When Weak Features Hide: Separating Statistical Detection from Sparse Autoencoder Recovery

Anonymous authors
Paper under double-blind review

Abstract

To understand weak features missed by sparse autoencoders (SAEs), we distinguish three questions: whether a statistical test detects a feature, whether a decoder can represent it, and whether training recovers it. We analyze covariance tests that refit coordinate-wise variances and derive a closed-form visibility factor that vanishes for features aligned with the detector's coordinate axes. Such signals can be absorbed into the fitted variances, creating a test-specific blind spot that does not restrict an unconstrained SAE decoder's representational capacity. We also derive a held-out alignment certificate under rank-one Gaussian covariance and recovery guarantees for decoder families that explicitly preserve a data-estimated anchor. In a prespecified synthetic control, an isotropic-null-calibrated covariance statistic has power 0.081 for an axis-aligned feature and 1.000 for a dense feature, while matched TopK SAEs recover both in all eight seeds. A post-hoc correlation-based sensitivity analysis yields the same qualitative contrast. A separate post-hoc experiment holds the architecture and non-anchor initialization fixed: training-PCA anchors yield $183/192$ recoveries versus $21/192$ for random anchors. The same comparison yields no recovery-rate advantage on the tested injected GPT-2/Pythia residual backgrounds. These results separate detector-specific invisibility from decoder feasibility and clarify the assumptions required for certified recovery.

1 Introduction

Sparse autoencoders (SAEs) can leave residual variance even as their width increases (Gao et al., 2025; Engels et al., 2025b). A weak feature can become hard to detect when a test refits coordinate-wise variances. We show that this blind spot arises in a specified covariance test with a fixed detector frame, even though an unconstrained dictionary can include an axis-aligned atom. A detector failure therefore neither limits what an unconstrained decoder can represent nor accounts for all residuals left by trained SAEs.

Superposition lets many features share a low-dimensional representation (Elhage et al., 2022; Scherlis et al., 2022). SAEs recover such features (Huben et al., 2024; Bricken et al., 2023; Gao et al., 2025), but residual variance, absorption, multidimensional features, and non-canonical decompositions remain challenges (Engels et al., 2025b; Chanin et al., 2025; Engels et al., 2025a; Leask et al., 2025). We ask which evidence a detector loses when it refits nuisance variance.

A detector class is defined by its statistic and the nuisance parameters it may refit. With labels, a mean shift gives the quadratic benchmark $N^\star \asymp \lambda^{-2}$; covariance testing against a fixed background gives the quartic benchmark $N^\star \asymp \lambda^{-4}$. Refitting coordinate variances in frame B further reduces leading-order evidence by $1 - \|B^\top \widehat{W}_i\|_4^4$. This coefficient belongs to the detector, not to a free decoder. Reconstruction has separate benchmarks: a linear encoder has quartic sample complexity uniformly over feature directions, whereas conditional isolation has a quadratic benchmark once a separating direction is available. Consistent with this distinction, rotating B changes independence power, while TopK recovery varies with signal strength in both geometries.

Our contributions are a local projection calculus for a specified detector frame, an exact decoder-feasibility result, a held-out alignment certificate for frozen subspaces, and recovery guarantees for exact and bounded-drift anchor-preserving families.

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.

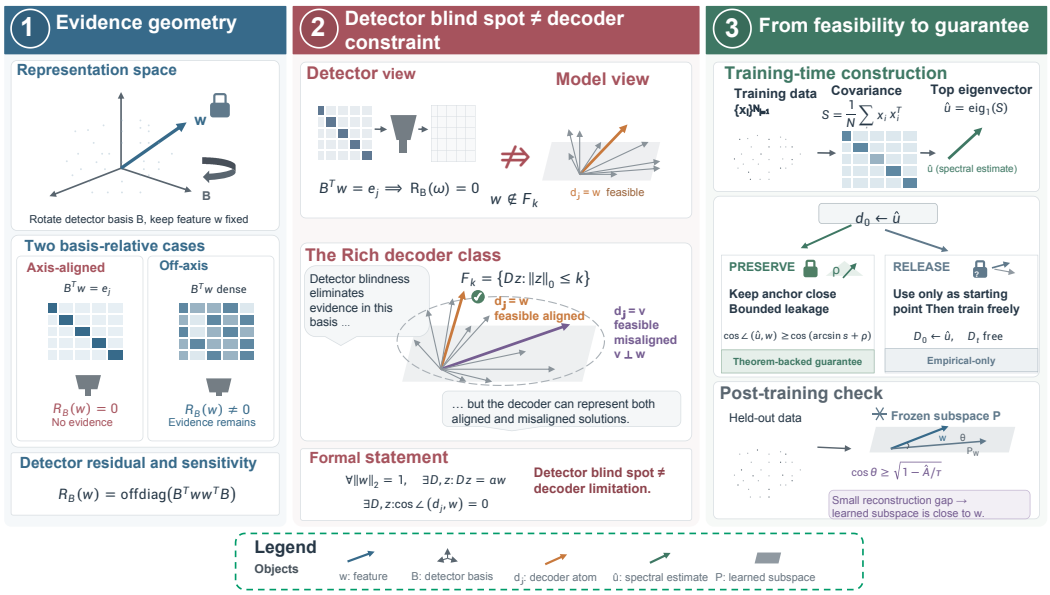


Figure 1: **Detector blind spots do not constrain decoder feasibility.** **(1)** Rotating the detector frame changes which covariance residuals remain visible for a fixed feature. **(2)** Detector invisibility does not prevent an unconstrained decoder from representing the feature; feasibility alone does not guarantee that training recovers it. **(3)** The recovery guarantee requires the stated anchor-preserving structure, while releasing that constraint is an empirical-only initialization. The panels summarize the paper's theory and experimental design; they are not additional measured results.

We test the detector–decoder distinction directly, then evaluate anchor-source effects in post-hoc matched Gaussian and injected-residual studies; results remain specific to their stated models. Figure 1 summarizes the framework.

2 Related Work

SAEs recover interpretable directions at scale, but residual floors, absorption, and non-canonical atoms persist (Huben et al., 2024; Bricken et al., 2023; Gao et al., 2025; Engels et al., 2025b; Chanin et al., 2025; Leask et al., 2025). Superposition and dictionary-learning work study overcomplete representations and global recovery (Elhage et al., 2022; Scherlis et al., 2022; Spielman et al., 2012; Arora et al., 2014); spiked-covariance theory studies spectral thresholds (Johnstone, 2001; Baik et al., 2005; Cai et al., 2013; Berthet & Rigollet, 2013). We instead quantify the evidence lost when a specified nuisance family refits diagonal variance, using local testing and projection (Le Cam, 1986; Amari & Nagaoka, 2000; Bickel et al., 1993).

3 Preliminaries

Residual variance and detector-limited benchmark. SAE scaling can lower reconstruction error without eliminating unexplained activation variance (Gao et al., 2025; Engels et al., 2025b). We distinguish this empirical residual from f_{dark}, the variance share assigned to features below a specified detector's threshold. We use f_{dark} only as an idealized recover-or-miss benchmark: it is neither a lower bound on trained-SAE error nor evidence that another decoder cannot recover those features.

Superposition model. Let n latent features share an m-dimensional representation with $n \gg m$. The linear encoding is $h = Wx + \sigma\varepsilon$, where $W = [W_1, \ldots, W_n] \in \mathbb{R}^{m \times n}$, $\varepsilon \sim \mathcal{N}(0, I_m)$, and $\sigma > 0$. Feature i activates with probability p_i along W_i. Let ν_i denote its centered second moment;

$\nu_i = p_i(1-p_i)$ and $\nu_i \approx p_i$ when rare. The representation covariance is

$$\Sigma_h = \sigma^2 I_m + \sum_{i=1}^{n} \nu_i W_i W_i^\top. \tag{1}$$

The scalar $\lambda_i := \sqrt{\nu_i}\,\|W_i\|$ is the effective strength of feature i. Under the idealized recover-or-miss rule, a missed feature contributes $\nu_i\|W_i\|^2$ to residual variance; this identity need not hold for a trained decoder that partially represents or absorbs the feature.

Feature recovery as testing. Fix a feature index i. Given N independent samples $h^{(1)}, \dots, h^{(N)}$, test

$$H_0^{(i)} :\ \Sigma_h = \sigma^2 I_m + \sum_{j\neq i} \nu_j W_j W_j^\top, \qquad H_1^{(i)} :\ \Sigma_h = \sigma^2 I_m + \sum_{j=1}^{n} \nu_j W_j W_j^\top. \tag{2}$$

Let N_i^* be the smallest sample size at which the specified test achieves vanishing error. With the activation indicator observed, the mean-shift benchmark is $N_i^* \asymp \lambda_i^{-2}$; covariance testing against a *fixed* background gives $N_i^* \asymp \lambda_i^{-4}$. A detector that refits coordinate variances can also absorb the spike's diagonal component. The remaining visibility depends on the feature loading in the detector frame. This nuisance choice does not assume that SAE features or activation coordinates are independent. Reconstruction is a separate task. Theorem 6 gives a quartic rate for a linear encoder and a conditional quadratic isolation benchmark; neither rate contains the detector's axis coefficient.

4 A LOCAL DETECTOR-CLASS VISIBILITY CALCULUS

4.1 LOCAL GAUSSIAN COVARIANCE SLICE

Write $\Sigma_h = \Sigma_{-i} + \nu_i W_i W_i^\top$, where Σ_{-i} contains all other features. For weak features, we use the local isotropy reduction $\Sigma_{-i} \approx \sigma^2_{\text{eff}} I_m$, with $\sigma^2_{\text{eff}} = \sigma^2 + m^{-1}\sum_{j\neq i}\nu_j\|W_j\|^2$. This approximates the background at the scale relevant to detection; it does not assert global isotropy. The resulting hypotheses are

$$H_0^{(i)} :\ h \sim \mathcal{N}(0, \sigma^2_{\text{eff}} I_m), \qquad H_1^{(i)} :\ h \sim \mathcal{N}\big(0, \sigma^2_{\text{eff}} I_m + \lambda_i^2 \widehat{W}_i \widehat{W}_i^\top\big), \tag{3}$$

where $\widehat{W}_i := W_i/\|W_i\|$.

4.2 PROJECTION CALCULUS

Proposition 1 (Detector-class projection)**.** *Let $\mathcal{M}_D$ be the regular background family that detector class D may refit, with score tangent space T_D at $P_0 = \mathcal{N}(0, s^2 I_m)$. In the local Gaussian covariance slice, for $P_\lambda = \mathcal{N}(0, s^2 I_m + \lambda^2 w w^\top)$ and $\psi_w(h) = ((w^\top h)^2 - s^2)/(2s^4)$,*

$$\inf_{Q\in\mathcal{M}_D} D_{\mathrm{KL}}(P_\lambda\|Q) = \frac{\lambda^4}{2}\big\|\psi_w - \Pi_{T_D}\psi_w\big\|^2_{L^2(P_0)} + O(\lambda^6). \tag{4}$$

Feature visibility is the squared norm of the feature score after removing the component explained by the detector's nuisance directions. The proof is in Appendix B.4.

4.3 FRAME-DIAGONAL VARIANCE DETECTOR

To isolate the effect of nuisance refitting, we define a detector that may refit coordinate variances but not correlations. In an orthonormal detector frame B, its local Gaussian family is

$$\mathcal{M}_{\mathrm{ind},B} := \big\{\mathcal{N}\big(0, B\,\mathrm{diag}(d_1,\dots,d_m)B^\top\big) : d_j > 0\big\}. \tag{5}$$

This is a specified detector class, not a claim that GPT-2 activations or trained SAEs factorize coordinate-wise. It can absorb variance inflation along the axes of B, but it cannot create correlations between those coordinates.

Which coordinates? All directions use isotropized analysis coordinates. In the synthetic rotation experiment, B changes the orthonormal detector frame. In the GPT-2 bridge, residuals are first PCA-whitened on the training split and the detector uses those whitened coordinates ($B = I$). This whitening map is generally non-orthonormal and is distinct from B. With $w_B := B^\top \widehat{W}_i$, axis-aligned means $w_B = e_j$ in the analysis frame, not alignment with a neuron or SAE atom. Rotating B at fixed strength changes $1 - \|w_B\|_4^4$ and independence power, but not label power.

Let $s^2 = \sigma_{\text{eff}}^2$, $w = \widehat{W}_i$, $w_B = B^\top w$, and $t = \lambda_i^2/s^2$. In detector coordinates, the best diagonal null matches the diagonal of the spiked covariance, so the trace term cancels and only the off-diagonal residual remains. Equivalently, $w_B w_B^\top = \mathrm{diag}(w_{B,1}^2, \ldots, w_{B,m}^2) + R_{w_B}$, with $\|R_{w_B}\|_F^2 = 1 - \|w_B\|_4^4$.

Theorem 1 (Quartic visibility after frame-diagonal refitting)**.** *Under the local isotropy reduction above, the minimum local KL divergence between the alternative in $H_1^{(i)}$ and $\mathcal{M}_{\mathrm{ind},B}$ satisfies*

$$D_{\mathrm{loc}}^{(i)}(\lambda_i) = C_i(B)\lambda_i^4 + O(\lambda_i^6), \qquad C_i(B) = \frac{1 - \|B^\top \widehat{W}_i\|_4^4}{4\sigma_{\text{eff}}^4}. \tag{6}$$

When $C_i(B) > 0$, the sample complexity is $N_i^* \asymp C_i(B)^{-1}\lambda_i^{-4}$. If $C_i(B) = 0$, the detector family absorbs the spike; for an exactly frame-aligned loading, the KL gap is identically zero.

Corollary 1 (Whitened form)**.** *For a general background $\Sigma_{-i} \succ 0$, let $\tilde{h} = \Sigma_{-i}^{-1/2} h$, $\tilde{w} = \Sigma_{-i}^{-1/2} W_i / \|\Sigma_{-i}^{-1/2} W_i\|$, and $\tilde{\lambda}_i := \sqrt{\nu_i}\|\Sigma_{-i}^{-1/2} W_i\|$. If the detector, after whitening, refits diagonal variances in an orthonormal frame B, then $D_{\mathrm{loc}}^{(i)} = (1 - \|B^\top \tilde{w}\|_4^4)\tilde{\lambda}_i^4/4 + O(\tilde{\lambda}_i^6)$. The exponent is unchanged, but strength and geometry are evaluated after whitening. This is generally a different nuisance family from one fixed as diagonal before a non-uniform whitening transform; for that case, use Proposition 1 with the transformed nuisance tangent space. Local isotropy is the special case $\Sigma_{-i} = \sigma_{\text{eff}}^2 I_m$, where $\tilde{\lambda}_i = \lambda_i/\sigma_{\text{eff}}$.*

A two-dimensional toy makes the geometry concrete. For $m = 2$, $w_B = (1, 0)$ has $C_B = 0$, whereas $w_B = (1/\sqrt{2}, 1/\sqrt{2})$ has $C_B = 1/2$. At the same λ and N, the first feature has $E_{\mathrm{ind}} = 0$ at this order and the second has $E_{\mathrm{ind}} \asymp N\lambda^4/2$. Equal strength is not equal evidence. Figure 1 summarizes the detector–decoder distinction and the separate recovery guarantee.

4.4 Detector-class hierarchy

Corollary 2 (Feature visibility principle)**.** *Up to constants and slowly varying thresholds,*

$$E_{\mathrm{label}}(i) \asymp N\lambda_i^2, \qquad E_{\mathrm{cov}}(i) \asymp N\lambda_i^4, \qquad E_{\mathrm{ind}}(i) \asymp N\big(1 - \|B^\top \widehat{W}_i\|_4^4\big)\lambda_i^4. \tag{7}$$

Corollary 3 (Detector-class hierarchy)**.** *In the weak-signal limit, $E_{\mathrm{label}}(i) \geq E_{\mathrm{cov}}(i) \geq E_{\mathrm{ind}}(i)$. If $B^\top \widehat{W}_i = e_j$, then $E_{\mathrm{ind}} = 0$ at leading quartic order, while labels and fixed-background covariance still see the feature.*

If the feature were absent, all three tests would fail; if its diagonal variance were not absorbed, covariance and independence tests would agree. The predicted pattern is a present feature that remains visible to label and covariance tests but becomes invisible to the independence detector. This split belongs to $\mathcal{M}_{\mathrm{ind},B}$, not to every sparse decoder.

Theorem 2 (The axis hole is not a free-decoder constraint)**.** *Let $\mathcal{F}_k = \{Dz : \|z\|_0 \leq k\}$ be a k-sparse linear decoder with unconstrained dictionary $D \in \mathbb{R}^{m \times d}$ and $d \geq 1$. For every unit w, including a frame-aligned w for which $C_i(B) = 0$, some D and 1-sparse z realize $Dz = \alpha w$. Thus the axis direction is available to the decoder class even when the detector's quartic coefficient vanishes; that coefficient is not a constraint on this model class. This is a representation statement, not a claim about optimization or about projecting probability models onto $\mathcal{F}_k$.*

Proposition 2 (Why a structural constraint is necessary)**.** *Let $m \geq 2$, $k \geq 1$, and let $\mathcal{F}_k$ be the unconstrained k-sparse decoder class above. Membership in $\mathcal{F}_k$ alone gives no positive uniform alignment bound: for every unit target w, the active atom of a feasible decoder can be any prescribed unit vector v, including one orthogonal to w. Thus the feasible class includes decoders with zero target alignment. This statement concerns the model class, not every learning procedure; it does*

not rule out a data-dependent algorithm that uses samples to select v. A positive recovery theorem therefore needs an explicit data–target linkage assumption, such as the training-only spectral anchor and anchor-preserving constraints used below.

Theorem 3 (Objective-agnostic reconstruction certificate)**.** *Let X be centered with covariance $\Sigma = \sigma^2 I_m + \tau w w^\top$, where $\tau > 0$ and $\|w\| = 1$. Let P be the orthogonal projector onto any rank-r decoder subspace, and let P_w be any rank-r projector whose range contains w. Define $R(P) = \mathbb{E}\|X - PX\|^2$. Then*

$$R(P) - R(P_w) = \tau\big(1 - \|Pw\|^2\big) = \tau \sin^2\angle(w, \operatorname{range} P). \tag{8}$$

Consequently, the population excess risk gives an alignment bound. The plug-in value $\widehat{\Delta}$ is an empirical diagnostic, not by itself a finite-sample certificate. If X is Gaussian and the independent validation set has n samples, then with probability at least $1 - \delta$, the stated concentration bound gives $\cos\angle(w, \operatorname{range} P) \ge \sqrt{1 - (\widehat{\Delta} + \epsilon_{n,r,\delta})/\tau}$ whenever $\widehat{\Delta} + \epsilon_{n,r,\delta} \le \tau$, where

$$\epsilon_{n,r,\delta} = 2(\sigma^2 + \tau)\left(\sqrt{\frac{4r\log(2/\delta)}{n}} + \frac{\log(2/\delta)}{n}\right). \tag{9}$$

The certificate is objective-agnostic but conditional on the rank, covariance model, target-specific oracle comparison, and independent validation split; it is not an optimization or unknown-direction discovery guarantee.

Theorem 4 (Objective/capacity-uniform spectral-anchor recovery)**.** *Let $X_1, \dots, X_N$ be independent centered Gaussian samples with covariance $\Sigma = \sigma^2 I_m + \tau w w^\top$, where $\tau > 0$ and $\|w\| = 1$. Let $\widehat{u}$ be the leading unit eigenvector of the training second moment $S = N^{-1}\sum_i X_i X_i^\top$, with its sign fixed by a deterministic coordinate convention. If $\|S - \Sigma\|_{\mathrm{op}} \le \epsilon < \tau/2$, then $\sin\angle(\widehat{u}, w) \le \epsilon/(\tau - \epsilon)$. Construct a signed anchor pair $\{\widehat{u}, -\widehat{u}\}$ and restrict every residual decoder column to $\widehat{u}^\perp$, reserving one active TopK slot for the signed anchor. Then, for all anchor-preserving objectives and capacities, and for every optimizer and initialization that preserves these constraints, the anchor recovery is identical and satisfies*

$$\cos\angle(\widehat{u}, w) \ge \sqrt{1 - \big(\epsilon/(\tau - \epsilon)\big)^2}.$$

With the Gaussian-Wishart bound $\epsilon = \|\Sigma\|_{\mathrm{op}}(2q + q^2)$, $q = (\sqrt{m} + \sqrt{2\log(2/\delta')})/\sqrt{N}$, and a family-wise allocation $\delta' = \delta/G$ over G training anchors, this holds simultaneously with probability at least $1 - \delta$. This is a theorem for the anchor-preserving family, not for unconstrained SAE training or unknown-direction discovery.

Theorem 5 (Bounded-drift spectral-anchor recovery)**.** *Let $\widehat{u}_0$ be the training-only leading-eigenvector anchor in Theorem 4, and let $\widehat{u}$ be a trainable unit anchor satisfying $\angle(\widehat{u}, \widehat{u}_0) \le \rho$. Allow each unit residual decoder column v_j to have bounded anchor leakage $|v_j^\top \widehat{u}| \le \eta < 1$. If the spectral estimate obeys $\sin\angle(\widehat{u}_0, w) \le s < 1$ and $\arcsin(s) + \rho < \pi/2$, then, for all bounded-drift anchor-preserving objectives and capacities, and for every optimizer and initialization that enforces these constraints,*

$$\cos\angle(\widehat{u}, w) \ge \cos\big(\arcsin(s) + \rho\big).$$

The residual leakage cap η may be nonzero and does not weaken this anchor bound. With the family-wise Gaussian-Wishart/Davis–Kahan choice $s = \epsilon/(\tau - \epsilon)$ and $\delta' = \delta/G$, the statement holds simultaneously with probability at least $1 - \delta$ over G training anchors.

The training objective is reconstruction, not a likelihood ratio against the specified detector family $\mathcal{M}_{\mathrm{ind},B}$. The next theorem gives a linear-encoder rate and a conditional isolation benchmark; it does not give an end-to-end convergence theorem for trained TopK. Write $h = \varepsilon + a(z - p)w$ with $\varepsilon \sim \mathcal{N}(0, I_m)$, $z \sim \mathrm{Bern}(p)$, $\lambda = a\sqrt{p}$, and $\beta = \lambda^2(1 - p)$.

Theorem 6 (Reconstruction two-regime law)**.** *Consider one-sparse reconstruction $\hat{h} = g(u^\top h)\, u$ with $\|u\| = 1$. (i) Linear encoder $g(t) = \alpha t$: the top eigenvector $\hat{u}$ of the sample covariance satisfies $\mathbb{E}[\sin^2\theta] = (m-1)(1+\beta)/(N\beta^2) + O(N^{-2})$ with $\theta = \angle(\hat{u}, w)$, hence $N^\star \asymp \lambda^{-4}$ uniformly in w. (ii) Conditional isolation benchmark: assuming a direction u that separates the mixture is already available and a threshold on $u^\top h$ isolates $\{z = 1\}$ as $a \to \infty$, the resulting difference-of-means estimator satisfies $\mathbb{E}[\cos\theta] \approx \kappa(\kappa^2 + m - 1)^{-1/2}$ with $\kappa = \lambda\sqrt{(1-p)N}$, hence the isolation stage*

has $N^\star \asymp \lambda^{-2}$ when both groups are sufficiently sampled ($Np \to \infty$ and $N(1-p) \to \infty$). If the positive group is rare and Np remains bounded, the probability of observing no positive examples imposes an additional sampling floor. This is not an end-to-end trained-SAE guarantee. Neither rate contains $1 - \|w\|_4^4$.

Proposition 3 (Oracle column is a residual mean)**.** *If code coordinate 0 is clamped to $z - p$ and the remaining dictionary D_{-0} is free, the population minimizer of decoder column 0 is proportional to $\mathbb{E}[(z-p)(h - D_{-0}z_{-0})]$. Its alignment with w is strictly less than one when the absorbed term has a nonzero component orthogonal to w and the residual mean is nonzero. Thus the oracle column is conditional on fitting the rest of the dictionary; its finite-N recovery cost cannot be interpreted without that stage.*

These results separate representation from learning: Theorem 2 concerns feasibility, Theorem 6 gives a linear-encoder rate and a conditional isolation benchmark, and Proposition 3 explains why clamping z is not the mean estimator. Gradient-descent recovery is evaluated empirically in Section 5.3.

The recover-or-miss calculation in Appendix A.3 is detector-specific, not a lower bound on decoder recovery. It equals residual variance only under the stipulated binary rule that supercritical features are recovered and subcritical ones are missed. In the 80-feature control, trained-TopK residuals vary with capacity and are poorly tracked by this benchmark.

5 Experiments

The primary E4-v2 control uses centered Bernoulli features ($p = 0.05$, $m = 8$), $\lambda_{\text{eff}} \in \{0.35, 0.55, 0.80\}$, $N = 4096$, and eight seeds. Appendix C reports the detector-trial counts, TopK settings, the 0.80 recovery threshold, the bootstrap protocol, and the separate E1 calibration.

Because frame rotation changes only the coordinates used by the independence statistic, label power remains 1; we use it as a protocol check.

5.1 Primary matched detector–decoder control

Figure 2 shows the pre-specified E4-v2 endpoint and the paired seed-by-strength audit.

5.2 Same strength, opposite sides of the boundary

At matched λ, label power is 1.000 for both geometries and covariance power is $0.969/0.974$ across the two geometries. Independence power is 0.096 $[0.090, 0.102]$ for axis loadings and 0.969 $[0.965, 0.974]$ for dense loadings. Under frame rotation, independence power tracks the predicted visibility coefficient ($r = 0.977$), while label power remains one. In the two-sparse control, power correlates more strongly with the visibility coefficient ($r = 0.972$) than with $\|u\|_1$ ($r = 0.808$). Appendix Figure 5 shows the full calibration.

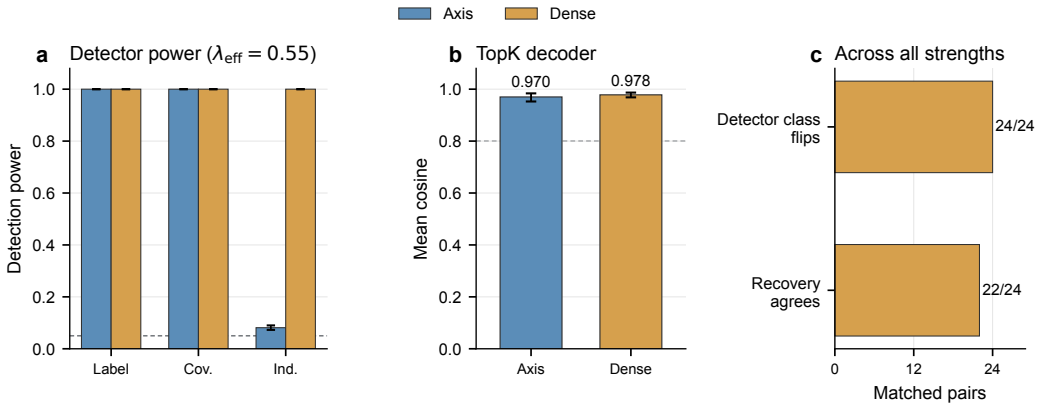


Figure 2: **A detector-class visibility split does not transfer to matched TopK recovery. a**, Pre-specified E4-v2 endpoint at $\lambda_{\text{eff}} = 0.55$: grouped bars compare label, fixed-background covariance, and feature-independence detector power for axis and dense loadings under isotropic-null calibration. Error bars are seed-level percentile-bootstrap 95% confidence intervals from 1,000 resamples; the dashed line marks the nominal 0.05 level for that calibration, not demonstrated size control over the full composite diagonal-covariance null. **b**, Mean decoder cosine from the same runs, with seed-level bootstrap 95% confidence intervals and the frozen 0.80 recovery threshold. All eight seeds recover both geometries (mean cosine 0.970 and 0.978 for axis and dense). **c**, Descriptive audit across all 24 matched seed–strength pairs: detector class flips in $24/24$ pairs and binary recovery agrees in $22/24$. The experiment uses $N = 4096$ detector samples per trial, 400 alternative trials, a 16-atom TopK decoder with $k = 2$, 20,000 training samples, and 4,000 optimization steps.

5.3 A DETECTOR SPLIT DOES NOT IMPLY A FREE-SAE RECOVERY GAP

Theorem 1 describes a specified detector nuisance family. Theorems 2 and 6 instead address decoder feasibility and reconstruction. In the pre-specified E4-v2 control ($p = 0.05$, $\lambda_{\text{eff}} = 0.55$, $N = 4096$), label and fixed-background covariance power are 1.000 for both geometries. Under the original isotropic-null calibration, independence power is 0.081 $[0.073, 0.090]$ for the axis loading and 1.000 $[1.000, 1.000]$ for the dense loading; this fixed-threshold covariance statistic does not control size over the full diagonal-covariance null. A post-hoc audit instead studentizes each sample covariance by its coordinate variances and uses the Frobenius norm of the off-diagonal sample correlation matrix. Its empirical size is 0.056–0.062 across four tested diagonal-Gaussian scales, and power at $\lambda_{\text{eff}} = 0.55$ is 0.064 $[0.050, 0.077]$ for the axis loading and 1.000 for the dense loading (Appendix A.2). This sensitivity analysis narrows the calibration concern for the correlation-based test in the tested settings; it does not retroactively validate the pre-specified covariance statistic or establish uniform size over every diagonal covariance. The same TopK protocol recovers both in all eight seeds, with mean cosines of 0.970 and 0.978, respectively. Detector classification differs between the two geometries for all 24 matched seed–strength pairs; binary recovery agrees in 22. This paired audit is descriptive, not a second pre-specified endpoint. An independent capacity confirmation fails its frozen all-condition gate because dense $d = 32, k = 4$ recovery is $0.625 < 0.75$. A post-hoc trace reaches 0.875 at 16k steps but does not change that decision (Appendix D).

Across the strength sweep, TopK recovery rises in both geometries, while independence retains an axis–dense gap. At $\lambda_{\text{eff}} = 0.32$, recovery remains near chance ($0.74/0.78$ versus 0.68); at 0.38, both geometries reach cosine 0.85. The full sweep is in Appendix Figure 6.

5.4 A SOURCE-MATCHED SYNTHETIC CONTROL ISOLATES ANCHOR-SOURCE EFFECTS

We compared training-PCA and independent random anchors in 192 matched pairs (384 SAE runs), holding data, seed, optimizer, and non-anchor initialization fixed. At the cosine-0.80 threshold, training-PCA recovered $183/192$ conditions (95.3%) versus $21/192$ (10.9%) for random anchors. The seed-cluster mean paired recovery difference was $+0.844$ (95% bootstrap CI $[0.797, 0.891]$),

and the target-cosine difference was $+0.641$ ($[0.597, 0.686]$). The objective-loss difference was $+0.00031$ ($[-0.00167, 0.00231]$), an interval that includes zero. This post-hoc study isolates an anchor-source effect on recovery in the tested rank-one Gaussian setting; it does not show an objective-loss benefit. Figure 3 shows the cluster-level outcomes, and Appendix D gives the protocol and paired estimates. The earlier correction-driven, structurally unmatched comparison is retained in Appendix A.1.

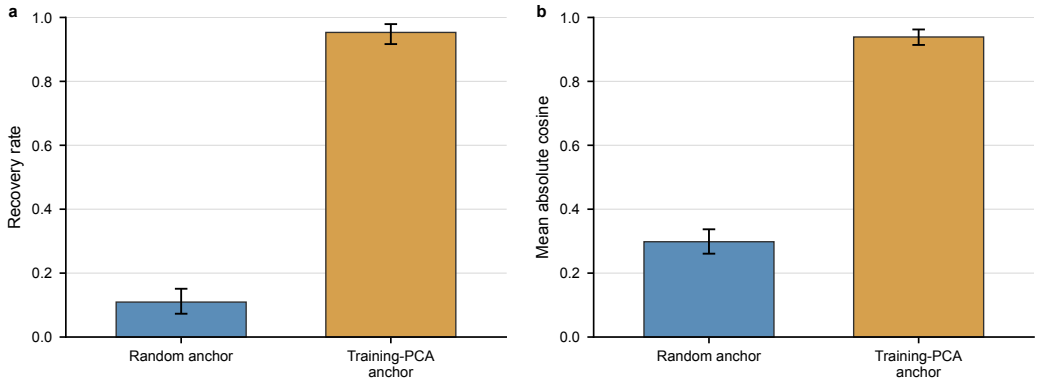


Figure 3: **Matched anchor-source comparison on synthetic rank-one Gaussian data.** (a) Recovery at absolute target cosine ≥ 0.80; the observed counts are 21/192 for random anchors and 183/192 for training-PCA anchors. (b) Mean absolute target cosine. Bars show means across 16 seed-cluster means (12 conditions per cluster); error bars show percentile 95% bootstrap confidence intervals from 20,000 resamples of the 16 seed clusters. The two arms contain 192 matched conditions each. Paired bootstrap intervals and the objective-loss comparison are reported in Appendix Table 7.

5.5 Known-direction injections in real residuals

We extracted GPT-2-small layer-6 and Pythia-70M layer-3 residuals from WikiText-2, fitted whitening on training sequences, and injected three known directions at two strengths. Within each model, 48 matched pairs shared the injected examples, signed-pair SAE structure, and all non-anchor initialization; only the anchor source differed. The evaluated atom was selected by training activation usage, while the injected direction shaped the training examples but was not given directly to the optimizer or selection rule. At the cosine-0.80 threshold, both arms recovered $0/48$ GPT-2 conditions and $1/48$ Pythia conditions. The PCA-minus-random difference in mean selected cosine was $+0.0101$ (95% seed-cluster CI $[0.0022, 0.0185]$) for GPT-2 and -0.0108 ($[-0.0734, 0.0413]$) for Pythia; held-out reconstruction differences were inconclusive. This post-hoc intervention tests known injected directions on empirical residual backgrounds, not native-feature discovery. Full protocol and model-specific estimates are in Appendix D.

6 Discussion

The E4-v2 control separates detector visibility from free-TopK recovery: under isotropic-null calibration, the axis loading has low independence-test power, yet matched TopK recovers it in all eight seeds. The statistic is not calibrated uniformly over the full composite diagonal-covariance null, and the independent $d = 32, k = 4$ confirmation fails its frozen all-condition gate. The held-out certificate assumes rank-one Gaussian covariance and a frozen decoder subspace; the spectral-anchor guarantees require explicit constraints. Post-hoc matched experiments find an anchor-source effect on synthetic Gaussian data but no recovery-rate advantage on injected GPT-2/Pythia residual backgrounds. Those injections test known directions and do not establish recovery of native residual features. The detector-specific recover-or-miss benchmark also fails to track all missed mass in trained SAEs. The evidence supports a detector-specific theoretical blind spot and bounded behavior in the tested controls, not universal decoder recovery.

AI USE STATEMENT

We used generative AI for manuscript editing and compression, experiment-design and validation planning, code implementation and debugging, scientific-figure planning and editing, and summarizing intermediate outputs or suggesting analyses. For writing support, we consulted procedural guidance from the Scientific Agent Skills library (Kassis et al., 2026). The authors made the final decisions about the hypotheses, mathematical claims, experimental protocols, and reported numerical results. AI-assisted suggestions were treated as drafts, not experimental evidence. The authors are responsible for verifying the proofs, protocols, numerical results, citations, figures, and final manuscript, and for the text, claims, and artifacts produced with AI assistance.

ETHICS STATEMENT

This study uses controlled synthetic features and residual activations from publicly available language-model data. It does not involve human participants, private personal data, or interventions. The experiments use known or externally specified directions where stated; these diagnostics are reported as such and are not presented as semantic discovery of named concepts.

REPRODUCIBILITY STATEMENT

The assumptions and complete proofs are provided in Appendices B–C. The experimental protocols, data splits, seeds, train–validation separation, held-out gates, and scope limitations are documented in Appendices C–D. The accompanying artifact contains the experiment scripts, machine-readable source-data tables, figure build scripts, and audit manifests needed to regenerate the reported figures and checks. The anonymous submission package will provide these materials without author-identifying metadata.

REFERENCES

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000. doi: 10.1090/mmono/191.

Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Conference on Learning Theory*, pp. 779–806, 2014. URL https://proceedings.mlr.press/v35/arora14.html.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory*, pp. 1046–1066, 2013.

Peter J Bickel, Chris AJ Klaassen, Ya'acov Ritov, and Jon A Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, 1993.

Joseph Bloom. GPT-2 small residual stream sparse autoencoders. Neuronpedia, 2024. URL https://www.neuronpedia.org/gpt2-small/6-res-jb.

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

T Tony Cai, Zongming Ma, and Yihong Wu. Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems*, volume 38, pp. 91489–91526, 2025. doi: 10.52202/085713-2755. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/764ff7477b8e24dbe01531f6791e8bdf-Abstract-Conference.html.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. Transformer Circuits Thread, 2022. URL https://transformer-circuits.pub/2022/toy_model/.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d3221cdb27e49d9c1cd35ad254feccfe-Abstract-Conference.html.

Joshua Engels, Logan Smith, and Max Tegmark. Decomposing the dark matter of sparse autoencoders. *Transactions on Machine Learning Research*, 2025b. URL https://openreview.net/forum?id=sXq3Wb3vef.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

Timothy Kassis, Vinayak Agarwal, Yuhuan He, Darshil Patel, and Aubrey M. Brueckner. Scientific agent skills: A library of procedural knowledge for research agents, 2026. URL https://arxiv.org/abs/2609.00065.

Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer, 1986. doi: 10.1007/978-1-4612-4946-7.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/84ca3f2d9d9bfca13f69b48ea63eb4a5-Abstract-Conference.html.

Adam Scherlis, Kshitij Sachan, Adam S. Jermyn, Joe Benton, and Buck Shlegeris. Polysemanticity and capacity in neural networks. *arXiv preprint arXiv:2210.01892*, 2022. URL https://arxiv.org/abs/2210.01892.

Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Conference on Learning Theory*, pp. 37.1–37.18, 2012. URL https://proceedings.mlr.press/v23/spielman12.html.

A Additional empirical details

This appendix collects secondary diagnostics for auditing the claim boundaries; they are not needed to understand the main detector–decoder comparison.

A.1 Correction-driven unmatched warm-start comparison

The training-only spectral warm start in this earlier follow-up uses the leading eigenvector of the training second moment. We recomputed PCA on its 96 training matrices ($m = 8$, $N_{\text{train}} = 20{,}000$, signal variance 0.55, three objectives, two capacities, two geometries, and eight fresh seeds). PCA reaches cosine ≥ 0.80 in $96/96$ conditions (mean 0.9991, minimum 0.9982). The corrected SAE comparison contains 96 pairs: vanilla initialization recovers in $45/96$ conditions (mean cosine 0.705), and spectral warm start in $89/96$ (mean cosine 0.920). The seed-cluster mean paired recovery difference is 0.458 (95% CI $[0.385, 0.531]$), target-cosine difference is 0.215 ($[0.169, 0.257]$), and objective-loss difference is -0.0032 ($[-0.0072, 0.0004]$). This correction-driven follow-up was designed after inspection of the invalid historical runs, and its vanilla arm does not match the spectral arm's signed-pair structure. It therefore does not isolate anchor source. The result is consistent with maintaining a direction already estimated by PCA through SAE training. The source-matched comparison is reported in Section 5.4.

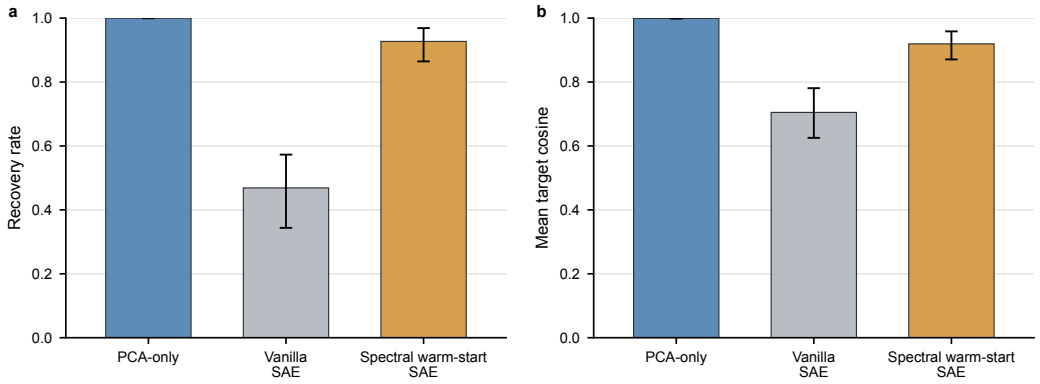


Figure 4: **PCA estimates the synthetic direction before SAE training.** (a) Recovery at target cosine ≥ 0.80; observed counts are 96/96 for PCA-only, 45/96 for vanilla SAE, and 89/96 for spectral warm start. (b) Mean target cosine. PCA uses training samples only; the target is used only for evaluation. The corrected fresh-seed SAE follow-up has 96 conditions per arm. Bars show means across eight base-seed cluster means (12 conditions per cluster); error bars show percentile 95% bootstrap confidence intervals from 20,000 resamples of the eight clusters. The vanilla initialization does not match the spectral arm's signed-pair structure.

A.2 Post-hoc diagonal-null calibration sensitivity

The pre-specified E4 statistic is $T_{\mathrm{cov}}(H) = \|\operatorname{offdiag}(\widehat{\Sigma})\|_F$, with its threshold estimated under an isotropic Gaussian null. Its null distribution depends on the coordinate variances, so that threshold need not control size for the full diagonal-covariance family. As a post-hoc sensitivity analysis, we used $T_{\mathrm{corr}}(H) = \|\operatorname{offdiag}(\widehat{R})\|_F$, where $\widehat{R}_{ij} = \widehat{\Sigma}_{ij}/\sqrt{\widehat{\Sigma}_{ii}\widehat{\Sigma}_{jj}}$, with a threshold calibrated under $\mathcal{N}(0, I)$. Under a Gaussian diagonal covariance, this statistic is invariant to coordinate-wise rescaling. The audit uses eight seed blocks; each threshold uses 800 calibration draws, and each diagonal-null scale is evaluated on 2,400 draws pooled across blocks. For the alternatives, each seed block contributes 400 trials per geometry and strength. The protocol and complete outputs are included in the supplement. This was not the pre-specified endpoint and reuses the E4 design, so it is a calibration sensitivity analysis rather than independent confirmation.

Table 1: Empirical null rejection rates under diagonal Gaussian covariance. Each variance scale modifies coordinate 1 and leaves the remaining seven variances at one. There are 2,400 evaluation draws per row; each seed-specific critical value is calibrated from 800 isotropic Gaussian draws at nominal $\alpha = 0.05$.

Variance scale	Original covariance statistic	Correlation statistic	Trials
1.000000	140/2400 (5.83%)	140/2400 (5.83%)	2,400
1.287375	231/2400 (9.63%)	135/2400 (5.63%)	2,400
2.000000	619/2400 (25.79%)	149/2400 (6.21%)	2,400
10.000000	2273/2400 (94.71%)	139/2400 (5.79%)	2,400

At $\lambda_{\mathrm{eff}} = 0.55$ under the original Bernoulli feature alternatives, the correlation statistic's seed-block mean power is 0.064 for the axis loading (seed-bootstrap 95% interval $[0.050, 0.077]$) and 1.000 for the dense loading. Thus the corrected statistic retains the axis–dense power contrast while keeping its size near nominal over the four tested diagonal-Gaussian scales. The tested scales are finite and do not establish a uniform finite-sample size guarantee over the continuum of diagonal covariance matrices.

A.3 RESIDUAL-FLOOR DERIVATION

Let β_N be the decision threshold induced by the model-selection rule. For $C_i(B) > 0$, define

$$\lambda_{c,i}(N) = \Big(\frac{\beta_N}{NC_i(B)}\Big)^{1/4}, \qquad f_{\mathrm{dark}}(N) = \frac{\sum_i \lambda_i^2\, 1\{\lambda_i < \lambda_{c,i}(N)\}}{\sum_i \lambda_i^2}. \tag{10}$$

If $C_i(B) = 0$, set $\lambda_{c,i}(N) = \infty$ for this detector class: the feature is invisible to this local test at every sample size under the exact frame-aligned model. For unconstrained reconstruction, Theorem 6 gives a separate linear-encoder rate and a conditional λ^{-2} isolation benchmark after a separating direction is available; neither follows by substituting a reconstruction coefficient into this detector threshold.

Theorem 7 (Critical boundary and dark-matter fraction)**.** *Under Theorem 1, the subcritical set at sample size N is $I_{\mathrm{dark}}(N) = \{i : \lambda_i < \lambda_{c,i}(N)\}$, and the finite-feature $f_{\mathrm{dark}}(N)$ is its variance share. For an i.i.d. feature population, define the corresponding population fraction by $f^{\mathrm{pop}}_{\mathrm{dark}}(N) = \mathbb{E}[\lambda^2 \mathbf{1}\{NC(B)\lambda^4 < \beta_N\}]/\mathbb{E}[\lambda^2]$. Suppose $\mathbb{E}[\lambda^2] \in (0,\infty)$, $\beta_N \asymp \log N$, $\Pr(\lambda \le x) \asymp x^\alpha$ near zero, and $0 < c_- \le C(B) \le c_+ < \infty$ uniformly over the features in this tail regime. If the quartic expansion is uniform there, then $f^{\mathrm{pop}}_{\mathrm{dark}}(N) \asymp (\log N/N)^{(\alpha+2)/4}$. Without the coefficient bounds, frame-aligned or near-aligned features can change this rate; it is not implied by the strength distribution alone.*

Proposition 4 (Recover-or-miss residual identity)**.** *Under the recover-or-miss idealization that supercritical features contribute negligible residual variance and subcritical features contribute their full variance, the residual-variance share equals $f_{\mathrm{dark}}(N)$.*

Under the recover-or-miss idealization, the residual-variance share equals $f_{\mathrm{dark}}(N)$. This is a conditional accounting identity, not a prediction or lower bound for trained-SAE residuals. Optimization failure, absorption, correlated latents, and nonlinear features can change which features are recovered and how much residual variance remains.

Table 2: Synthetic suite. Intervals are seed-level bootstrap 95% confidence intervals. These rows summarize secondary controls; the matched E4-v2 endpoint remains in the main text.

Check	**Result**
Detector $N^\star(\lambda)$ slope	-1.95 vs. -3.72 (theory: -2 vs. -4)
Same-λ independence power	axis 0.096 $[0.090, 0.102]$; dense 0.969 $[0.965, 0.974]$
Rotate B, hold λ	ind vs. $1 - \|B^\top w\|_4^4$: $r = 0.977$; label $\equiv 1$
2-sparse imbalance ε	0.123 (ε=0.01) $\to$ 0.810 (ε=0.50)
Ind recover-or-miss benchmark	predicted vs. missed variance $r = 0.999$, MAE 0.005
TopK cosine vs. λ	axis/dense $0.85/0.85$ at 0.38; 0.32 remains near chance
Label vs. SAE $N^\star$ (m=32)	$-1.96/-1.99$ vs. $-2.97/-4.33$ (Theorem 6)

A.4 STRENGTH SWEEP

Figure 5 reports the matched-strength detector calibration, and Figure 6 reports TopK recovery and independence power across the full strength sweep.

A.5 RECONSTRUCTION-RATE AND RESIDUAL-FLOOR DIAGNOSTICS

The reconstruction-rate controls test Theorem 6 with a chance cosine of approximately 0.36 and a recovery threshold of 0.80. The labeled estimator follows the λ^{-2} rate, while empirical PCA is close to λ^{-4}. Trained linear and TopK encoders show no corresponding C=0 axis hole.

The residual diagnostics use an 80-feature mixture (8 axis-aligned and 72 random directions) with heavy-tailed strengths. With dictionary size 192 and $k = 8$, mean recovery is 0.91 for axis features and 0.85 for random directions; axis features account for 6% of misses. At fixed $N = 16384$ and $k = 4$, the missed-variance share falls from 0.79 at $d = 8$ to 0.06 at $d = 256$. The fixed-basis independence predictor remains near 0.11 and does not track capacity. The detector-specific quartic benchmark has correlation $r = 0.013$ with missed mass; it is neither a lower bound nor an explanation of the full trained-TopK residual share.

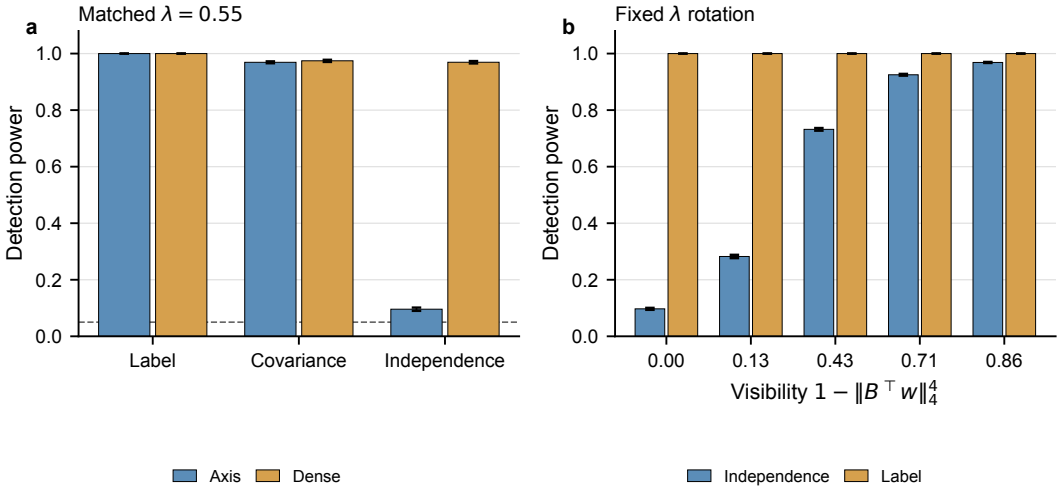


Figure 5: Matched-strength detector calibration. (a) Detector power for axis and dense loadings at $\lambda = 0.55$; error bars are seed-level bootstrap 95% intervals. (b) Power across basis orientations at fixed strength; bars are seed means and error bars are s.e.m. Independence power changes with visibility while label power remains one.

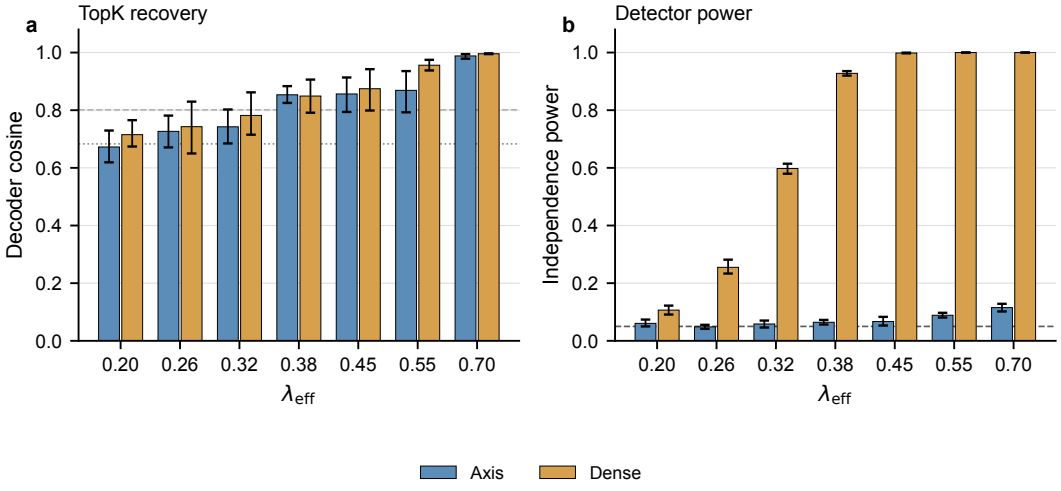


Figure 6: Grouped-bar strength sweep for the free TopK decoder and fixed-basis independence detector. Recovery rises with signal strength in both geometries, while independence power retains an axis–dense gap; intervals are seed-level bootstrap 95% intervals.

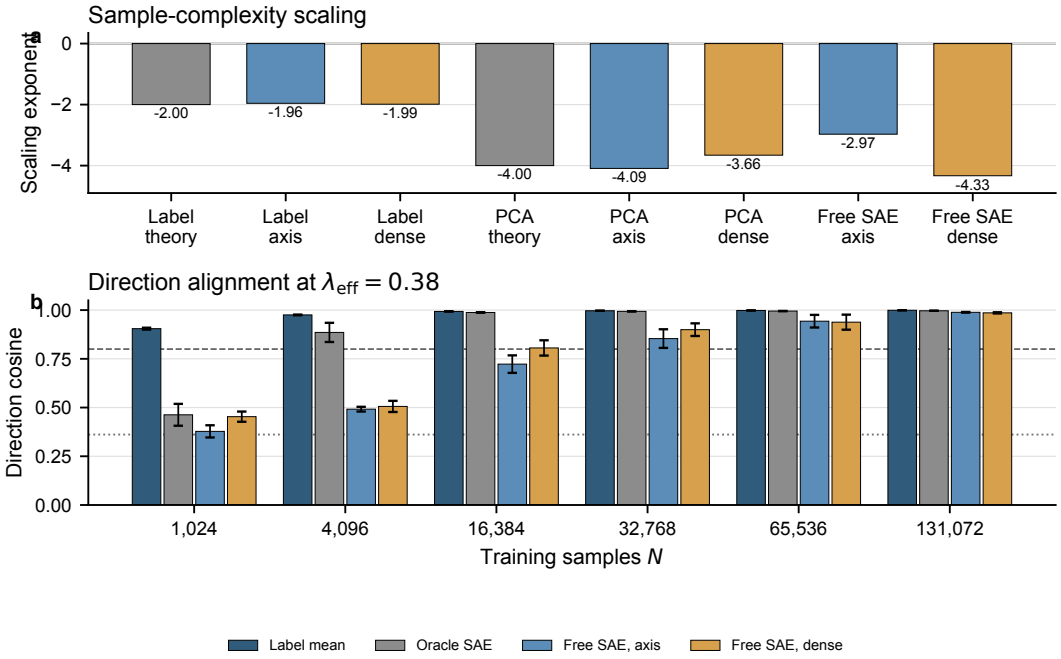


Figure 7: Reconstruction-rate controls in $m = 32$. (a) Empirical log–log slopes of $N^\star$ versus λ across seven strengths, alongside the theoretical -2 and -4 exponents. (b) Mean direction cosine for label-mean, oracle, and free-SAE estimates at $\lambda = 0.38$ (error bars: s.e.m.). Lines mark chance and the shared 0.80 threshold; sample counts are shown exactly. The empirical and trained encoders show no axis-specific $C{=}0$ floor.

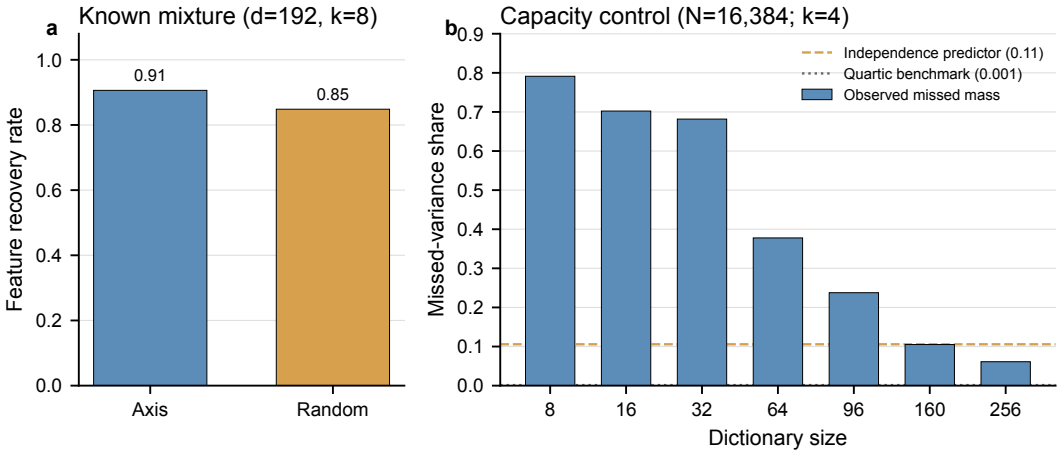


Figure 8: Residual diagnostics on the known 80-feature mixture. Capacity changes the observed missed-variance share, whereas the fixed-basis independence predictor does not track the sweep. The quartic curve is a detector-specific recover-or-miss benchmark, not a lower bound or a trained-SAE identity.

A.6 GPT-2/Pythia residual transfer and public atoms

The corrected bridge uses an explicit WikiText-2 file without a synthetic-text fallback and fits PCA whitening on the training split. GPT-2-small layer-6 and Pythia-70M layer-3 receive Bernoulli injections along two whitened axes and one dense loading. For the implemented quadratic statistic, an exactly one-hot loading yields an algebraic zero by construction; this is not an empirical finding that the detector has calibrated power or size for those axes. Fixed-background covariance misses only PC1 in the reported runs. A separate post-hoc matched-anchor test on these empirical residual backgrounds, with known directions injected into its training examples, reached the prespecified 0.80 recovery threshold in only $0/48$ conditions for either arm on GPT-2 and $1/48$ for either arm on Pythia; the small selected-cosine differences did not yield a consistent cross-model recovery benefit (Appendix D). Historical real-residual warm-start runs, including the Pythia-160M scale and budget extensions, are excluded because they used the flawed initialization helper described in Appendix D. Without injection, the token difference-of-means and public-SAE atom audits evaluate specified directions rather than named-concept discovery.

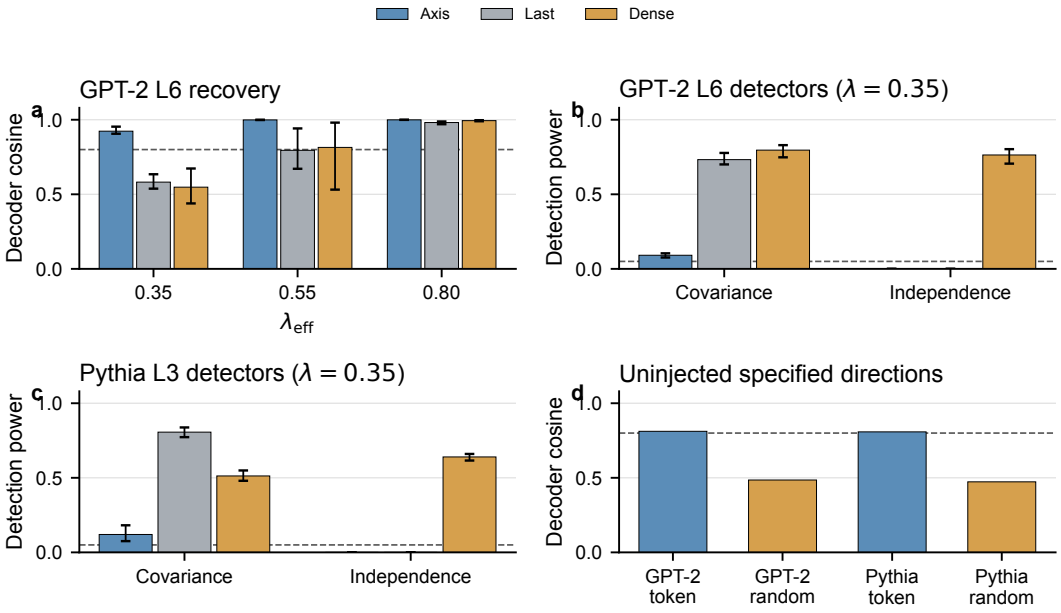


Figure 9: Corrected real-WikiText GPT-2/Pythia residual bridge. Bars show GPT-2 recovery across strength, detector power for both models, and uninjected specified-direction cosine. Detector and recovery summaries use four bridge seeds where available. Labels denote known injected or externally specified directions, not discovered concepts.

A published GPT-2 residual SAE (Bloom, 2024) with $24{,}576$ atoms has a minimum residual-coordinate C of 0.140. Ten near-duplicate columns concentrated on residual coordinate 447 account for the 0.04% of atoms below 0.20. This audit evaluates the published decoder without retraining and does not establish named-concept discovery.

Table 3: Representative Bloom layer-6 atoms. The reported cosine is alignment to the encoder-induced firing/non-firing contrast direction, not an independent label or concept-discovery score.

Atom	C	C_{32}	Documented as	Enc.-contrast cos.
23123	0.140	0.122	residual 447 ($\times$10)	0.998
1	0.996	0.862	grants / funding	0.065
1834	0.996	0.904	VAT / tax terms	0.078
0	0.996	0.902	notice / realization	0.087

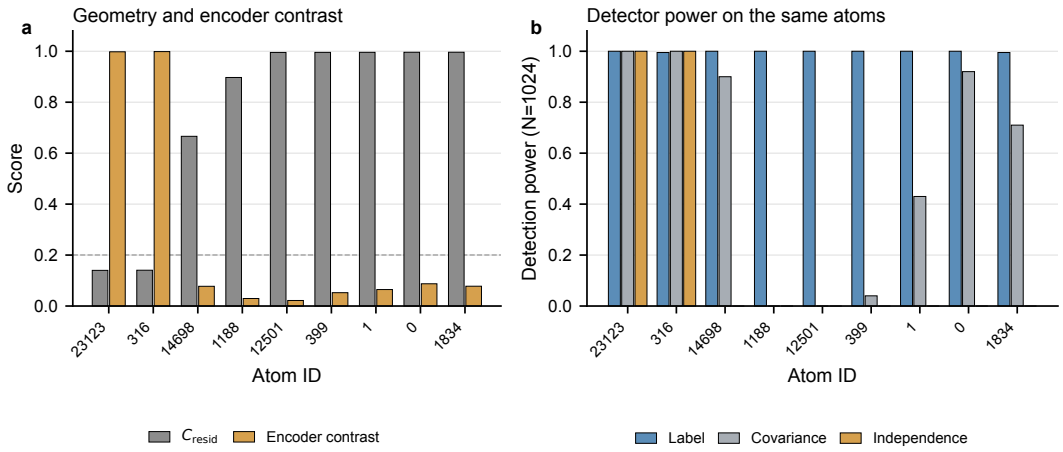


Figure 10: Published Bloom GPT-2-small layer-6 residual SAE evaluated without injection or retraining. In the first panel, gray bars show residual-coordinate C and amber bars cosine with the encoder-induced firing/non-firing contrast direction across nine specified atoms. The second panel orders detector bars as label, covariance, and independence (blue, gray, amber). This contrast is derived from the same encoder activations, not an independent external semantic label; the analysis is not a concept-discovery result.

B Proofs

B.1 Gaussian identities

Lemma 1. *If* $P = \mathcal{N}(0, \Sigma_1)$ *and* $Q = \mathcal{N}(0, \Sigma_0)$ *are nondegenerate,* $D_{\mathrm{KL}}(P\|Q) = \frac{1}{2}\big(\mathrm{tr}(\Sigma_0^{-1}\Sigma_1) - m + \log\det\Sigma_0 - \log\det\Sigma_1\big)$.

Lemma 2. *For* $s > 0$, $\lambda \geq 0$, *and* $\|w\| = 1$, $\det(s^2 I_m + \lambda^2 w w^\top) = s^{2m}(1 + \lambda^2/s^2)$.

B.2 Labeled quadratic benchmark

Assume the activation indicator z_i is observed and $h = z_i W_i + u$ with $u \sim \mathcal{N}(0, \sigma_{\text{eff}}^2 I_m)$. Then $D_{\mathrm{KL}}(\mathcal{N}(W_i, \sigma_{\text{eff}}^2 I)\|\mathcal{N}(0, \sigma_{\text{eff}}^2 I)) = \|W_i\|^2/(2\sigma_{\text{eff}}^2)$. The expected information per observation is of order λ_i^2, so $N_i^* \asymp \sigma_{\text{eff}}^2 \lambda_i^{-2}$.

B.3 Proof of Theorem 1

Work in detector coordinates $y = B^\top h$ and write $v = B^\top w$. Because B is orthonormal, the alternative covariance in these coordinates is $\Sigma_1 = s^2 I + \lambda^2 v v^\top$, with $\|v\| = 1$; the detector's null covariance is $D(d) = \mathrm{diag}(d)$. The KL $\Phi(d) = \frac{1}{2}\{\mathrm{tr}(D(d)^{-1}\Sigma_1) - m + \log\det D(d) - \log\det\Sigma_1\}$ decouples across coordinates. Its unique minimizer is $d_j^\star = (\Sigma_1)_{jj} = s^2 + \lambda^2 v_j^2$. Substitution gives $\mathrm{tr}(D_\star^{-1}\Sigma_1) = m$, so

$$D_{\mathrm{loc}}(\lambda) = \tfrac{1}{2}\Big(\sum_{j=1}^{m} \log(1 + t v_j^2) - \log(1+t)\Big), \qquad t = \lambda^2/s^2. \tag{11}$$

The expansion $\log(1+u) = u - u^2/2 + O(u^3)$ and $\sum_j v_j^2 = 1$ give $\sum_j \log(1+tv_j^2) - \log(1+t) = \frac{t^2}{2}(1 - \|v\|_4^4) + O(t^3)$. Multiplying by $1/2$ yields $D_{\mathrm{loc}} = C_i(B)\lambda^4 + O(\lambda^6)$ with $C_i(B) = (1 - \|v\|_4^4)/(4s^4)$. Thus the local sample-complexity scale is $N^* \asymp C_i(B)^{-1}\lambda^{-4}$ when $C_i(B) > 0$. If $v = e_j$, the alternative lies in the diagonal null family and the exact KL is zero.

B.4 PROJECTION IDENTITY

Write the local perturbation parameter as $\eta = \lambda^2$. The log-likelihood of P_η against P_0 has score ψ_v. A regular local parameterization of $\mathcal{M}$ has scores in T. The second-order local KL is the least-squares residual $\eta^2/2\,\|\psi_v - \Pi_T\psi_v\|^2 + O(\eta^3)$. Substituting $\eta = \lambda^2$ yields Proposition 1. In detector coordinates, the diagonal-family tangent space is spanned by $e_j(y) = (y_j^2 - s^2)/(2s^4)$. Then $\Pi_T\psi_v = \sum_j v_j^2 e_j$ and the residual is $(2s^4)^{-1}\sum_{j\neq\ell} v_j v_\ell y_j y_\ell$, with squared L^2 norm $(1 - \|v\|_4^4)/(2s^4)$, recovering Theorem 1.

B.5 PROOF OF THEOREM 2

Take D with first column w and remaining columns arbitrary unit vectors, and $z = \alpha e_1$. Then $Dz = \alpha w$ for every unit w, including a frame-aligned loading with $B^\top w = e_j$. The rank-one mean-shift family is therefore in the closure of $\mathcal{F}_k$ even when $C_i(B) = 0$. This establishes availability of that direction; it makes no claim about the optimization path or about projecting probability models onto $\mathcal{F}_k$. By contrast, $\mathcal{M}_{\mathrm{ind},B}$ contains every frame-diagonal covariance spike, which is why the detector's quartic coefficient vanishes in this case.

B.6 PROOF OF PROPOSITION 2

Fix a unit target w and choose any unit $v \perp w$, which exists because $m \geq 2$. Let the first decoder column be v and choose a one-sparse code $z = \alpha e_1$. Then $Dz = \alpha v$ is feasible in $\mathcal{F}_k$, but its alignment with w is $|v^\top w| = 0$. Since the construction is feasible for every target and the class contains both aligned and orthogonal choices, membership in $\mathcal{F}_k$ cannot imply a positive target-uniform alignment lower bound. The argument quantifies over feasible decoders only; it does not describe a learner that observes samples and selects v.

B.7 PROOF OF THEOREM 3

For any orthogonal projector P, $R(P) = \mathrm{tr}((I - P)\Sigma)$. Since P_w has rank r and contains w, $\mathrm{tr}(P_w\Sigma) = r\sigma^2 + \tau$. For P, $\mathrm{tr}(P\Sigma) = r\sigma^2 + \tau w^\top P w$, so subtraction gives $R(P) - R(P_w) = \tau(1 - w^\top P w) = \tau(1 - \|Pw\|^2)$.

For the validation statement, the difference of the two projected quadratic losses is a centered Gaussian quadratic form after subtracting its expectation. The standard Gaussian quadratic-form tail bound for the difference of two rank-r projectors (whose difference has rank at most $2r$) gives, with probability at least $1 - \delta$, an error at most $2(\sigma^2 + \tau)(\sqrt{4r\log(2/\delta)/n} + \log(2/\delta)/n)$. Substituting this upper confidence bound for the excess risk into the exact identity yields the stated cosine lower bound. The proof uses only the frozen projector and the held-out samples; it does not use how the projector was produced.

B.8 PROOF OF THEOREM 4

For the population covariance, w is the simple leading eigenvector and the eigengap is τ. The standard eigenvector perturbation bound for a symmetric perturbation $E = S - \Sigma$ gives $\sin\angle(\widehat{u}, w) \leq \|E\|_{\mathrm{op}}/(\tau - \|E\|_{\mathrm{op}})$ whenever $\|E\|_{\mathrm{op}} < \tau/2$. A Gaussian sample-second-moment bound follows by applying the extreme singular-value inequality to $\Sigma^{-1/2}[X_1, \ldots, X_N]$: with probability at least $1-\delta'$, $\|S-\Sigma\|_{\mathrm{op}} \leq \|\Sigma\|_{\mathrm{op}}(2q+q^2)$ for $q = (\sqrt{m}+\sqrt{2\log(2/\delta')})/\sqrt{N}$. A union bound over the G training anchors uses $\delta' = \delta/G$. The signed pair is implemented by nonnegative anchor activations, so it uses one active feature slot; projecting each residual decoder column onto $\widehat{u}^\perp$ makes residual updates unable to change the anchor component. Hence residual objective, optimizer, initialization, and residual capacity do not enter the anchor direction or its bound.

B.9 PROOF OF THEOREM 5

Let $\theta_0 = \angle(\widehat{u}_0, w)$ and $\theta = \angle(\widehat{u}, w)$. The spherical triangle inequality gives $\theta \leq \theta_0 + \angle(\widehat{u}, \widehat{u}_0) \leq \arcsin(s) + \rho$. Since the right-hand side is below $\pi/2$ and cosine is decreasing there, $\cos\theta \geq \cos(\arcsin(s) + \rho)$. The residual leakage constraint is allowed to be nonzero, but it does not change

the direction of the anchor or the preceding angular inequality. The anchor and residual constraints are enforced after each optimizer update, so the conclusion is uniform over the objective, capacity, optimizer, and initialization within this family. Applying the same union bound over the G training anchors gives the stated simultaneous probability.

B.10 Proof of Theorem 6

Write $h = \varepsilon + a(z-p)w$ with $\varepsilon \sim \mathcal{N}(0, I_m)$ independent of $z \sim \mathrm{Bern}(p)$. Then $\mathbb{E}[h] = 0$ and $\Sigma = \mathbb{E}[hh^\top] = I_m + \beta ww^\top$ with $\beta = \lambda^2(1-p)$ and $\lambda = a\sqrt{p}$.

Linear encoder. One-sparse linear reconstruction $\hat{h} = \alpha(u^\top h)u$ is rank-one PCA. The population maximizer of $u^\top \Sigma u$ is $u = \pm w$ for every unit w. For the sample covariance $S = N^{-1}\sum_{i=1}^N h^{(i)}h^{(i)\top}$, write $P = I_m - ww^\top$. Then $P\Sigma w = 0$ and

$$PSw = \frac{1}{N}\sum_{i=1}^{N}(Ph^{(i)})(w^\top h^{(i)}).$$

The orthogonal component $Ph \sim \mathcal{N}(0, P)$ is independent of $w^\top h$, whose mean is zero and variance is $1+\beta$. Thus $\mathbb{E}\|PSw\|^2 = (m-1)(1+\beta)/N$. First-order perturbation along the eigengap β yields $\hat{u} = w + \beta^{-1}PSw + O_P(N^{-1})$ and

$$\mathbb{E}[\sin^2\theta] = \frac{(m-1)(1+\beta)}{N\beta^2} + O(N^{-2}).$$

The right-hand side does not depend on the coordinates of w. Setting $\sin^2\theta \le 1 - 0.80^2$ gives $N^\star \asymp \lambda^{-4}$.

Conditional isolation benchmark. Assume a separating direction u is already available and a threshold on $u^\top h$ recovers the labels as $a \to \infty$. By the multivariate central limit theorem, the difference of the two sample means is asymptotically distributed as $\mathcal{N}(aw, (n_1^{-1} + n_0^{-1})I_m)$ with $n_1 \approx pN$ and $n_0 \approx (1-p)N$. Its noncentrality is $\kappa = a\sqrt{p(1-p)N} = \lambda\sqrt{(1-p)N}$, so $\mathbb{E}[\cos\theta] \approx \kappa(\kappa^2 + m - 1)^{-1/2}$. Setting this to 0.80 produces $N^\star = (16/9)(m-1)/[(1-p)\lambda^2]$. This is the cost of the isolation stage only; it does not include acquiring u and is not an end-to-end trained-SAE guarantee.

No detector coefficient. Both derivations use an unconstrained direction $u \in \mathbb{S}^{m-1}$. The diagonal Gaussian tangent space never appears, so the detector-specific factor $1 - \|B^\top w\|_4^4$ does not multiply either reconstruction rate.

B.11 Proof of Proposition 3

Clamp coordinate 0 of the code to $z - p$ and write the reconstruction as $(z-p)u + D_{-0}z_{-0}(h)$. Expanding the population MSE in u and taking the gradient yields the normal equation

$$u \propto \mathbb{E}\big[(z-p)\big(h - D_{-0}z_{-0}(h)\big)\big] = a\,p(1-p)\,w - \mathbb{E}\big[(z-p)D_{-0}z_{-0}(h)\big].$$

The second term is the absorbed component. It vanishes if and only if the remaining dictionary is uncorrelated with $(z-p)w$. Only its component orthogonal to w changes the direction: if that component is nonzero and remains non-negligible, the normalized residual mean has alignment strictly below one. A collinear absorbed term changes scale but not direction, so finite-N recovery must be analyzed jointly with the cost of fitting D_{-0}.

B.12 Proof of Theorem 7

Detectability requires $ND_{\mathrm{loc}}^{(i)} \gtrsim \beta_N$. For positive detector coefficient, the leading-order inversion of $NC_i(B)\lambda_i^4 \gtrsim \beta_N$ gives $\lambda_{c,i}(N)$. Summing λ_i^2 over $\{\lambda_i < \lambda_{c,i}(N)\}$ and dividing by total feature variance gives the finite-population version of $f_{\mathrm{dark}}(N)$. Under the theorem's uniform coefficient bounds, each cutoff is within constant factors of $(\log N/N)^{1/4}$. The lower-tail condition then gives $\mathbb{E}[\lambda^2 1\{\lambda < \lambda_c\}] \asymp \lambda_c^{\alpha+2}$, yielding the stated population rate after normalization by $\mathbb{E}[\lambda^2]$. The recover-or-miss identity follows only after imposing its binary reconstruction rule; it is not a lower-bound argument.

C EXPERIMENTAL PROTOCOL

Detector calibration (E1 and E4-v2). Both tests use $m = 8$, significance level $\alpha = 0.05$, eight seeds with base 20260824, and null quantiles estimated from H_0 draws only. E1 uses Gaussian latent activations, $N = 1024$, 700 alternative trials, and 2,000 null draws; the $N = 1024$, $\lambda = 0.55$ setting was locked on seed 0 to compare axis and dense independence power. E4-v2 uses centered Bernoulli activations with $p = 0.05$, $\lambda_{\text{eff}} \in \{0.35, 0.55, 0.80\}$, $N = 4096$, 400 alternative trials, and 800 null draws. The covariance background is I and is never refit on the alternative. Rotating B changes only the coordinates used by the independence statistic; label power is a protocol check and remains one. The matched E4-v2 TopK decoder uses $d = 16$, $k = 2$, 20,000 training samples, and 4,000 optimization steps. Decoder cosine ≥ 0.80 is the recovery criterion and was fixed before inspecting results.

Synthetic reconstruction-rate grids. The fixed-step SAE grid uses $N \in \{500, \ldots, 64000\}$, six seeds, the same cosine criterion, and 4,000 Adam steps. Small-N runs therefore receive more epochs, so that slope is not presented as an epoch-controlled comparison. The constant-epoch labeled/unlabeled contrast in the main text uses $m = 32$, dictionary size 16, and 80 epochs. Its shared cosine threshold is 0.80, above the $m = 32$ chance floor of approximately 0.36. The labeled difference-of-means channel measures $N^\star$ over $N \in \{8, \ldots, 1024\}$ at strengths $\{0.30, 0.34, 0.38, 0.42, 0.48, 0.55, 0.65\}$. A parallel $m = 8$ run has a 16-atom chance floor of approximately 0.68, leaving little separation at the same threshold.

The wide SAE grid uses $m = 32$, $N \leq 131$k, the same seven strengths, six seeds, and 80 epochs. It compares oracle-encoder TopK (the target code is clamped to atom 0) with free TopK. The plotted first-principles references are the conditional-isolation benchmark $N^\star_{\text{label}} = (16/9)(m-1)/[(1-p)\lambda^2]$ and the PCA benchmark $N^\star_{\text{PCA}} = (m-1)(1+\beta)/(0.36\beta^2)$, where $\beta = \lambda^2(1-p)$ and the target cosine is 0.80. Regime controls include empirical PCA, rank-one linear gradient descent, and one-atom labeled/unlabeled TopK ($d = k = 1$).

Transformer-residual and capacity diagnostics. The GPT-2 bridge uses layer-6 residuals from WikiText-2, PCA whitening fit on the training split, and known-direction injections along e_1, e_{32}, and a dense loading in the leading 32 whitened coordinates. It uses four seeds, $N = 1024$ known-direction tests, and a TopK dictionary of size 64 with $k = 4$. The Pythia-70M layer-3 residual (hidden size 512) uses the same injection protocol. Uninjected GPT-2 TopK is evaluated on 40 token difference-of-means directions and 40 random unit directions. The capacity diagnostic uses the known 80-feature mixture, $k = 4$, $N = 16384$, and dictionary sizes from 8 to 256.

The public residual-SAE audit evaluates the unretrained Bloom GPT-2-small layer-6 decoder (Bloom, 2024) with 24,576 atoms. Its encoder is $\text{ReLU}((h - b_{\text{dec}})W_{\text{enc}} + b_{\text{enc}})$; atom scores use the same WikiText residuals at $N = 1024$ without injection. A separate uninjected GPT-2 SAE uses dictionary size 128 and $k = 8$ on the same 32-dimensional slice. Code and machine-readable JSON artifacts accompany the submission.

Follow-up validation protocols. The corrected synthetic follow-up uses eight fresh base seeds (20270109–20270116), three objectives, two capacities, and two geometries, yielding 96 matched pairs. It fixes the signed-pair encoder-bias assignment bug, but was initiated after inspection of the earlier invalid results and is not preregistered confirmation. A second post-hoc synthetic follow-up (seeds 20270117–20270132) matches signed-pair architecture and all non-anchor initial parameters while varying only whether the signed pair is anchored to training PCA or an independent random unit vector; its complete 192-pair result is reported in Appendix D. Neither follow-up is confirmatory. The historical eight-seed synthetic runs and real-residual matched runs used the defective helper and remain excluded (Appendix D). A new post-hoc empirical-background transfer study uses three known injected directions, two strengths, eight seed clusters per model, 24k training and 3k validation samples per condition, and matched 3,000-step d=64, k=4 TopK training on GPT-2-small L6 and Pythia-70M L3 residuals. It uses an explicit WikiText-2 file, disjoint training and validation sequences, and training-only whitening. The injected direction shapes the training examples but is not directly supplied to the optimizer or atom-selection rule. The model-specific results and their limited scope are reported in Appendix D; they do not support native semantic discovery or transfer of the Gaussian certificate.

D HELD-OUT CAPACITY/SPARSITY CONFIRMATION

After E4-v2, we froze an independent confirmation before execution: $m = 8$, centered Bernoulli $p = 0.05$, reported $\lambda_{\text{eff}} = 0.55$, $N = 4096$ detector samples, and eight new seeds beginning at 20260901. The detector gate passed (axis ind power 0.097; dense ind power 1.000; label and cov power 1.000 for both). We trained the same 20k-sample, 4000-step TopK protocol over $(d, k) \in \{(8, 1), (16, 2), (32, 4)\}$. Every configuration was required to have axis and dense recovery at least 0.75, absolute recovery gap at most 0.25, and axis cosine penalty at most 0.05. The complete artifact has 16 detector and 48 training rows.

Table 4: Independent held-out capacity/sparsity confirmation (eight new seeds). The composite gate requires every row to pass. The d=32, k=4 dense recovery rate misses its frozen minimum, so all-pass is false. This table is a limitation result, not support for an expanded claim.

(d, k)	axis rec.	dense rec.	axis–dense	cosine gap	gate
$(8, 1)$	1.000	1.000	0.000	-0.023	pass
$(16, 2)$	1.000	1.000	0.000	-0.027	pass
$(32, 4)$	0.750	0.625	0.125	$+0.005$	fail (dense rec.)

The failure is not axis-specific: at $(32, 4)$, the axis recovery rate is higher, and both the recovery-gap and cosine-penalty gates pass. Instead, this result shows that the tested budget does not support a capacity- or sparsity-invariant recovery claim. We therefore limit the main-text non-implication to E4-v2 and provide the complete raw and summary JSON for audit.

Post-hoc budget diagnostic. To distinguish a fixed-budget failure from a persistent configuration effect, we later froze a separate d=32, k=4 checkpoint trace using the same eight held-out seeds, data seeds, and initialization. The 4k checkpoint exactly reproduces all 16 original d32-k4 trajectories; the 8k and 16k checkpoints continue those paths without restarting. Because this diagnostic is post hoc, it does not change the original composite-gate outcome in Table 4. Table 5 reports the checkpoint trajectory.

Table 5: Post-hoc d32-k4 budget trace on the same held-out trajectories. Recovery is decoder cosine ≥ 0.80 across eight seeds. At 16k steps both geometry recovery rates pass the frozen follow-up threshold, but this is a training-budget diagnostic rather than a new independent confirmation.

steps	axis rec.	dense rec.	axis cos.	dense cos.	axis–dense rec.
4k	0.750	0.625	0.841	0.837	0.125
8k	0.875	0.750	0.936	0.887	0.125
16k	1.000	0.875	0.948	0.925	0.125

At 16k steps, the follow-up passes the recovery, recovery-gap, axis-cosine-penalty, and dense-recovery-gain gates. A dense miss remains on one seed, so the result supports a budget explanation for the earlier d32-k4 failure, not an all-seed equivalence guarantee or a revision of the original confirmation decision.

Objective and capacity stress test. To test transfer of the certificate rather than of a training algorithm, we pre-specified a separate Gaussian rank-one stress protocol for this corrected rerun. It covers three objectives (MSE, MSE+ℓ_1 code penalty, and MSE+ℓ_2 code penalty), two overcomplete TopK capacities, axis and dense geometries, four new seeds, and 100k independent validation samples per run. Candidate atoms were selected using training activation statistics only. Table 6 summarizes the 48 completed rows. This stress test does not show that every objective or capacity optimizes successfully. Table 9 summarizes the separate target-free anchor-preserving matrix.

Table 6: Corrected cross-objective and capacity stress test for the reconstruction certificate. Entries are certificate pass rates over four seeds, pooling axis and dense geometries, under the v3 projector-difference concentration bound. Although the certificate is objective-agnostic under its assumptions, training does not reach it uniformly across objectives or capacities.

Objective	d=16, k=2	d=32, k=4	Overall
MSE	$5/8$	$2/8$	$7/16$
MSE+ℓ_1	$8/8$	$0/8$	$8/16$
MSE+ℓ_2	$3/8$	$1/8$	$4/16$
Overall	$16/24$	$3/24$	$19/48$

The stress matrix separates the objective-independent certificate from objective-dependent optimization success. Under the corrected v3 bound, the overall certificate pass rate is $19/48 = 0.3958$, and the direct recovery rate is 0.4792. MSE+ℓ_1 at d=16, k=2 passes all eight cases, whereas only $3/24$ cases pass at d=32, k=4. We therefore report the certificate as an objective-independent post-training guarantee, not as a universal training-recovery claim.

Warm-start implementation audit and follow-up. The earlier signed-pair bias bug invalidates the historical warm-start runs. The correction-driven unmatched comparison is described in Appendix A.1. The separate source-matched result varies only the anchor source and is reported in Section 5.4; its full estimates are in Table 7. The GPT-2/Pythia transfer test injects known directions into training examples and is detailed in Table 8. Neither study tests native-feature discovery or transfers the Gaussian certificate to empirical residuals.

Table 7: Post-hoc matched-structure anchor-source control on synthetic rank-one Gaussian data. Arm-level recovery and target cosine pool 192 conditions descriptively. Paired differences are PCA minus random, averaged within each of 16 base-seed clusters and bootstrapped over clusters (20,000 resamples). Objective loss pools the three specified objectives and two capacities; its interval includes zero and is not an equivalence test.

Outcome	Arm-level summaries; paired difference (95% seed-cluster CI)
Recovery	Random: $21/192$ (10.9%); PCA: $183/192$ (95.3%); difference: $+0.844$ $[0.797, 0.891]$
Absolute target cosine	Random: 0.298; PCA: 0.939; difference: $+0.641$ $[0.597, 0.686]$
Objective loss	Random: 0.2648; PCA: 0.2651; difference: $+0.00031$ $[-0.00167, 0.00231]$

Table 8: Post-hoc anchor-source comparison on injected GPT-2/Pythia residual backgrounds. Each model has 48 matched conditions (three known directions, two strengths, eight seed clusters). Recovery is the count with selected-atom cosine ≥ 0.80. Differences are PCA minus random, with 95% bootstrap intervals over the eight seed clusters (20,000 resamples); the cross-model pool is not used for inference.

Model	Outcome	Estimate
GPT-2-small L6	Recovery, random / PCA	$0/48$ / $0/48$ 0.169 / 0.179; $\Delta = +0.0101$
	Mean selected cosine, random / PCA; Δ (PCA$-$random)	95% CI $[0.0022, 0.0185]$ $\Delta = -0.000150$
	Held-out MSE difference, PCA$-$random	95% CI $[-0.000818, 0.000526]$
Pythia-70M L3	Recovery, random / PCA	$1/48$ / $1/48$ 0.085 / 0.074; $\Delta = -0.0108$
	Mean selected cosine, random / PCA; Δ (PCA$-$random)	95% CI $[-0.0734, 0.0413]$ $\Delta = -0.000680$
	Held-out MSE difference, PCA$-$random	95% CI $[-0.001299, 0.000077]$

Target-free spectral-anchor recovery. In a separate target-free protocol, the leading eigenvector of the training second moment supplies a signed anchor pair. One active slot is reserved for this pair, and every residual decoder column is projected onto the anchor's orthogonal complement. The protocol covers three residual objectives, two total dictionary capacities, two geometries, four seeds, and 100k independent validation samples per row. Across all 48 rows, anchor recovery is $48/48$, the minimum observed target cosine is 0.9989, and the family-wise finite-sample certificate lower bound is 0.9405 at 95% confidence. Table 9 reports this result for the structured family in the objective/capacity-uniform theorem; it is not a guarantee for unconstrained SAE training.

Table 9: Target-free spectral-anchor recovery across all residual objectives and capacities. Entries are anchor recovery rates over four seeds, pooling axis and dense geometries; every row inherits the same training-only anchor and the same family-wise lower bound.

Residual objective	d=16, k=2	d=32, k=4	Overall
MSE	$8/8$	$8/8$	$16/16$
MSE+ℓ_1	$8/8$	$8/8$	$16/16$
MSE+ℓ_2	$8/8$	$8/8$	$16/16$
Overall	$24/24$	$24/24$	$48/48$

Bounded-drift spectral-anchor recovery. We ran a separate relaxed v4 protocol in which the training-only spectral anchor is trainable. After each optimizer step, the anchor is projected into a spherical cap of radius $\rho = 0.20$ radians around the reference spectral anchor. Residual decoder columns are normalized subject to the leakage bound $|v_j^\top \widehat{u}| \leq \eta = 0.10$, rather than exact orthogonality. The matrix covers the same three residual objectives, two total capacities, two geometries, four seeds, and 100k independent validation samples per row, for 48 rows. The family-wise spectral estimate gives $s = 0.3397$ and the angular bound $\cos(\arcsin(s) + \rho) = 0.8543$. Table 10 reports the results. This protocol relaxes the v3 constraint but retains an explicit trust region and data–target linkage; it therefore remains a structured-family result rather than a guarantee for unconstrained SAEs.

Table 10: Relaxed bounded-drift spectral-anchor recovery. Entries are recovery counts over eight seed–geometry cases for each objective/capacity pair; the final columns report the minimum target cosine and maximum anchor drift across those cases.

Residual objective	d=16, k=2	d=32, k=4	min. cos.	max. drift
MSE	8/8	8/8	0.9899	0.1287
MSE+ℓ_1	8/8	8/8	0.9930	0.1060
MSE+ℓ_2	8/8	8/8	0.9933	0.1078
Overall	24/24	24/24	0.9899	0.1287

Dose-response and matched-compute baselines. To separate what the constraints guarantee from what training alone achieves at this scale, we pre-specified a dose-response protocol with a fixed MSE objective, d=32, k=4, and nine mechanisms: two frozen arms with $\eta \in \{0, 0.10\}$; five hard trust-region arms with $\eta = 0.10$ for $\rho \in \{0.05, 0.10, 0.40\}$ and $(\rho, \eta) \in \{(0.20, 0), (0.20, 0.10)\}$; one soft-penalty arm at $(\rho, \eta) = (0.20, 0.10)$; and an unconstrained control. Each mechanism uses the same eight independent training-only anchor groups, for $9 \times 8 = 72$ rows, with 20k training samples, 100k validation samples, and 4,000 steps.

Every mechanism recovers all $8/8$ anchor groups, and the worst-group cosine is at or above 0.9788. The mechanisms differ in leakage and loss stability. The unconstrained control reaches a maximum residual–anchor dot product of 0.734 and has the largest, most variable held-out loss (0.1414 ± 0.0077). Every constrained mechanism keeps leakage at or below 0.102 (hard and frozen arms remain within their nominal 0.10 cap; the soft arm reaches 0.10125) and loss at or below 0.1373. The monotone trend is confined to within the $\eta = 0.10$ hard series: the worst-group cosine improves as the trust region tightens; the frozen anchor performs best (0.9983).

The unbounded control's worst-group cosine (0.9901) beats four of the seven constrained arms. Its mean loss is not significantly higher than the frozen-anchor arm at $n = 8$ (Welch $t \approx 1.5$). The clearest differences are leakage and loss variability, not mean loss. Table 11 reports the complete results.

The artifact metadata labels this control `evidence_eligible=false` and identifies it as an anchor-group ablation. Its recovery criterion is also weaker than Table 4's: the reserved anchor slot starts from the spectral estimate, so the two recovery columns are not directly comparable. At this scale, the unconstrained control also recovers all $8/8$ groups. The observed advantage of the constraints is lower leakage and held-out loss variability, not higher recovery.

Table 11: Dose-response and matched-compute baselines. Each mechanism runs over eight independent anchor groups. Entries report the recovery count, worst-group anchor cosine, maximum residual–anchor dot product, and held-out reconstruction loss (mean ± s.d.).

Mechanism	rec.	worst cos.	max leak	val. loss
frozen, $\eta = 0$	8/8	0.9983	0.000	0.1370 ± 0.0014
frozen, $\eta = 0.10$	8/8	0.9983	0.100	0.1367 ± 0.0016
hard ρ=0.05, η=0.10	8/8	0.9964	0.099	0.1369 ± 0.0018
hard ρ=0.10, η=0.10	8/8	0.9910	0.099	0.1367 ± 0.0014
hard ρ=0.20, η=0.10	8/8	0.9856	0.100	0.1370 ± 0.0020
hard ρ=0.20, η=0	8/8	0.9788	0.000	0.1373 ± 0.0018
hard ρ=0.40, η=0.10	8/8	0.9810	0.100	0.1372 ± 0.0018
soft ρ=0.20, η=0.10	8/8	0.9872	0.101	0.1373 ± 0.0017
unbounded	8/8	0.9901	0.734	0.1414 ± 0.0077