanned by the target and the 31 leading principal directions of the arm's fit-half probes (widths 16 and 64 keep 15 and 63). Probes

Table 18: Target-independent slices give larger median real-target ratios at every width and budget. For each architecture–layer cell we take the ratio of its target-independent to its target-conditioned median H_M on the preservation-projected arm, for the real target and for the worst target found. Entries are geometric means of this ratio over cells, with 95% intervals from 4,000 bootstrap resamples of cells. Width 32 pools all 17 cells, and widths 16 and 64 pool the five cells measured at every width; width 64 starts at M=32. Cells above one counts the cells whose real-target ratio exceeds one.

Width	M	Real target [95%]	Cells above one	Worst target [95%]
32	8	1.0067 [1.0059, 1.0077]	17/17	1.0083 [0.9943, 1.0275]
	32	1.0525 [1.0419, 1.0642]	17/17	1.0055 [0.9995, 1.0121]
	128	1.1089 [1.0741, 1.1447]	16/17	0.9991 [0.9882, 1.0090]
	512	1.1362 [1.0806, 1.2000]	16/17	0.9913 [0.9744, 1.0070]
16	8	1.0075 [1.0063, 1.0090]	5/5	1.0172 [0.9982, 1.0408]
	32	1.0576 [1.0418, 1.0735]	5/5	1.0078 [0.9948, 1.0206]
	128	1.1379 [1.0748, 1.2028]	5/5	0.9941 [0.9428, 1.0284]
	512	1.1859 [1.0703, 1.3104]	5/5	1.0202 [0.9864, 1.0542]
64	32	1.0476 [1.0334, 1.0620]	5/5	0.9766 [0.9394, 1.0056]
	128	1.1043 [1.0624, 1.1480]	5/5	1.0216 [0.9758, 1.0680]
	512	1.1478 [1.0787, 1.2190]	5/5	1.0217 [1.0095, 1.0339]

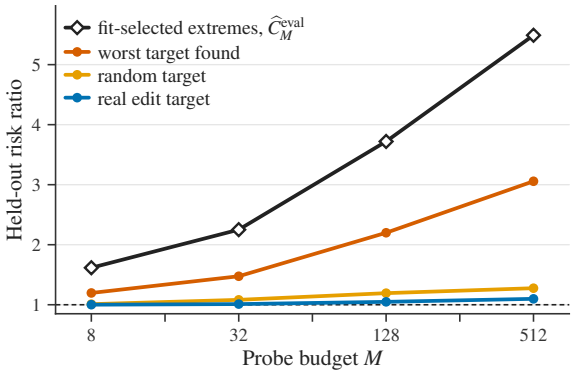

Figure 11: Every held-out median grows with the budget; the real-edit-target median grows least. Median held-out ratio H_M for each target class, and median $\widehat{C}_M^{\text{eval}}$, the held-out ratio between the two extreme directions selected on the fit half. Preservation-projected arm, 680 cell–target measurements in 17 cells, 32-dimensional slices, article-disjoint halves. Random-target medians pool six random targets per slice.

are mapped into the slice and whitened by the inverse square root of their fit-half second moment, so the fit half satisfies $N^{-1}X^\top X = I$. The arms differ only in what is projected off the probes and the target before the slice is built: the 100 preservation keys, nothing, 100 random Gaussian directions, the mean preservation key, or the 100 leading principal directions of the fit half. The last is the energy-maximal projection of that rank. On GPT-J L_5 its directions carry 22.0% of fit-half energy, and the preservation keys carry 2.6%. Two control targets use the preservation-projected geometry: a random ambient direction projected off the preservation keys, and a fit-half probe row, which the evaluation half does not contain. Using a fixed per-cell seed, every cell draws its own 40 CounterFact targets from the first 400 records and its own random directions. The principal-direction arm and the probe-row target come from a separate run with its own 40 targets per cell from the same 400 records (341 facts, 282 shared with the main run).

Target conditioning of the slice. The slices above are target-conditioned. A target-independent slice spans the leading principal directions (32 at width 32). For $v = Uz$, unit gain on the target is the linear constraint $(U^\top k_\star)^\top z = 1$, feasible whenever $U^\top k_\star \neq 0$ (Section J). We ran it on all 17 width-32 cells and, at widths 16 and 64, on the five cells measured at every width. Its runs draw their own targets from the same 400 CounterFact records, so the two slice types are compared cell by cell through their medians (Tables 17 and 18). The target-independent slice gives the larger real-target

median in 16 of the 17 cells at $M{=}128$, and at every budget in the two cells of Table 17. At the worst target found, every width-32 cell-bootstrap interval covers one.

Targets measured on each slice. The preservation-projected and no-projection arms evaluate the real target, six targets uniform on the whitened sphere, the worst target found by search, and $\widehat{C}_M$. The other arms evaluate their target and $\widehat{C}_M$. The unit directions $u_{\max}$ and $u_{\min}$ are the best of 32 projected-gradient restarts for the maximum and the minimum of the fit-half R_M on the unit sphere; $\widehat{C}_M = R_M(u_{\max})/R_M(u_{\min})$, and $\widehat{C}_M^{\text{eval}}$ is the same ratio on held-out probes. The search starts from the real and random targets and the six unit-normalized mixtures $\cos t\, u_{\max} \pm \sin t\, u_{\min}$ at $t \in \{30°, 45°, 60°\}$; for an ellipsoid the worst target is the $45°$ mixture. It then runs up to twelve steps of projected ascent on $\log \kappa_M$ from the two best starts. Write $R_M^*(a)$ for the constrained optimum $\min_{a^\top z=1} R_M(z)$, $z_{\text{opt}}(a)$ for its minimizer and $c(a)$ for the supporting cut the solver returns at a. The ascent direction at a unit target is $c(a)/R_M(a) + z_{\text{opt}}(a) - 2a$; it uses $\nabla \log R_M^*(a) = -z_{\text{opt}}(a)$, which holds where the minimizer is unique. The empirical R_M is polyhedral, so $c(a)$ is a subgradient and the ascent is a heuristic search. Candidates are ranked by the certified lower end of their ratio. The mixtures and ascent steps are solved to relative gap 10^{-3}, and the worst target found is re-solved at the reporting gap given below. The lower end of its ratio bounds the fit-half $\kappa_M = \sup_{a\neq 0} \kappa_M(a)$ from below. At width 64 no ascent is run, and the worst target found is the best start.

Worst-target floor. The floor in Table 1 is $\exp \Phi^{-1}(\log \widehat{C}_M)$. Since Φ is increasing and the searched $\widehat{C}_M$ is at most the fit-half C_M, the floor bounds the fit-half κ_M from below (Theorem 2). At the real-slice medians of Table 1 the infimum in Φ is attained at $N{=}3$ for $M{=}32$, giving 1.140 at $\widehat{C}_M{=}2.28$. For $M{=}128$ and 512 it is attained at $N{=}2$, where the bound reads $C_M \le 2\kappa_M^2$ and the floor is $\sqrt{\widehat{C}_M/2}$ (1.404 at $\widehat{C}_M{=}3.94$). Figure 12 compares every width-32 slice of the one-token-per-article bank with the floor from its own $\widehat{C}_M$. At each budget, the lower end of the ratio at the worst target found exceeds this floor in all 680 slices. On held-out probes the ratio at the worst target found also exceeds its slice's floor in all 680 slices at every budget, with smallest ratio to the floor 1.07 at $M{=}8$ and 1.48 at $M{=}512$. In the spherical reference it exceeds the floor in 9 of the 136 slices at $M{=}8$ and in none at $M{\ge}32$. Over the same range of $\widehat{C}_M$, the constant-factor bound $C_M \le 12\kappa_M + 3$ gives the floor $(\widehat{C}_M - 3)/12$, which lies below one.

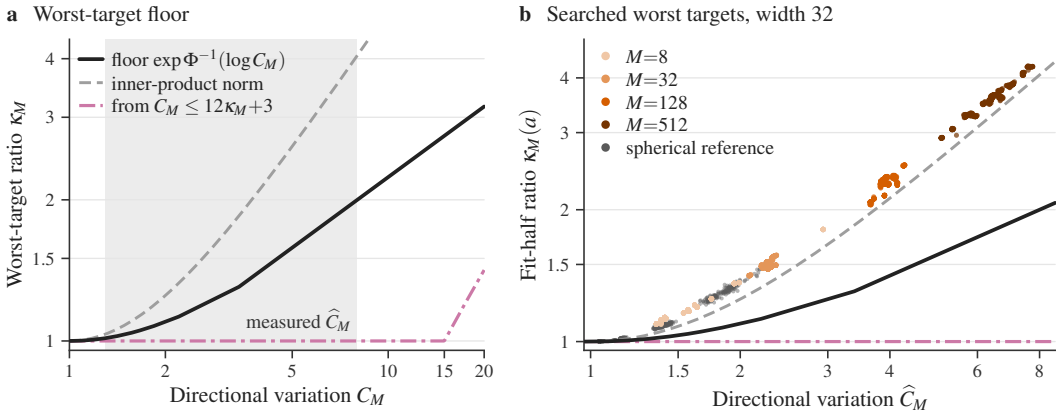

Figure 12: Over the measured range of $\widehat{C}_M$, the floor of the quantitative converse exceeds one. Every searched worst target lies above its floor. a The floor $\exp \Phi^{-1}(\log C_M)$ of Theorem 2, the worst-target ratio $(C_M + C_M^{-1})/2$ of an inner-product norm, and the floor implied by $C_M \le 12\kappa_M + 3$ (Khurana, 2024); shading marks the range of fit-half $\widehat{C}_M$ on real activations. **b** Certified lower end of the fit-half ratio at the searched worst target against $\widehat{C}_M$ for the 680 width-32 slices, by budget, and for the spherical-reference slices (grey); curves as in **a**.

Solver and certificate. On a sample $X \in \mathbb{R}^{N\times r}$, the expected maximum of M draws with replacement is exactly $R_M(z) = \sum_j w_j |Xz|_{(j)}$, where $|Xz|_{(1)} \le \cdots \le |Xz|_{(N)}$ are the sorted

absolute responses and the weights $w_j = (j^M - (j-1)^M)/N^M$ are non-decreasing and sum to one. R_M is the support function of $\mathrm{conv}\{X^\top(\epsilon \odot Pw)\}$ over sign vectors ϵ and permutations P, and the maximizing vertex at z is read off the order and signs of Xz. Hence $\min_{a^\top z=1} R_M(z)$ is the linear program $\min t$ subject to $a^\top z = 1$ and $c^\top z \le t$ for every vertex c. We solve it by cutting planes with in-out separation at the point $0.4\, z_{\mathrm{rel}} + 0.6\, z_{\mathrm{best}}$ between the restricted solution z_{rel} and the incumbent z_{best}. Each solve is warm-started from the cuts active in the previous solve on the same sample. Every restricted linear program is a relaxation over the box $\|z\|_\infty \le 1.01 \max_i \|x_i\|_2^2$, which contains the optimum: isotropy gives $R_M(z) \ge N^{-1} \sum_i |x_i^\top z| \ge \|z\|_2 / \max_i \|x_i\|_2$, and for a unit target $R_M(z_{\mathrm{opt}}) \le R_M(a) \le \max_i \|x_i\|_2$. The relaxation value is therefore a valid lower bound and the incumbent a valid upper bound. Every ratio reported at a target comes from a solve stopped at relative gap 2×10^{-5} (10^{-3} for the target-conditioned slices at width 64). The bracket $[\underline{t}, \bar{t}]$ on the optimum places the fit-half ratio at each target in a certified interval of that relative width, $[R_M(z_\Sigma)/\bar{t},\ R_M(z_\Sigma)/\underline{t}]$.

Raw ratios and spherical-reference diagnostics. For each target, the held-out risk ratio $H_M(a) = R_M^{\mathrm{eval}}(z_\Sigma)/R_M^{\mathrm{eval}}(\hat{z}_{\mathrm{fit}})$ scores covariance placement against the fit-half minimizer $\hat{z}_{\mathrm{fit}}$ on the same evaluation probes. In all 17 cells at M=128, the median held-out ratio rises from real edit target to random target to worst target found (Figure 13).

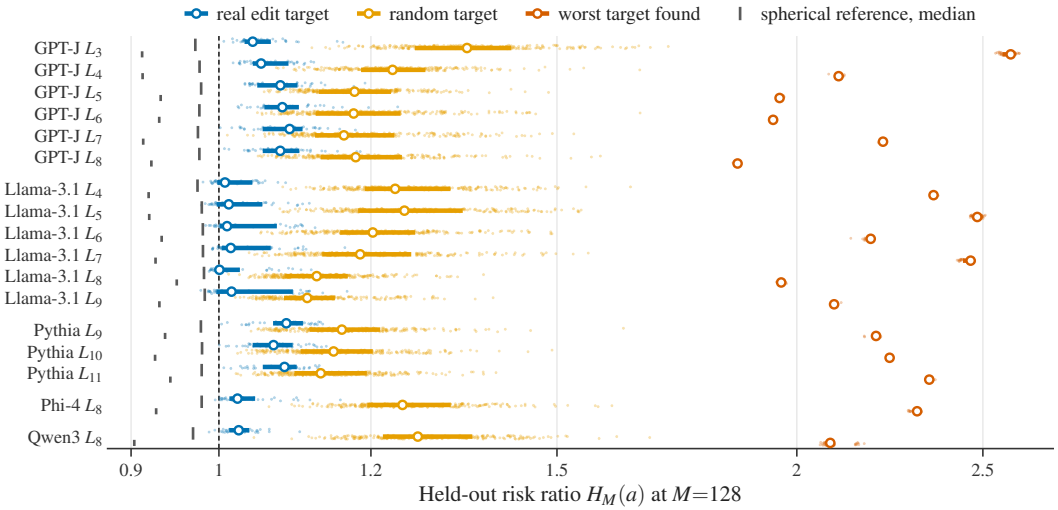

Figure 13: In every cell, the median held-out ratio rises from real edit target to random target to worst target found. Per-target $H_M(a)$ at M=128 on the preservation-projected 32-dimensional slices: 40 real edit targets, 240 random targets and 40 worst targets found per cell, with interquartile bars and medians (open circles). Grey ticks mark each cell's spherical-reference medians: its real and random slots (upper tick) and its worst targets found (lower tick), the pools that enter D_M.

Each arm of each cell has eight spherical-reference slices (four at width 64). Each spherical-reference slice starts from the slice built for the arm's first target, keeps each row's whitened radius and gives every row a uniform random direction. Its rows are then recentered across the full reference bank and whitened with the fit half. Random targets fill the real-target slot. At every budget the reference's median ratios exceed one on the fit half and fall below one on held-out probes (Table 1). Per target at M=128 (Figure 14a), fit-half ratios on real slices carry over to held-out probes, with Spearman correlation 0.93 for real and 0.94 for random targets. In the reference, fit-half gains become held-out losses for 99.6% of targets.

The *null-normalized diagnostic* D_M is the pooled median of H_M over a target class divided by the pooled median of the matching spherical-reference ratios. Diagnostic excess is $100(D_M - 1)$ in percent. For real and random targets the reference pool is the reference's random targets, including its real-target slot. For worst targets it is the reference's worst targets found. At width 32 the real-target diagnostic divides the median of 680 real ratios by the median of 952 reference ratios (17 cells, eight slices and seven random targets each). The in-sample diagnostic uses the corresponding fit-half ratios.

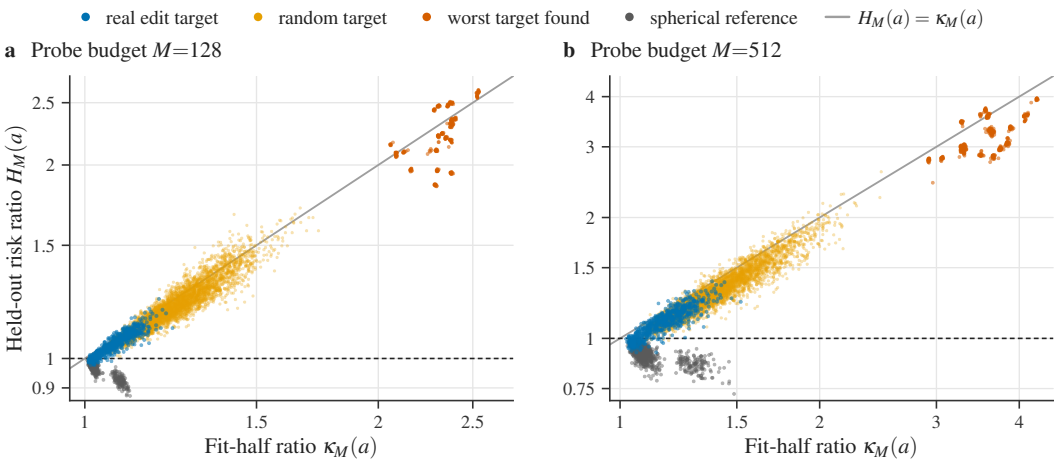

Figure 14: On real slices, fit-half gains carry over to held-out probes. In the spherical reference, they become held-out losses. Fit-half ratio $\kappa_M(a)$, the midpoint of its certified interval, against held-out $H_M(a)$ for every target on the preservation-projected 32-dimensional slices of the 17 cells, and for the real, random and worst-target slots of the 136 spherical-reference slices, at **a** M=128 and **b** M=512. Solid line: $H_M(a) = \kappa_M(a)$; dashed line: $H_M = 1$.

Table 19: Held-out diagnostic excess rises from M=128 to M=512 in every architecture. Diagnostic excess $100(D_M - 1)$, in percent, for the real target and the worst target found on the preservation-projected arm, with 40 targets per cell and 32-dimensional slices. Brackets are 95% intervals from 4,000 resamples of targets and reference records within each cell.

Architecture / layers	M	Real, held out	Real, in sample	Worst, held out
GPT-J (L_3–L_8)	128	+9.5 [8.7, 10.3]	+6.4 [5.8, 6.9]	+120.1 [119.2, 122.2]
	512	+22.6 [21.4, 23.6]	+11.0 [9.5, 12.1]	+247.2 [242.6, 253.4]
Llama-3.1 (L_4–L_9)	128	+2.9 [2.5, 3.4]	+1.1 [0.7, 1.6]	+144.5 [142.7, 145.4]
	512	+8.2 [7.1, 10.0]	+1.4 [0.5, 3.4]	+286.2 [283.2, 291.5]
Pythia (L_9–L_{11})	128	+10.2 [9.5, 10.8]	+8.0 [7.4, 8.6]	+139.1 [137.0, 139.9]
	512	+21.4 [19.5, 23.9]	+13.9 [12.2, 15.7]	+234.9 [232.0, 245.4]
Phi-4 (L_8)	128	+4.4 [3.7, 5.5]	+1.8 [1.1, 2.9]	+149.1 [146.4, 154.7]
	512	+13.0 [10.6, 15.1]	+2.8 [1.4, 5.8]	+327.7 [311.0, 344.9]
Qwen3-14B (L_8)	128	+5.6 [4.9, 6.0]	+1.8 [1.1, 2.5]	+130.4 [127.8, 133.7]
	512	+17.9 [14.4, 20.8]	−1.8 [−3.1, 0.2]	+306.9 [287.2, 323.3]
all 17	128	+7.2 [6.6, 7.6]	+4.8 [4.4, 5.2]	+137.4 [136.8, 138.5]
	512	+17.8 [17.2, 18.6]	+8.2 [7.4, 8.9]	+252.1 [249.4, 254.6]

The 95% cell-bootstrap intervals for pooled diagnostic excess in Figure 2(b) and Table 20 come from 1,000 resamples of the 17 cells. Each resample keeps a selected cell's real and reference pools together and then takes their separate medians. The 17 cells are six GPT-J layers, six Llama-3.1 layers, three Pythia layers and one layer each of Phi-4 and Qwen3-14B. Resampling the five architectures widens the real-target intervals (Table 20).

Table 20: Architecture-bootstrap intervals are wider than cell-bootstrap intervals. Held-out diagnostic excess $100(D_M - 1)$, in percent, for preservation-projected real targets, with 95% intervals that resample the 17 cells or the five architectures.

M	Diagnostic excess	Cell bootstrap	Architecture bootstrap
32	+1.7	[1.3, 2.1]	[1.0, 2.3]
128	+7.2	[5.4, 8.8]	[3.4, 9.6]
512	+17.8	[14.0, 20.4]	[9.8, 22.2]

Per-architecture values and slice width. The held-out diagnostic excess rises from $M{=}128$ to $M{=}512$ in every architecture (Table 19). Pooled over the five cells measured at every width, it grows with slice width for the real target and the worst target found (Table 16).

Budget-fitted placement on ROME edits. On GPT-J and Llama-3.1 we use the layer-5 ROME configuration and take 150 requests per model, drawn with the same fixed seed from the first 2,000 CounterFact records. For each request, $\bar{k}$ is the subject key of ROME (Meng et al., 2022) averaged over its context templates, and $v = C^{-1}\bar{k}/\|C^{-1}\bar{k}\|$ is ROME's input direction (its left vector), where C is ROME's uncentered second-moment estimate. The slice U spans $\bar{k}$ and the 31 leading principal directions of the uncentered fit-half probes of the one-token-per-article bank. It is whitened by Γ, the inverse square root of the probes' uncentered fit-half second moment, so the whitened rows $y_i = \Gamma U^\top x_i$, stacked in Y, are isotropic on the fit half. With $a = \Gamma U^\top \bar{k}$, $z_0 = a/\|a\|_2^2$ and $v_n = v/(\bar{k}^\top v)$, LP searches the affine family $\tilde{d}(z) = v_n + U\Gamma(z - z_0)$ with $a^\top z = 1$. Every member has $\bar{k}^\top \tilde{d}(z) = 1$, and $\tilde{d}(z_0) = v_n$ is ROME's direction. The fit-half response of $\tilde{d}(z)$ is $Yz + \eta$ with $\eta = Xv_n - Yz_0$, and LP takes the minimizer z^* of $R_M(Yz + \eta)$ at $M{=}M_{\text{fit}}$ subject to $a^\top z = 1$. This is again a linear program, generated by cuts $g_j^\top z + b_j \le t$ over the box $\|z\|_\infty \le 10^3 \max(1, \|z_0\|_\infty)$ and solved to the same gap (2×10^{-5}). In every edit run the box is inactive at the final restricted solve of each request. The value of that solve therefore bounds the unrestricted optimum from below, and the reported gap brackets the unrestricted problem.

Edit runs. Each run applies ROME twice per request, once with v and once with $\tilde{d}(z^*)/\|\tilde{d}(z^*)\|$ as the input direction, and alternates their order across requests. A run samples ROME's context templates once and keeps the model and the deterministic GPU settings fixed, so the two edits of a request differ only in their input direction. On each model, a replay run that applies ROME's direction in both edits reproduces every probability and gives factor ratios of exactly one on all 150 requests. Risk ratios are ROME / LP on the 10,000 held-out probes, summarized by geometric means with a bootstrap over requests. On the fit half of GPT-J, LP lowers the certified directional objective by median factors 1.04 at $M_{\text{fit}}{=}128$ and 1.10 at $M_{\text{fit}}{=}512$.

Squared response of LP and the quadratic refit. In the runs of Table 3, the quadratic refit has a fit-half mean squared response 0.58 to 0.70% below ROME's. On the fit half, LP's mean squared response exceeds the quadratic refit's by 22.1% for GPT-J at $M_{\text{fit}}{=}128$, 96.9% at $M_{\text{fit}}{=}512$ and 14.3% for Llama-3.1; on held-out probes, by 21.7%, 92.5% and 13.9% (geometric means over 150 requests). At every evaluated $M \ge M_{\text{fit}}$, LP's held-out expected-maximum risk is below the quadratic refit's (Table 3).

Normalization and measured endpoints. Let k_0 be the realized prompt key in ROME's update denominator and d a unit input direction; $\bar{k}$ is the averaged subject key defined above. Write b_d for the optimized target layer output minus the current output. In exact arithmetic, ROME (Meng et al., 2022) forms

$$q_d = \frac{b_d}{k_0^\top d}, \qquad \Delta W(d) = q_d d^\top,$$

$$V_M(d) = \mathbb{E} \max_{j \le M} \|\Delta W(d) x_j\|_2 = \frac{\|b_d\|_2 R_M(d)}{|k_0^\top d|}.$$

The directional comparison normalizes to unit gain on the averaged key, $\tilde{d} = d/(\bar{k}^\top d)$, and scores $R_M(\tilde{d}_{\text{ROME}})/R_M(\tilde{d}_{\text{LP}})$. With $g_d = (k_0^\top d)/(\bar{k}^\top d)$, homogeneity gives

$$\frac{V_M(d_{\text{ROME}})}{V_M(d_{\text{LP}})} = \frac{R_M(\tilde{d}_{\text{ROME}})}{R_M(\tilde{d}_{\text{LP}})} \frac{\|b_{\text{ROME}}\|_2}{\|b_{\text{LP}}\|_2} \left| \frac{g_{\text{LP}}}{g_{\text{ROME}}} \right|.$$

We call $\|b_d\|_2/|g_d|$ the output multiplier of d. Two directions with equal gain on $\bar{k}$ can have different output multipliers.

The *factor reconstruction* scores the layer-output norm risk $\|q_d\|_2 R_M(d)$ from the realized output-vector norm and the supplied input direction, before the model's precision cast. This score includes the realized output multiplier. The runs of Table 22 retained the realized $W_{\text{after}} - W_{\text{before}}$ for every request. The written-matrix score evaluates the same risk on these matrices, in float64 on the same

probes. In each of those runs it agrees with that run's factor reconstruction to within 4×10^{-5} in every geometric-mean ratio. In every setting the realized output multiplier changes each geometric-mean ROME / LP ratio by less than 0.3% (first two blocks of Table 2).

Behavior and article resampling. Table 21 counts successful prompts before and after the replacement, and Table 22 scores the written matrices with an interval over resampled evaluation articles.

Table 21: All 150 rewrites succeed with both directions; paraphrase and neighborhood counts change by at most 11 prompts. Successful prompts before and after replacing ROME's direction by LP on the 150 GPT-J L_5 requests of Table 2. The two fit budgets are separate runs, each with its own sampled ROME context templates. Their ROME baselines differ on 12 neighborhood prompts and one paraphrase prompt. In the fit-128 run, the median log-probability margin is 0.076 on those 12 prompts and 3.43 on all 1,500. The paraphrase prompt has the lowest margin of the 300 paraphrase prompts.

Endpoint	Prompts	Fit M=128, ROME → LP	Fit M=512, ROME → LP
Rewrite	150	150 → 150	150 → 150
Paraphrase	300	300 → 299	299 → 299
Neighborhood	1,500	1,203 → 1,210	1,201 → 1,212

Table 22: At every evaluated $M \geq M_{\text{fit}}$, the article-resampling interval of the written-matrix ROME / LP risk ratio lies above one. Directional is the geometric-mean ROME / LP risk ratio at unit gain on the averaged subject key, with a 95% interval from 4,000 bootstrap resamples of requests. Requests lower is the share of the 150 requests on which LP's directional risk is lower. Written matrix scores the matrices written to the model, with a 95% interval from 400 bootstrap resamples of the 10,000 evaluation articles, each applied to both directions and every request. Llama-3.1 directional and requests-lower entries come from the runs behind the unit-gain block of Table 2. All other entries come from the runs that retained the written matrices, which give that table's written-matrix block.

Cell, fit budget	Eval. M	Directional	Requests lower	Written matrix	Article interval
GPT-J L_5, fit 128	32	0.9781 [0.9763, 0.9798]	1.3%	0.9766	[0.9740, 0.9786]
	128	1.0159 [1.0139, 1.0178]	92.0%	1.0143	[1.0116, 1.0166]
	512	1.0408 [1.0380, 1.0436]	100.0%	1.0392	[1.0358, 1.0418]
	2,048	1.0493 [1.0455, 1.0531]	98.7%	1.0476	[1.0453, 1.0515]
GPT-J L_5, fit 512	32	0.8589 [0.8524, 0.8650]	0.0%	0.8569	[0.8507, 0.8615]
	128	0.9556 [0.9508, 0.9598]	4.0%	0.9534	[0.9471, 0.9586]
	512	1.0286 [1.0244, 1.0330]	88.0%	1.0262	[1.0186, 1.0307]
	2,048	1.0671 [1.0605, 1.0736]	96.0%	1.0646	[1.0577, 1.0693]
Llama-3.1 L_5, fit 128	32	0.9828 [0.9814, 0.9841]	0.7%	0.9843	[0.9811, 0.9874]
	128	1.0112 [1.0093, 1.0132]	83.3%	1.0120	[1.0074, 1.0169]
	512	1.0330 [1.0297, 1.0363]	98.7%	1.0326	[1.0266, 1.0391]
	2,048	1.0425 [1.0381, 1.0471]	100.0%	1.0406	[1.0336, 1.0507]

I SELF-CONTAINED PROOF OF THE SUB-WEIBULL TAIL BOUND

The proof is the classical Markov–Orlicz argument (Boucheron et al., 2013, Ch. 2) followed by a union bound, written here with every constant explicit. Its tail step is the standard exponential-Orlicz bound under the threshold-2 convention (Kuchibhotla & Chakrabortty, 2022, Definition 2.2 and Eq. (2.3)). Vladimirova et al. (2020, Thm. 2.1) show that a finite $\|Y\|_{\psi_\theta}$ is equivalent, up to constants, to a tail bound of the form $2\exp(-(t/K')^\theta)$ for some $K' > 0$. Their tail parameter is $1/\theta$ in our notation, so their sub-Gaussian case $1/2$ is our $\theta = 2$.

Definition (threshold-2 Orlicz functional). For $\theta > 0$ let $\psi_\theta(x) = \exp(x^\theta) - 1$ on $x \geq 0$ (so $\psi_\theta(0) = 0$, ψ_θ increasing). For a real random variable Y define

$$\|Y\|_{\psi_\theta} \;:=\; \inf\Big\{c > 0 :\; \mathbb{E}\,\exp\big((|Y|/c)^\theta\big) \leq 2\Big\}.$$

For $\theta \geq 1$ this is a norm. For $0 < \theta < 1$ it is a quasi-norm, whose triangle inequality holds up to a factor $2^{1/\theta-1}$. The proof uses only the defining inequality $\mathbb{E}\,\exp((|Y|/K)^\theta) \leq 2$ at $K = \|Y\|_{\psi_\theta}$, which holds for all $\theta > 0$.

Lemma 8 (Sub-Weibull tail from Orlicz finiteness)**.** *If $0 < K := \|Y\|_{\psi_\theta} < \infty$ then for every $t > 0$, $\Pr(|Y| > t) \leq 2\exp(-(t/K)^\theta)$. If $K = 0$, then $Y = 0$ almost surely and its positive-threshold tail is zero.*

Proof. The map $x \mapsto \exp((x/K)^\theta)$ is non-negative and strictly increasing on $x \geq 0$, so $\{|Y| > t\} = \{\exp((|Y|/K)^\theta) > \exp((t/K)^\theta)\}$. Markov's inequality applied to the non-negative random variable $\exp((|Y|/K)^\theta)$ gives

$$\Pr(|Y| > t) \leq \frac{\mathbb{E}\,\exp((|Y|/K)^\theta)}{\exp((t/K)^\theta)} \leq \frac{2}{\exp((t/K)^\theta)} = 2\exp(-(t/K)^\theta),$$

using $\mathbb{E}\,\exp((|Y|/K)^\theta) \leq 2$ at $K = \|Y\|_{\psi_\theta}$. □

Lemma 9 (Maximum over M probes)**.** *Let $k_1, \ldots, k_M$ be probe keys, possibly dependent, and put $Y_i := v^\top k_i$ with $\|Y_i\|_{\psi_\theta} \leq K$ for all i, where $K > 0$ and $M \geq 1$. Then for every $t > 0$, $\Pr(\max_{i \leq M} |Y_i| > t) \leq 2M\exp(-(t/K)^\theta)$, and setting the right-hand side to $\delta \in (0, 1)$ yields the deterministic threshold $t = K(\log(2M/\delta))^{1/\theta}$, which is $K(2\log M)^{1/\theta}$ at $\delta = 2/M$ (admissible for $M \geq 3$).*

Proof. A union bound over Lemma 8 gives $\Pr(\max_i |Y_i| > t) \leq \sum_{i=1}^M \Pr(|Y_i| > t) \leq 2M\exp(-(t/K)^\theta)$, which is valid under arbitrary dependence among the Y_i. Solving $2M\exp(-(t/K)^\theta) = \delta$ gives $(t/K)^\theta = \log(2M/\delta)$, which is $\log(M^2) = 2\log M$ at $\delta = 2/M$. □

Proof of Theorem 3. By linearity of $\Delta W = \alpha\, uv^\top$ and of the readout, the per-probe response of a rank-one edit (α, u, v) read out along r is $s(k_i) = \alpha(r^\top u)(v^\top k_i)$. Hence $|s(k_i)| = |\alpha(r^\top u)|\,|v^\top k_i|$, and $\max_{i \leq M} |s(k_i)| = |\alpha(r^\top u)| \max_{i \leq M} |v^\top k_i|$. By (A1) and (A2), every $v^\top k_i$ satisfies $\|v^\top k_i\|_{\psi_\theta} \leq K$. If $K = 0$, every response vanishes almost surely (Lemma 8). Otherwise Lemma 9 gives, with probability $\geq 1 - \delta$,

$$\max_{i \leq M} |s(k_i)| \;\leq\; |\alpha(r^\top u)|\, K\, (\log(2M/\delta))^{1/\theta},$$

which is the statement of Theorem 3. □

J Covariance optimality across targets and probe budgets

This appendix proves Theorems 1 and 2. Fix a population probe law and linear editing constraints. Section J.2 characterizes when the covariance direction minimizes the expected maximum absolute response for every feasible target and every probe budget. Section J.3 ties the worst-target risk ratio to directional anisotropy at each fixed budget. Section J.4 gives bounded counterexamples, and Section J.5 explains why the characterization uses all budgets and iid probes. The characterization is the expected-maximum analogue of elliptical mean–variance characterizations (Chamberlain, 1983). Expected maxima of iid copies form the MAXVAR risk family (Sun & Yao, 2018). The converse combines two classical facts. A norm in which Euclidean orthogonality implies Birkhoff–James orthogonality is a multiple of the Euclidean norm (Koldobsky, 1993; Blanco & Turnšek, 2006). An integrable law is determined by its sequence of expected maxima (Hoeffding, 1953).

J.1 Reduction to the feasible whitened subspace

Let $k \in \mathbb{R}^d$ satisfy $\mathbb{E}k = 0$ and have finite population covariance Σ. Fix a retained subspace $\mathcal{R}$, preservation keys K_{pres}, a target key k_*, and a prescribed gain $\gamma \neq 0$. Define

$$\mathcal{S} = \mathcal{R} \cap \ker K_{\mathrm{pres}}^\top, \qquad p = \dim \mathcal{S} \geq 1.$$

Table 23: Four whitened probe laws and the optimality of covariance placement. Each row is the law of the probe X after projection onto the feasible subspace and population whitening (covariance I_2). Probes are iid copies of X, projection tails are those of $Y = u^\top X$ for unit u, and the symmetrized law is that of εX for an independent uniform sign ε. Optimal means $a/\|a\|_2^2$ minimizes $\mathbb{E}\max_{i\le M}|z^\top X_i|$ subject to $a^\top z = 1$ for every $a \neq 0$ and $M \ge 1$. The Student law is $X = \sqrt{(\nu-2)/V}\,G$ with independent $G \sim \mathcal{N}(0, I_2)$ and $V \sim \chi^2_\nu$, the axis law is uniform, and in the angular row Θ has density $(1+\beta\cos 3\phi)/(2\pi)$ with $0 < |\beta| < 1$.

Whitened law	Projection tails	Symmetrized law spherical?	Optimal for all targets and budgets?
Standard Gaussian $\mathcal{N}(0, I_2)$	Gaussian; finite ψ_2 scale.	Yes	Yes
Standardized spherical Student, $\nu > 2$	Polynomial, $\Pr(\lvert Y\rvert > t) \asymp t^{-\nu}$; infinite ψ_θ scale for every $\theta > 0$.	Yes	Yes
Four-point axis law $\{\pm\sqrt{2}e_1, \pm\sqrt{2}e_2\}$	Bounded; finite ψ_θ scale for every $\theta > 0$.	No	No
Asymmetric angular law $\sqrt{2}(\cos\Theta, \sin\Theta)$	Bounded; finite ψ_θ scale for every $\theta > 0$.	Yes; original law is not spherical	Yes

Let $B \in \mathbb{R}^{d\times p}$ have orthonormal columns spanning $\mathcal{S}$, and assume $H = B^\top \Sigma B \succ 0$. Put

$$L = BH^{-1/2}, \qquad X = L^\top k, \qquad a = L^\top k_*. \tag{4}$$

Then $\mathrm{Cov}(X) = I_p$, every $v \in \mathcal{S}$ has a unique representation $v = Lz$, and

$$v^\top \Sigma v = \|z\|_2^2, \qquad v^\top k_* = a^\top z, \qquad v^\top k = z^\top X. \tag{5}$$

The target is feasible if and only if $a \neq 0$. Its covariance minimizer is

$$z_\Sigma = \frac{\gamma a}{\|a\|_2^2}, \qquad v_\Sigma = \frac{\gamma L L^\top k_*}{\|L^\top k_*\|_2^2}. \tag{6}$$

For iid copies $X_1, \dots, X_M$, define the population risk

$$R_M(z) = \mathbb{E}\max_{1\le i\le M} |z^\top X_i|. \tag{7}$$

For a rank-one update $\tilde{u}v^\top$ with fixed $\tilde{u}$, the expected maximum local activation displacement is $\|\tilde{u}\|_2 R_M(z)$. For the fixed readout r, the corresponding scalar risk is $|r^\top \tilde{u}| R_M(z)$. A nonzero fixed factor leaves the risk ordering of feasible directions unchanged. When $r^\top \tilde{u} = 0$, every feasible direction has zero scalar risk.

J.2 Exact characterization

A random vector is *spherical* if its law is invariant under every orthogonal transformation. Its *symmetrized law* is the law of εX, with ε an independent uniform sign.

Theorem 10 (Covariance sufficiency, full form of Theorem 1)**.** *Let $X \in \mathbb{R}^p$, $p \ge 1$, satisfy $\mathbb{E}\|X\|_2 < \infty$ and $\Pr(u^\top X \neq 0) > 0$ for every nonzero u. Use iid copies of X in eq. (7). The following are equivalent:*

1. *For every $a \neq 0$ and every integer $M \ge 1$, the vector $a/\|a\|_2^2$ minimizes $R_M(z)$ subject to $a^\top z = 1$.*
2. *For every $M \ge 1$, there is $c_M \in (0, \infty)$ such that $R_M(z) = c_M\|z\|_2$ for all z.*
3. *The symmetrized law εX is spherical.*

Under these conditions the minimizer is unique, and at gain $\gamma \neq 0$ it is $\gamma a/\|a\|_2^2$. If $X \overset{d}{=} -X$, item 3 is equivalent to sphericity of X itself.

Proof. For each M, R_M is an even, positively homogeneous norm. The triangle inequality holds samplewise, positivity follows from nondegenerate projections, and

$$R_M(z) \le M\mathbb{E}\|X\|_2 \, \|z\|_2.$$

The risk R_M is therefore finite and continuous.

Assume item 1 and fix M. For unit u, v with $c = u^\top v > 0$, v/c is feasible for target u. Applying optimality in both directions gives

$$c \le \frac{R_M(u)}{R_M(v)} \le c^{-1}, \qquad |\log R_M(u) - \log R_M(v)| \le -\log\cos\angle(u, v). \tag{8}$$

For $p \ge 2$, join any two unit vectors by a great-circle arc of length $\varphi \le \pi$ and divide it into N equal segments, with $\varphi/N < \pi/2$. Telescoping eq. (8) yields

$$|\log R_M(u) - \log R_M(v)| \le -N \log\cos(\varphi/N) \longrightarrow 0.$$

Thus R_M is constant on the unit sphere. For $p = 1$, the same conclusion follows from evenness. Homogeneity proves item 2.

This radial step is classical. In normed-space language, item 1 states that Euclidean orthogonality implies Birkhoff–James orthogonality in the norm R_M ($z \perp_B w$ whenever $z \neq 0$ and $w^\top z = 0$). Radiality therefore also follows from Precupanu (2013, Thm. 10, (i)⇔(iv)), whose argument applies to nonsmooth and nonrotund norms on spaces of any dimension. A second route is the orthogonality-preserver theorem of Koldobsky (1993), extended to linear maps between normed spaces by Blanco & Turnšek (2006), applied to the identity map $(\mathbb{R}^p, \|\cdot\|_2) \to (\mathbb{R}^p, R_M)$.

The next step is Hoeffding's classical identification of an integrable law by its expected maxima (Hoeffding, 1953), applied to every absolute projection. For unit u, let $Q_u \in L^1(0, 1)$ denote the quantile function of $|u^\top X|$. The iid order-statistic identity is

$$\frac{R_M(u)}{M} = \int_0^1 Q_u(t) t^{M-1} \, dt. \tag{9}$$

To verify it, transform iid uniform random variables by Q_u and use the density Mt^{M-1} of their maximum. Under item 2, for any unit u, v, all polynomial moments of the finite signed measure $(Q_u - Q_v)\, dt$ vanish. Density of polynomials in $C[0, 1]$ implies that this measure is zero, hence $Q_u = Q_v$ almost everywhere. Thus all unit absolute projections have the same law.

Set $Y = \varepsilon X$. Its one-dimensional projections are symmetric, so their laws are determined by their absolute values. Consequently, the characteristic function of Y depends only on the Euclidean norm of its argument. By uniqueness of characteristic functions, the law of Y is invariant under every orthogonal transformation, which proves item 3. This passage from projections to the vector is the Cramér–Wold step used in Molchanov & Turin (2021, Prop. 5.7), which their Ex. 5.8 applies to expected maxima.

Conversely, item 3 implies $|z^\top X| \overset{d}{=} \|z\|_2 |e_1^\top X|$. Independence across probes then gives item 2, with $c_M = R_M(e_1) > 0$. Cauchy–Schwarz shows that $a^\top z = 1$ has the unique minimum-Euclidean-norm point $a/\|a\|_2^2$. This proves item 1 and uniqueness. Scaling proves the assertion for general gain. If $X \overset{d}{=} -X$, then $\varepsilon X \overset{d}{=} X$, which proves the last assertion. □

Under eq. (4), finite covariance guarantees the integrability and nondegeneracy assumptions. With the bijection $v = Lz$ and the full column rank of L, it follows that v_Σ minimizes the expected maximum absolute response for every feasible target and every number M of iid probes if and only if the symmetrized law of $L^\top k$ is spherical. The hypotheses of Theorem 10 are integrability and nondegenerate projections. Uncentered probes whitened by the second-moment matrix $\Sigma_\psi = \mathbb{E}\, kk^\top$ meet them, and the same equivalence then holds for second-moment placement. Centered elliptical probes with finite covariance satisfy the sphericity condition, including Student distributions with more than two degrees of freedom. For these laws $R_M(z) = c_M\|z\|_2$, and the tail weight enters the risk through the constant c_M. If the symmetrized law is nonspherical, then for some target and some finite budget a feasible direction has lower risk than v_Σ.

Verification of the entries. For the Student representation, $\mathbb{E}[(\nu-2)/V] = 1$ gives covariance I_2, and a direction-independent Gaussian scale mixture is spherical. Polynomial projection tails make the ψ_θ scale infinite for every $\theta > 0$. For the angular law, Fourier orthogonality gives zero mean and covariance I_2. Random sign multiplication cancels the odd harmonic and produces the uniform law on the circle. The first, second and fourth optimality entries therefore follow from Theorem 10. The third follows from the explicit target $a = (2, 1)$ and minimizers in eq. (14). In the bounded rows, $|Y| \le \sqrt{2}$ for every unit u, so the ψ_θ scale is finite for every $\theta > 0$.

J.3 Approximate optimality and directional anisotropy

Theorem 11 (Quantitative converse, full form of Theorem 2)**.** *Let $R : \mathbb{R}^p \to [0, \infty)$ be continuous, even, positively homogeneous and positive off the origin; every norm qualifies. Define*

$$\kappa_R = \sup_{\|u\|_2=1} \kappa(u), \qquad \kappa(u) = \frac{R(u)}{\inf_{u^\top z=1} R(z)}, \qquad C_R = \frac{\max_{\|u\|_2=1} R(u)}{\min_{\|u\|_2=1} R(u)}.$$

Write $\lambda = \log \kappa_R$ and

$$\Phi(\lambda) = \inf_{\substack{N\in\mathbb{N}\\ N\ge 2}} N\Big(\lambda - \log\cos\tfrac{\pi}{2N}\Big). \tag{10}$$

For $\lambda > 0$ the infimum is attained. At $\lambda = 0$, $\Phi(0) = 0$, approached as $N \to \infty$ but not attained by any finite integer N. Then

$$\kappa_R \le C_R \le \exp\Phi(\lambda) \le \exp\Big(\pi\sqrt{\lambda} + 2\lambda\Big), \tag{11}$$

and $\Phi(\lambda) = \frac{\pi}{\sqrt{2}}\sqrt{\lambda}\,(1 + o(1))$ as $\lambda \to 0$. This applies to $R = R_M$, population or empirical, at any single fixed budget. The empirical R_M is positive off the origin exactly when the probes span the space, which whitening on the fit half guarantees.

Proof. Let m and b be the minimum and maximum of R on the unit sphere; both are finite and positive. At a unit target u, every feasible z has $\|z\|_2 \ge 1$, so $R(z) \ge m$, and $R(u) \le b$. This proves $\kappa_R \le C_R$.

For $\kappa_R = 1$, the exact cosine argument in the preceding proof gives $C_R = 1$. Suppose $\kappa_R > 1$ and put $\lambda = \log \kappa_R$. Approximate optimality gives, for $u^\top v > 0$,

$$|\log R(u) - \log R(v)| \le \lambda - \log\cos\angle(u, v).$$

For $p = 1$, evenness gives $C_R = 1$. For $p \ge 2$, replacing one endpoint by its negative if needed gives a connecting arc of length $\varphi \le \pi/2$. Split that arc into N equal steps and telescope. Since $-\log\cos$ is increasing on $[0, \pi/2)$ and $\varphi/N \le \pi/(2N)$,

$$|\log R(u) - \log R(v)| \le N\Big(\lambda - \log\cos\tfrac{\varphi}{N}\Big) \le N\Big(\lambda - \log\cos\tfrac{\pi}{2N}\Big)$$

for every $N \ge 2$. Taking the infimum over N, with u and v the maximizing and minimizing directions of R on the unit sphere, gives $\log C_R \le \Phi(\lambda)$. For the closed form, set $\varphi = \pi/2$ and $N = \max\{2, \lceil\varphi/\sqrt{\lambda}\rceil\}$, so that the minimand is that of eq. (10). Every step is then at most $\pi/4$. On $[0, \pi/4]$, $-\log\cos t \le t^2$, because the derivative of $t^2 + \log\cos t$ is $2t - \tan t \ge 0$. The minimand is therefore at most $N\lambda + \varphi^2/N$, and

$$N\lambda + \varphi^2/N \le 2\varphi\sqrt{\lambda} + 2\lambda \le \pi\sqrt{\lambda} + 2\lambda.$$

The asymptotic follows from two bounds. Since $-\log\cos t \ge t^2/2$ on $[0, \pi/2)$, every admissible N gives $N\big(\lambda - \log\cos\frac{\pi}{2N}\big) \ge N\lambda + \frac{\pi^2}{8N} \ge \frac{\pi}{\sqrt{2}}\sqrt{\lambda}$ by the arithmetic–geometric mean inequality, so $\Phi(\lambda) \ge \frac{\pi}{\sqrt{2}}\sqrt{\lambda}$. Conversely, take $N_\lambda = \lceil\pi/(2\sqrt{2\lambda})\rceil$, which is admissible once $\lambda < \pi^2/8$. Since $-\log\cos t = t^2/2 + O(t^4)$, the minimand at N_λ is at most $N_\lambda\lambda + \pi^2/(8N_\lambda) + O(\lambda^{3/2})$. With $N_\lambda\lambda \le \frac{\pi}{2\sqrt{2}}\sqrt{\lambda} + \lambda$ and $\pi^2/(8N_\lambda) \le \frac{\pi}{2\sqrt{2}}\sqrt{\lambda}$, this gives $\Phi(\lambda) \le \frac{\pi}{\sqrt{2}}\sqrt{\lambda} + \lambda + O(\lambda^{3/2})$. Together the two bounds give $\Phi(\lambda) = \frac{\pi}{\sqrt{2}}\sqrt{\lambda}\,(1 + o(1))$. □

Inner-product norms show that the bound of Theorem 11 is order-sharp as $\kappa_R \to 1$. For $R(z) = \sqrt{z_1^2 + d^2 z_2^2}$, $d \geq 1$, one has $C_R = d$ and $\kappa_R = (d + d^{-1})/2$, because $\kappa(u)^2 = (u^\top Q u)(u^\top Q^{-1} u)$ for $Q = \operatorname{diag}(1, d^2)$, which is maximized at $u = (1, 1)/\sqrt{2}$. Hence $\log \kappa_R \sim (\log d)^2/2$ as $d \downarrow 1$, so $\log C_R \sim \sqrt{2\lambda}$. For inner-product norms the sharp constant is $C_R = \kappa_R + \sqrt{\kappa_R^2 - 1}$ (Zhang et al., 2014, Thm. 2.3), and the ellipsoid above attains it. Zhang et al. (2014) state it in Chmieliński's $\hat{\varepsilon}$, which converts by $1/\kappa_R = \sqrt{1 - \hat{\varepsilon}^2}$. The constant $\pi/\sqrt{2}$ of Φ is $\pi/2$ times that family's sharp constant $\sqrt{2}$.

In operator language, $\kappa_R \leq \kappa$ says that the identity map from Euclidean space to $(\mathbb{R}^p, R)$ sends orthogonal pairs to pairs that are approximately Birkhoff–James orthogonal in Dragomir's sense with $1 - \varepsilon = 1/\kappa$, because for unit u, $\min_{w \perp u} R(u + w) = \min_{u^\top z = 1} R(z) = R(u)/\kappa(u)$. For maps between arbitrary normed spaces under this hypothesis, Khurana (2024, arXiv v4, Lem. 2.6 and Eq. (2.1)) gives the constant-factor bound $C_R \leq 12\kappa_R + 3$. Taking $N = 2$ in eq. (10) gives $\exp \Phi(\lambda) \leq 2\kappa_R^2$, so $\exp \Phi(\lambda) < 12\kappa_R + 3$ whenever $\kappa_R < 6.24$. As $\kappa_R \to 1$, $\exp \Phi(\lambda)$ tends to 1 and $12\kappa_R + 3$ tends to 15. Stability of relation-preserving maps in finite-dimensional normed spaces is studied by Wójcik (2014). Bounds that tend to 1 hold in two other settings. Under Chmieliński's ε-orthogonality, $(1 - 8\varepsilon)\|T\|\|x\| \leq \|Tx\|$ (Mojškerc & Turnšek, 2010, Thm. 3.5, Rem. 3.1). Under Dragomir's condition, a similar bound holds when the codomain is uniformly smooth (Mojškerc & Turnšek, 2010, Prop. 3.10); see Chmieliński & Stypka (2025, Thm. 4.4, Rem. 4.1). The empirical risk R_M of a finite probe sample is a polyhedral norm and hence nonsmooth. Theorem 11 treats Dragomir's condition with a Euclidean domain and an arbitrary codomain gauge, including this polyhedral norm. Its bound is explicit and dimension-free.

For a direct improvement witness at a fixed budget, take unit u, v with $c = u^\top v > 0$. The candidate v/c is gain-feasible for target u, and

$$\frac{R_M(v/c)}{R_M(u)} = \frac{R_M(v)}{c R_M(u)}, \qquad \|v/c\|_2^2 = c^{-2}. \tag{12}$$

Hence v/c has strictly lower risk than the covariance solution u whenever $R_M(v) < c\, R_M(u)$, and its quadratic cost is c^{-2}.

J.4 Bounded counterexamples at fixed covariance

An exact budget transition. Let X be uniform on $\{\pm\sqrt{2}e_1, \pm\sqrt{2}e_2\}$. It is centered, centrally symmetric, bounded, and has covariance I_2. For $b_M = 2^{-M}$,

$$R_M(z) = \sqrt{2}\big((1 - b_M)\max(|z_1|, |z_2|) + b_M \min(|z_1|, |z_2|)\big), \tag{13}$$

because with probability 2^{-M} every probe lies on the axis of the smaller coordinate. At target $a = (2, 1)$, the covariance solution is $(2/5, 1/5)$, and the exact minimizer is

$$z_M^* = \begin{cases} (1/2, 0), & M = 1, \\ (1/3, 1/3), & M \geq 2. \end{cases} \tag{14}$$

To prove global optimality, replace any feasible z by $|z|/(2|z_1| + |z_2|)$. The denominator is at least one, so this operation cannot increase the risk and yields a nonnegative feasible point. Write such points as $(x, 1 - 2x)$, $0 \leq x \leq 1/2$. Up to the factor $\sqrt{2}$, the piecewise-affine risk has slope $-2 + 3b_M < 0$ for $x < 1/3$ and $1 - 3b_M$ for $x > 1/3$, proving eq. (14). Covariance risk equals $\sqrt{2}(2/5 - b_M/5)$. The relative improvement is $1/21$ at $M = 2$ and tends to $1/6$ as $M \to \infty$. For this bounded law, the optimal direction changes between $M = 1$ and $M = 2$.

An unbounded covariance approximation factor across budgets. For an integer $A \geq 3$, let

$$\Pr(X_1 = A) = \Pr(X_1 = -A) = \frac{1}{2A^2}, \qquad \Pr(X_1 = 0) = 1 - A^{-2},$$

and let X_2 be an independent uniform sign. This law is centered, centrally symmetric and bounded, with covariance I_2. At target $(1, 1)$, $z_\Sigma = (1/2, 1/2)$, and the feasible direction e_2 has $R_M(e_2) = 1$. For $M = A^2$, the event that some probe has a nonzero first coordinate has probability $1 - (1 -$

$A^{-2})^{A^2}$ and yields an absolute response at least $(A-1)/2$. Consequently,

$$\frac{R_M(z_\Sigma)}{\inf_{z_1+z_2=1} R_M(z)} \geq \frac{A-1}{2}\left[1-(1-A^{-2})^{A^2}\right] \geq \frac{A-1}{2}(1-e^{-1}) \longrightarrow \infty. \tag{15}$$

For arbitrary $M \geq 1$, the three possible response magnitudes give

$$R_M(z_\Sigma) = \frac{A+1}{2} - \left(1-\frac{1}{2A^2}\right)^M - \frac{A-2}{2}\left(1-\frac{1}{A^2}\right)^M. \tag{16}$$

Also $R_M(e_1) = A[1-(1-A^{-2})^M]$, so e_1 has lower risk than e_2 at $M=1$ and higher risk at large M. For isotropic Gaussian probes, which have the same covariance, z_Σ is optimal at every budget. Figure 15a traces the exact ratio of this spike family across budgets.

An unbounded approximation factor at a fixed budget. Fix any integer $M \geq 1$ and let $A \to \infty$ in the preceding family. The feasible direction e_1 satisfies

$$R_M(e_1) = A[1-(1-A^{-2})^M] \leq M/A.$$

The covariance response is $z_\Sigma^\top X = (X_1+X_2)/2$, and $|X_1+X_2| \geq 1$ almost surely for $A \geq 3$, so $R_M(z_\Sigma) \geq 1/2$. Consequently,

$$\frac{R_M(z_\Sigma)}{\inf_{z_1+z_2=1} R_M(z)} \geq \frac{A}{2M} \longrightarrow \infty.$$

Each law is centered, bounded and centrally symmetric with covariance I_2, and its support bound grows with A. At one prespecified budget, a constant-factor guarantee that holds uniformly over laws with covariance I_2 therefore requires a further restriction, such as a shared support or tail-scale bound or the directional symmetry of Theorem 1.

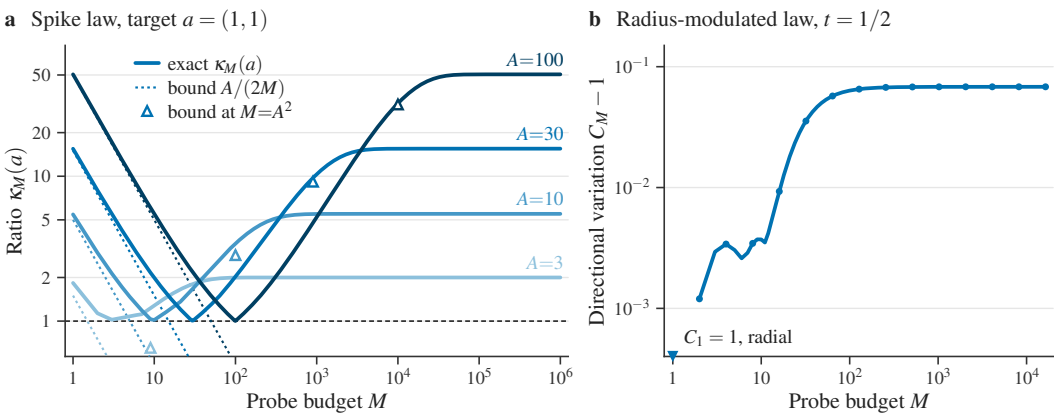

Figure 15: Bounded laws with identity covariance make covariance placement budget-dependent. a Exact $\kappa_M(a)$ at $a=(1,1)$ for the spike law of Section J.4 at four support scales A, with the bound $A/(2M)$ and the bound of eq. (15) at $M=A^2$. **b** Directional variation minus one, C_M-1, of the law of Section J.5 at $t=1/2$; R_1 is radial, so $C_1=1$.

J.5 Why the characterization uses all budgets and iid probes

Theorem 10 quantifies over every budget M and uses iid probes. The law below shows that its converse needs both. Let Θ be uniform on $[0,2\pi)$ and S an independent uniform sign. For $0<t<1$, set

$$r = 1+tS\cos 4\Theta, \qquad X = cr(\cos\Theta, \sin\Theta), \qquad c = (1/2+t^2/4)^{-1/2}.$$

The law is bounded, centrally symmetric and nonspherical, with covariance I_2. Its radius distribution depends on direction, and $\mathbb{E}[r \mid \Theta] = 1$, so $R_1(z) = (2c/\pi)\|z\|_2$. Thus a single budget does not identify sphericity. At $t=1/2$ the directional variation C_M exceeds one at every budget from M=2 to 16,384 (Figure 15b). For perfectly dependent probes $X_i = X$, $R_M = R_1$ for every M, so the risk is radial at every budget. The converse therefore needs iid probes.