DECODER-PRESERVING SPARSE AUTOENCODERS: WHICH READOUTS SURVIVE SPARSE COMPRESSION?

Aniket Deshpande
Department of Physics, University of Illinois Urbana-Champaign
aniket4@illinois.edu

ABSTRACT

Sparse autoencoders (SAEs) compress model activations into sparse codes, but equal reconstruction error and sparsity can preserve different linearly decodable signals. We formalize this ambiguity as a matrix-valued distortion between optimal ridge-prediction operators and train decoder-preserving SAEs by combining this distortion with reconstruction loss. In a rank relaxation, an isotropic task prior saturates per-mode omission costs without changing PCA's ordering, whereas a structured prior can change which modes are retained. A controlled sparse experiment shows that a declared prior protects held-out combinations from its task subspace. On GPT-2 small block 8, DPSAE reduces held-out decoder distortion by 10.6–11.4% across three paired runs while matching reconstruction NMSE. The same checkpoints pass an average natural-text output-KL noninferiority test, but one matched Pythia pair shows no improvement in probes restricted to a few sparse features. These results show that reconstruction quality does not determine which refitted linear readouts survive sparse compression, and that readout preservation is distinct from learning cleaner benchmark concepts or preserving every frozen-model behavior.

github.com/aniket-desh/decoder-preserving-sae

1 INTRODUCTION

Sparse autoencoders (SAEs) compress model activations into sparse feature dictionaries for analysis and intervention (Karvonen et al., 2025). When exact reconstruction is unavailable at a fixed sparsity budget, error must fall somewhere in activation space. MSE assigns capacity by squared displacement in that space. Many analyses instead ask whether linearly recoverable variables remain recoverable after compression. Two reconstructions produced from equally sparse codes can therefore match exactly in MSE while allocating their errors to different decoding tasks.

We ask whether an SAE can preserve a family of linear readouts more effectively without increasing reconstruction MSE. Harvey et al. interpret several representation-similarity measures through agreement across optimal linear readouts (Harvey et al., 2024). For a fixed batch, K_X maps any target vector to the predictions of the optimal regularized linear probe fitted to X. Comparing it with the corresponding operator for a reconstruction measures the taskwise pattern of ridge-prediction distortion. We use this prediction-operator geometry as a sparse-compression objective and analyze the taskwise underdetermination left by MSE and sparsity.

Readout fidelity is therefore task-indexed and matrix-valued. A task second moment weights the decoding tasks that contribute to its scalar summary. We combine decoder disagreement with MSE and fixed sparsity, and call the resulting model a decoder-preserving sparse autoencoder (DPSAE). MSE anchors the reconstruction in the coordinate system expected by the language model. The decoder term biases the sparse bottleneck toward readouts weighted by the task prior. An isotropic prior weights target directions equally, whereas a structured prior can emphasize a declared task subspace.

Figure 1 makes this freedom explicit and summarizes the training objective.

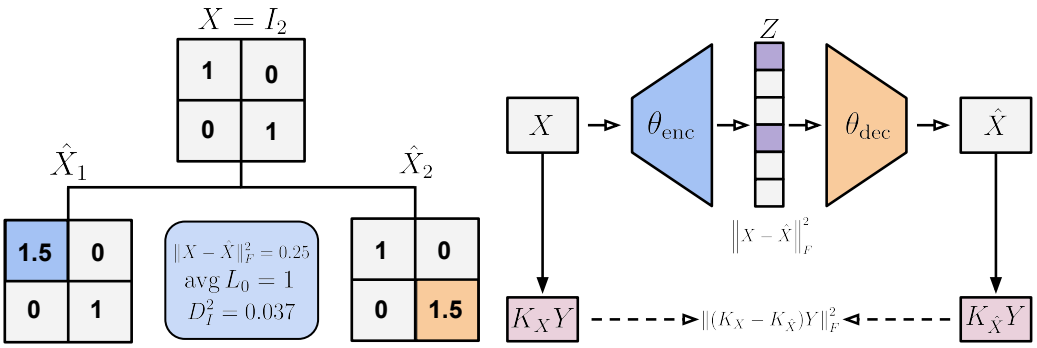

Figure 1: **Equal reconstruction error leaves readout fidelity underdetermined.** Left: with $W = I_2$ and $Z_j = \widehat{X}_j$, the two reconstructions have the same squared Frobenius error $\|X - \widehat{X}_j\|_F^2 = .25$, average sparsity $L_0(Z_j) = 1$, and isotropic decoder distance $D_I^2 = .037$. At $\lambda = .5$, their squared ridge-prediction distortions on targets (e_1, e_2) are $(.037, 0)$ and $(0, .037)$, respectively. Right: DPSAE combines reconstruction error with disagreement between separate ridge refits on X and $\widehat{X}$ using the same target matrix Y.

We make three contributions:

- **Taskwise underdetermination and its rank-relaxed limit.** We prove that equal MSE and code L_0 do not determine ridge-readout fidelity. Under a pure rank bottleneck, isotropic decoder preservation keeps PCA's singular-direction ordering but replaces variance-proportional omission cost with a ridge-saturated cost. Any qualitative departure from this rank-relaxed ordering must involve structure omitted by the relaxation, such as sparse support allocation.
- **Task-selective preservation under declared priors.** A task second moment scalarizes the distortion operator. In the commuting rank relaxation, a structured task prior retains modes according to a closed-form weighted score. In a controlled sparse generator, the prior reduces median distortion on unseen combinations from its declared subspace by 25.3% relative to paired MSE, with no observed NMSE penalty.
- **Language-model evidence.** On GPT-2 small block 8, DPSAE reduces exact finite-group decoder distortion by 10.6–11.4% across three paired runs at matched reconstruction quality. The same pairs pass average natural-text output-KL noninferiority but perform worse on IOI; one matched Pythia pair also shows no gain in sparse concept probes.

These claims concern regularized readouts refitted after compression. The frozen-network and concept-probe results show why they should not be read as evidence for uniform downstream fidelity, cleaner sparse concepts, or cross-model generality.

2 Background

2.1 Sparse autoencoders

Sparse autoencoders learn a dictionary for model activations. An encoder maps an activation $h \in \mathbb{R}^d$ to a sparse code $z \in \mathbb{R}^m$, and a decoder reconstructs $\hat{h}$ as a weighted combination of its dictionary vectors. Language-model applications use the active coordinates of z as candidate features for analysis and intervention (Huben et al., 2024; Bricken et al., 2023).

Standard SAE training balances reconstruction error against sparsity. An ℓ_1 penalty encourages small codes but controls sparsity only indirectly. TopK variants instead retain a fixed number of code entries, which makes the reconstruction–sparsity tradeoff easier to compare across models (Gao et al., 2025). In either case, squared reconstruction error assigns capacity according to Euclidean displacement in activation space. It does not directly measure whether a target remains predictable from the reconstruction.

2.2 LINEAR DECODABILITY AS REPRESENTATION FIDELITY

A linear decoding task fits a readout from a representation matrix $X \in \mathbb{R}^{n\times d}$, with one example per row, to targets defined on the same examples. Ridge regression makes this readout well posed by penalizing its weight norm. Each representation then induces a prediction operator K_X that maps any target vector to the predictions of its optimal ridge readout. Comparing K_X with the operator induced by a reconstruction asks whether the same family of targets remains accessible after refitting.

This prediction-based view also underlies existing representation distances. GULP compares representations through regularized linear prediction tasks and controls differences in their predictive performance (Boix-Adserà et al., 2022). Harvey et al. show that CKA, CCA, and related similarity measures can likewise be interpreted as average alignment between optimal linear readouts over distributions of decoding tasks (Harvey et al., 2024). These are statements about refitted decoders. They do not imply that a frozen downstream network will process a reconstructed activation in the same way, because its weights are not refitted.

3 READOUT FIDELITY UNDER COMPRESSION

3.1 READOUT DISTORTION IS MATRIX-VALUED

Let $X \in \mathbb{R}^{n\times d}$ contain one activation per row. For targets $y \in \mathbb{R}^n$, the prediction of the optimal ridge readout from X is

$$\hat{y}_X = K_X y, \qquad K_X = X(X^\top X + n\lambda I)^{-1}X^\top, \tag{1}$$

where $\lambda > 0$ is the penalty in the average ridge objective $n^{-1}\|Xw - y\|_2^2 + \lambda\|w\|_2^2$. Given a reconstruction $\widehat{X}$, define the signed prediction-operator error and its taskwise distortion operator as

$$\Delta_X(\widehat{X}) = K_X - K_{\widehat{X}}, \qquad \mathcal{A}_X(\widehat{X}) = \Delta_X(\widehat{X})^\top \Delta_X(\widehat{X}). \tag{2}$$

The distortion of a target y is the quadratic form

$$d_{\widehat{X}}(y) = y^\top \mathcal{A}_X(\widehat{X})y = \|K_X y - K_{\widehat{X}}y\|_2^2. \tag{3}$$

For a task distribution with second moment $\Sigma_y = \mathbb{E}[yy^\top] \succeq 0$, its exact decoder distance and source prediction energy are

$$D^2_{\Sigma_y}(X, \widehat{X}) = \operatorname{tr}\Big(\Sigma_y \mathcal{A}_X(\widehat{X})\Big), \quad S_{\Sigma_y}(X) = \operatorname{tr}\big(\Sigma_y K_X^\top K_X\big). \tag{4}$$

When $S_{\Sigma_y}(X) > 0$, the exact relative distortion is $R_{\Sigma_y}(X, \widehat{X}) = D^2_{\Sigma_y}(X, \widehat{X})/S_{\Sigma_y}(X)$. Readout distortion is therefore matrix-valued before a task second moment scalarizes it. The isotropic choice $\Sigma_y = I$ gives $D_I^2 = \|K_X - K_{\widehat{X}}\|_F^2$.

3.2 MSE AND SPARSITY DO NOT DETERMINE TASKWISE FIDELITY

Reserve $Z \in \mathbb{R}^{n\times p}$ for the sparse code and define average per-example sparsity as

$$L_0(Z) = \frac{1}{n}\sum_{i=1}^{n} \|Z_{i,:}\|_0. \tag{5}$$

The freedom left by MSE and L_0 is exact, even in two dimensions.

Proposition 1 (Equal-error taskwise separation)**.** *For every $\lambda > 0$, there exist a source $X \in \mathbb{R}^{2\times 2}$, nonnegative codes $Z_1, Z_2 \in \mathbb{R}_+^{2\times 2}$, and a common linear decoder W with reconstructions $\widehat{X}_j = Z_j W$ such that*

$$L_0(Z_1) = L_0(Z_2) = 1, \qquad \|X - \widehat{X}_1\|_F^2 = \|X - \widehat{X}_2\|_F^2,$$

and $D_I^2(X, \widehat{X}_1) = D_I^2(X, \widehat{X}_2)$. Nevertheless, $\mathcal{A}_X(\widehat{X}_1)$ and $\mathcal{A}_X(\widehat{X}_2)$ are incomparable in the Loewner order. Each reconstruction has lower distortion than the other on some unit target.

The construction changes one diagonal coordinate in each reconstruction. It gives the same MSE, code sparsity, and isotropic average distortion while exchanging which coordinate readout is preserved. Appendix B gives the construction and proof.

For two trained reconstructions $\widehat{X}_M$ and $\widehat{X}_D$, define their DPSAE-advantage operator

$$Q = \mathcal{A}_X(\widehat{X}_M) - \mathcal{A}_X(\widehat{X}_D). \tag{6}$$

Proposition 2 (Taskwise advantage characterization). *For every target y, $d_{\widehat{X}_M}(y) - d_{\widehat{X}_D}(y) = y^\top Q y$. Over unit targets, the smallest and largest advantages are $\lambda_{\min}(Q)$ and $\lambda_{\max}(Q)$. DPSAE has no greater distortion on every target exactly when $Q \succeq 0$, and it has lower Σ_y-weighted distortion exactly when $\mathrm{tr}(\Sigma_y Q) > 0$. If Q is indefinite, rank-one task priors can prefer either reconstruction.*

This proposition is a diagnostic rather than the explanation for why training changes a reconstruction. Its trace measures isotropic average advantage, while its eigenvalues distinguish average improvement from uniform taskwise dominance.

3.3 The rank relaxation isolates what isotropy can change

The simplest relaxation replaces an SAE by any reconstruction of rank at most r. This removes elementwise sparsity and nonlinear support selection while retaining the prediction-operator objective.

Theorem 3 (Isotropic rank relaxation). *Let $X = U \operatorname{diag}(\sigma_1, \dots, \sigma_s) V^\top$ be a compact SVD with $\sigma_1 \geq \cdots \geq \sigma_s > 0$, and define*

$$q_i = \frac{\sigma_i^2}{\sigma_i^2 + n\lambda}.$$

For $0 \leq r \leq s$,

$$\min_{\mathrm{rank}(\widehat{X}) \leq r} \|K_X - K_{\widehat{X}}\|_F^2 = \sum_{i>r} q_i^2. \tag{7}$$

The truncated SVD X_r attains the minimum.

Rank-r MSE and isotropic decoder preservation use the same mode ordering but assign different costs to each omitted mode:

$$c_i^{\mathrm{MSE}} = \sigma_i^2, \qquad c_i^{\mathrm{dec}} = \left(\frac{\sigma_i^2}{\sigma_i^2 + n\lambda}\right)^2. \tag{8}$$

The decoder cost saturates at one once a direction is strongly decodable. Since it remains monotone in σ_i, both objectives retain the same family of optimal top left singular subspaces under a pure rank bottleneck, up to choices within singular-value ties. Isotropic decoder preservation cannot reorder singular directions in this relaxation. Any qualitative departure from this rank-relaxed ordering must involve ingredients omitted by the relaxation, including sparse support, overcompleteness, nonorthogonal features, unequal firing rates, nonlinear selection, minibatch estimation, or optimization.

3.4 A structured task prior can change the retained modes

The monotone ordering above is specific to an isotropic task distribution. A structured prior can change it when the task and representation geometries align.

Corollary 4 (Structured-prior selection). *Suppose $\Sigma_y \succeq 0$ commutes with $XX^\top$. Choose a simultaneous eigenbasis in which*

$$XX^\top = U \operatorname{diag}(\sigma_i^2) U^\top, \qquad \Sigma_y = U \operatorname{diag}(\omega_i) U^\top,$$

and let $q_i = \sigma_i^2/(\sigma_i^2 + n\lambda)$, including $q_i = 0$ on the nullspace of X. For $0 \leq r \leq \mathrm{rank}(X)$, if S_r indexes the r largest scores $\omega_i q_i^2$, then

$$\min_{\mathrm{rank}(\widehat{X}) \leq r} D_{\Sigma_y}^2(X, \widehat{X}) = \sum_{i \notin S_r} \omega_i q_i^2. \tag{9}$$

An optimizer retains the source modes in S_r at their original singular values.

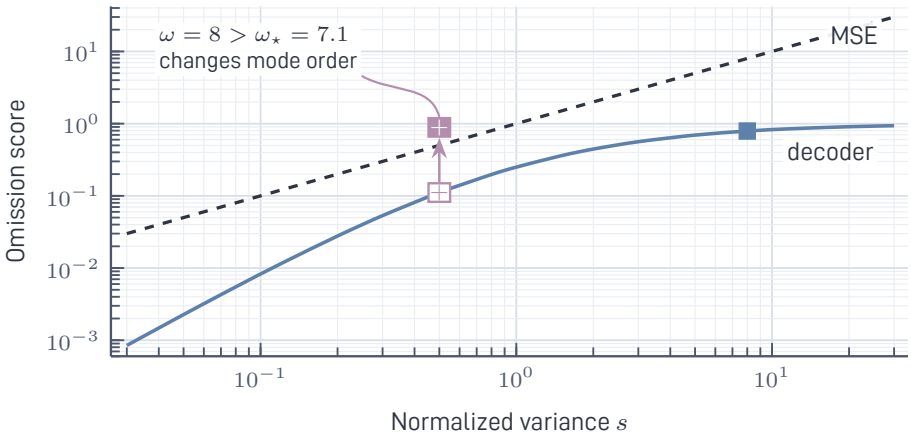

Figure 2: **Isotropic decoder omission cost saturates without changing the rank-relaxed mode order.** MSE omission cost grows with normalized variance s, whereas the isotropic decoder cost approaches one. Both remain monotone in s; the purple markers illustrate how a sufficiently strong structured task weight can cross the mode-selection threshold.

A protected direction j overtakes a nuisance direction i exactly when

$$\omega_j \left(\frac{\sigma_j^2}{\sigma_j^2 + n\lambda} \right)^2 > \omega_i \left(\frac{\sigma_i^2}{\sigma_i^2 + n\lambda} \right)^2 . \tag{10}$$

The exact two-direction construction tests this crossover directly. The sparse generator uses a concatenated stochastic task prior whose expected covariance need not commute with the sample Gram matrix, so the same threshold is only a scale reference there, not a predicted sparse transition. Without the commutation assumption, an optimal weighted rank approximation can mix source modes, and the scalar ranking in Equation 10 need not hold.

3.5 Why the objective remains hybrid

Decoder distance identifies a larger equivalence class than activation reconstruction.

Proposition 5 (Zero decoder distance)**.** *For fixed $\lambda > 0$,*

$$K_X = K_{\widehat{X}} \quad \Longleftrightarrow \quad XX^\top = \widehat{X}\widehat{X}^\top. \tag{11}$$

If $\Sigma_y \succ 0$, these conditions are also equivalent to $D^2_{\Sigma_y}(X, \widehat{X}) = 0$. When X and $\widehat{X}$ have the same feature dimension, equal row Gram matrices imply $\widehat{X} = XQ$ for some orthogonal Q.

Decoder distance can therefore be zero after a feature-coordinate rotation that would change the input seen by the frozen downstream network. MSE anchors a member of this rotation-equivalence class to the original coordinates. The task prior still supplies a finite-group guarantee: for $y \in \mathrm{range}(\Sigma_y)$,

$$\|\Delta_X(\widehat{X})y\|_2^2 \le D^2_{\Sigma_y}(X, \widehat{X})\, y^\top \Sigma_y^\dagger y. \tag{12}$$

This bound controls the protected task ellipsoid on the observed group. It is not a generalization guarantee for new activation samples.

3.6 A scalable decoder-preservation objective

For one geometry group, let $Y_m = m^{-1/2}[y_1, \ldots, y_m]$ contain task samples with $\mathbb{E}[y_j y_j^\top] = \Sigma_y$. We estimate relative decoder distortion as

$$\widehat{R}_m(X, \widehat{X}; Y_m) = \frac{\|\Delta_X(\widehat{X})Y_m\|_F^2}{\max\{\|K_X Y_m\|_F^2, \epsilon\}}. \tag{13}$$

The language-model experiments use fixed-radius Gaussian directions, while the controlled experiments use Rademacher probes and add scaled protected-subspace targets for the structured prior. Language-model training aggregates sampled numerators and denominators across geometry groups before taking their ratio. Identity targets compute the isotropic trace exactly on a fixed finite group. Appendix A.1 defines the corpus aggregations and records the finite-probe, scaling, and denominator-clamp audits.

For encoder E_θ, decoder D_θ, sparse code $Z = E_\theta(X)$, and reconstruction $\widehat{X} = D_\theta(Z)$, the hybrid objective is

$$\mathcal{L}(X) = \alpha \frac{\|X - \widehat{X}\|_F^2}{\|X\|_F^2} + \gamma \widehat{R}_{\mathcal{G},m}(X, \widehat{X}) + \mathcal{L}_{\text{sparse}}(Z), \tag{14}$$

where $\widehat{R}_{\mathcal{G},m}$ is the aggregation used by the corresponding experiment. The Euclidean term anchors coordinates, while the decoder term protects refittable readouts weighted by the declared task prior.

3.7 Scope of the guarantee

All formal claims above concern optimal regularized readouts refitted separately to the original and reconstructed representations. They do not imply that the original model's frozen downstream weights can consume the reconstruction, nor that labeled concepts become concentrated into a few sparse coordinates; Section 5.4 evaluates those questions separately. We also avoid describing the objective as mutual-information maximization or as a Fisher-information metric.

4 Experimental Setup

Controlled sparse generators. The controlled experiments use paired tied signed-TopK dictionaries on overcomplete, nonorthogonal sparse generators. The structured setting places lower-amplitude features in a declared protected subspace and evaluates unseen coefficient combinations from that same subspace. Comparisons share initialization, minibatches, architecture, sparsity, optimizer, and stopping rule. They include MSE, isotropic DPSAE, task-prior DPSAE, a frozen-task weighted-MSE loss, and a row-permuted task-prior control.

Language-model training and selection. The main language-model experiment trains 16,384-latent BatchTopK SAEs with target sparsity $k = 32$ on normalized GPT-2-small residual-stream activations after block 8 (Radford et al., 2019; Bussmann et al., 2024). A disjoint 25M-token sweep selects the smallest decoder-loss weight that reduces exact decoder distortion by at least 10% while keeping NMSE within 1% of paired MSE. We freeze the selected $\gamma = 0.03125$ before training three paired 100M-token runs and evaluating them on a reserved activation range. MSE and DPSAE pairs share initialization, token order, architecture, and optimization.

Metrics, uncertainty, and scope. The primary language-model metrics are exact held-out finite-group decoder distortion, activation NMSE, and inference L_0. Exact evaluation uses 16,384 tokens, groups of 128, identity targets, and 10,000 paired group-bootstrap resamples. Intervals quantify evaluation uncertainty conditional on each trained pair; three paired initializations test optimization repeatability rather than population generality. The frozen-network and Pythia concept evaluations use separately specified held-out data and are interpreted as boundary tests, not consequences of the decoder objective. Appendix A gives complete generators, splits, grids, evaluation geometry, controls, and provenance.

5 Experiments and Results

5.1 A structured prior protects its declared task family

The structured sparse generator tests whether a task covariance can direct capacity toward a lower-amplitude protected subspace. Held-out coefficients and examples are new, but their targets remain within the subspace that defines the prior. Task-prior DPSAE reduces median protected-task distortion from 0.1755 to 0.1332, a 25.3% reduction, and has lower distortion than paired MSE in all ten seeds.

Median NMSE changes from 0.0693 to 0.0667. A one-sided paired Wilcoxon test does not establish an NMSE improvement ($p = 0.097$), so the protected-task gain has no observed reconstruction penalty in this generator. Relative to isotropic DPSAE, the frozen-task loss, and the row-permuted-prior control, task-prior DPSAE reduces protected-task distortion by 18.8%, 16.4%, and 14.1%, respectively.

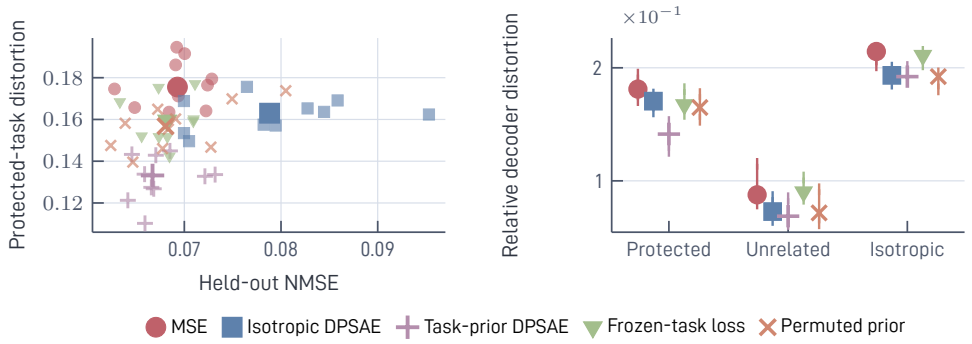

Figure 3: **Task-prior decoder preservation protects held-out combinations from the declared task subspace.** Left: individual paired seeds and medians for protected-task distortion versus reconstruction NMSE. Right: task-family distortions, with horizontally offset points showing medians and vertical whiskers showing the 10th–90th percentiles over ten seeds. Held-out coefficients and examples are fresh, but the targets remain within the protected latent span used to define the training prior.

The protected-task advantage persists across a 16-fold sweep of the prior weight: median reductions relative to paired MSE range from 23.9% to 26.6%, and every 10th-percentile seed reduction remains positive. The sparse generator does not reproduce the crossover from the separate commuting rank construction, so the sweep establishes robustness to task weight rather than a sparse phase transition. Appendix A.9 gives the full sweep, and Appendix A.8 reports the secondary isotropic sparse control.

5.2 DPSAE PRESERVES MORE GPT-2 READOUTS AT MATCHED RECONSTRUCTION QUALITY

The disjoint selection sweep traces the trade-off induced by the decoder weight (Figure 4, left). The predeclared rule selects $\gamma = 0.03125$, the smallest weight that reduces exact decoder distortion by at least 10% while keeping NMSE within 1% of paired MSE, before the three paired 100M-token runs are trained. The sweep uses shorter 25M-token models and a separate activation range, so its points are not pooled with the final evaluation.

At the selected weight, DPSAE has lower decoder distortion and lower NMSE than paired MSE in all three runs (Figure 4, right). Exact held-out decoder-distortion reductions are 10.61–11.38%, with every conditional 95% interval excluding zero. NMSE changes by -0.31 to -0.06%. Thus the tested operating point changes readout fidelity without exchanging it for worse reconstruction.

Inference sparsity remains close to the $k = 32$ training budget: MSE has $L_0 = 30.95$–31.27, and DPSAE has $L_0 = 30.51$–30.80. A tokenwise-TopK control, which retains exactly 32 activations for every token, reduces decoder distortion by 11.50% (95% CI 11.11–11.90%) at a 0.81% NMSE cost. This single 25M-token pair supports the narrower conclusion that the effect is not unique to BatchTopK's global support competition.

The sign of the reduction persists across the tested ridge strengths, group sizes, and grouping schemes. On cached activations, the decoder objective adds 21.3% to the isolated SAE optimization step and about 6 MiB of peak GPU allocation. Appendix A.7 reports these robustness and compute measurements.

5.3 THE READOUT GAIN IS BROAD BUT NOT UNIFORM

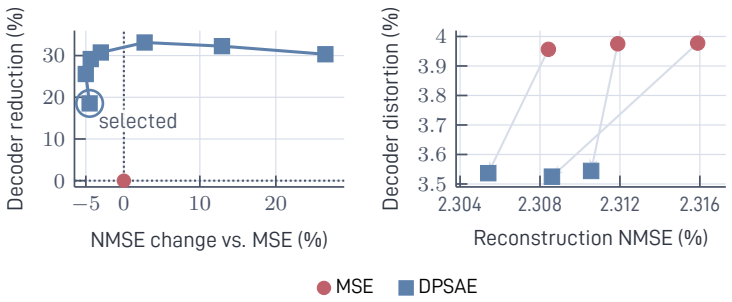

Figure 4: **DPSAE reduces held-out decoder distortion at matched reconstruction quality.** Left: one paired 25M-token run per weight on disjoint activations; the open ring marks the smallest weight that clears the predeclared reconstruction and decoder-distortion thresholds. Right: arrows connect three paired 100M-token MSE and DPSAE runs evaluated on a reserved activation range. Intervals resample held-out geometry groups and are conditional on the trained checkpoints.

Average decoder distortion hides how the gain is distributed across tasks. Every held-out group in all three paired runs has positive trace advantage, but every group also has material positive and negative eigenvalues; none of the advantage operators is positive semidefinite. Among shared random unit targets, 67.0–69.4% materially favor DPSAE, 8.7–10.0% favor MSE, and 21.8–22.9% are unresolved at the declared 5% scale (Figure 5). The improvement therefore spans many directions without providing uniform taskwise dominance. These finite-group directions are not semantic tasks, and Appendix A.4 reports their weak cross-seed recurrence.

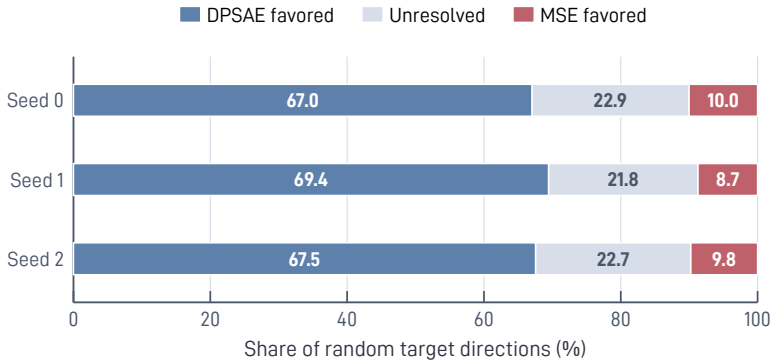

Figure 5: **The average readout gain is broad but not uniform.** Stacked bars show the share of 4,096 shared random target directions per held-out group that materially favor DPSAE, favor MSE, or remain unresolved for each paired run. A direction is material when its absolute advantage exceeds 5% of that group's baseline mean task error. The directions live in finite sample space and should not be interpreted as semantic concepts.

The tested static covariance metrics do not reproduce the DPSAE reduction, but none reaches the matched NMSE regime, so this comparison remains unresolved. Appendix A.3 gives the separately selected high-weight controls and the matched-quality feasibility screen.

5.4 Readout fidelity does not imply cleaner concepts or uniform frozen behavior

Refitted readout fidelity does not guarantee that the original downstream network will process the reconstruction in the same way. We insert each selected GPT-2 pair at block 8 of the frozen model and evaluate 2,048 fresh FineWeb sequences. The DPSAE/MSE output-KL ratios are 0.9905, 0.9744, and 0.9777, with 95% intervals $[0.9880, 0.9929]$, $[0.9720, 0.9768]$, and $[0.9755, 0.9800]$. Every upper endpoint lies below the predeclared 1.01 noninferiority margin (Figure 6). The intervals also lie

below equality, but lower output KL remains secondary because the experiment was designed to test noninferiority.

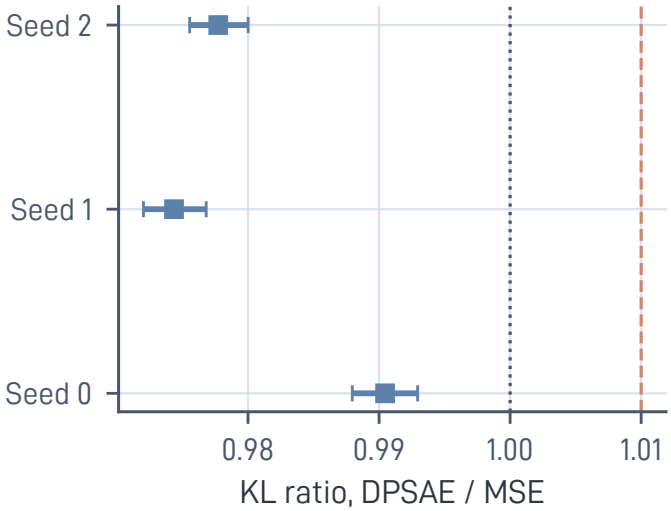

Figure 6: **All three GPT-2 pairs pass natural-text output-KL noninferiority.** Points show $\mathrm{KL}(p_{\mathrm{orig}} \| p_{\mathrm{DPSAE}})/\mathrm{KL}(p_{\mathrm{orig}} \| p_{\mathrm{MSE}})$ on 2,048 fresh FineWeb sequences. Whiskers are sequence-level paired-bootstrap 95% intervals conditional on each trained pair; vertical lines mark equality and the predeclared 1.01 margin.

This compatibility is limited to the average natural-text output metric. On 2,048 separately specified IOI prompts, DPSAE reconstruction has 0.88–2.39 percentage points lower accuracy than paired MSE reconstruction, with every paired-bootstrap interval below zero. Its error in the correct-minus-incorrect name logit difference is also higher in all three pairs. Natural-text output-KL compatibility therefore does not imply uniform preservation of frozen task behavior.

Nor does improved average readout geometry imply that benchmark concepts concentrate into a few sparse coordinates. In one matched Pythia-160M block-8 pair, probes restricted to $k = 5$ selected features have a family-macro DPSAE-minus-MSE AUROC difference of -0.00387 across 60 concept families, with a 95% family-block interval of $[-0.00691, -0.00115]$. All ten probe-seed aggregates are negative. Because the trained checkpoint pair is the replication unit, this is a conditional result rather than evidence for a population-average disadvantage. It nevertheless rules out the stronger interpretation that the measured readout gain automatically improves concentration into one, two, or five benchmark features. Appendix A.5 and Appendix A.6 give the complete boundary evaluations.

6 Related Work

Prediction-based representation comparison. GULP defines representation distances through regularized linear prediction tasks and controls differences in predictive performance across those tasks (Boix-Adserà et al., 2022). Harvey et al. show that CKA, CCA, and related measures can be interpreted as average agreement between optimal linear readouts over distributions of decoding tasks (Harvey et al., 2024). These works provide the prediction-operator geometry used here, but compare representations after they have been formed. Similarity-preserving distillation likewise uses representational relationships to train a compact student (Tung & Mori, 2019). DPSAE instead uses prediction-operator disagreement to train a sparse reconstruction and studies the taskwise freedom left by fixed MSE and sparsity.

SAE objectives and inductive biases. Language-model SAEs usually balance activation reconstruction and sparsity, with recent methods changing the support mechanism or encoder assumptions (Bricken et al., 2023; Gao et al., 2025; Rajamanoharan et al., 2024; Bussmann et al., 2024). MP-SAE

begins from hierarchical and nonlinear representational structure and introduces a residual-guided encoder suited to that structure (Costa et al., 2025). Related work shows more generally that an SAE's assumptions determine which concepts it exposes (Hindupur et al., 2025). Temporal SAE adds a contrastive prior over adjacent tokens (Bhalla et al., 2026), while Aligned SAE uses a cross-modal energy prior that changes the learned dictionary without degrading reconstruction (Dhimoïla et al., 2026). DPSAE studies a different ambiguity under a fixed sparse architecture: equal MSE and L_0 can induce different ridge-readout distortion profiles, and a target covariance weights that profile.

Compression and model selection. Ayonrinde et al. cast SAE selection as a two-part coding problem, preferring the shortest quantized activation-and-decoder description at a fixed fidelity tolerance (Ayonrinde et al., 2024). This addresses the rate and model-selection side of sparse compression. DPSAE instead fixes width and sparsity while changing the task-indexed distortion, without estimating operational description length or claiming MDL optimality.

Task-aware sparse representation learning. Supervised dictionary learning has long optimized sparse representations for specified classification or regression tasks rather than reconstruction alone (Mairal et al., 2012). Recent LLM methods bring related supervision into SAEs. Guided SAE binds labeled concepts to designated latent features (Härle et al., 2025); ClassifSAE jointly trains a classifier to concentrate task-relevant information in a sparse subspace (Le Bail et al., 2026); and AdaptiveK uses a supervised complexity probe to allocate input-dependent sparsity (Yao et al., 2026). These methods target specified labeled quantities. DPSAE instead weights a distribution of readouts that are refitted separately to the original and reconstructed activations. Its isotropic language-model objective requires no task labels.

Frozen-model objectives. End-to-end sparse dictionary learning adds output-distribution KL to activation reconstruction so that an SAE preserves the behavior of the frozen downstream model (Braun et al., 2024). A short KL-and-MSE fine-tune can recover much of this benefit at lower cost (Karvonen, 2025). These objectives preserve one realized downstream computation. DPSAE makes no frozen-network guarantee, so we test compatibility empirically rather than treating it as a consequence of the objective; it preserves a specified distribution of refittable regularized linear readouts. To our knowledge, prior SAE objectives have not directly penalized disagreement between the regularized prediction operators induced by original and reconstructed activations.

7 Discussion, Limitations, and Future Work

These results support a narrow but useful view of sparse compression: reconstruction error and sparsity leave substantial freedom in which refittable readouts survive, which DPSAE makes explicit by assigning a cost to disagreement between regularized prediction operators. A structured prior can steer capacity toward its declared task family in a controlled sparse generator, while the isotropic objective reduces held-out finite-group readout distortion by 10.6–11.4% across three paired GPT-2 runs at matched reconstruction quality. This improvement is broad across sampled directions but is not uniform, and its meaning should remain distinct from other desirable properties of an SAE. The same GPT-2 reconstructions satisfy natural-text output-KL noninferiority, yet they perform worse on IOI, and one matched Pythia pair shows no improvement in concentration of benchmark concepts into a few sparse features. Refittable readout fidelity, compatibility with a frozen downstream computation, and sparse concept recovery are therefore separate empirical properties rather than interchangeable notions of representation quality.

The evidence also leaves important limits unresolved. The main language-model result concerns one layer of GPT-2 small and three paired initializations, so its intervals quantify evaluation uncertainty conditional on those trained checkpoints rather than population-level generality. The structured-prior result uses a synthetic task family, the taskwise audit studies finite-sample directions rather than semantic tasks, the static covariance controls never reach a matched reconstruction regime, and the Pythia concept result is conditional on one checkpoint pair. Future work will replicate the readout effect across layers, model families, widths, and sparse architectures; construct task priors from independently specified real prediction families; and compare adaptive decoder preservation with static baselines along genuinely matched reconstruction–sparsity frontiers. It would also be useful to test nonlinear and frozen readouts, connect readout preservation to causal interventions and feature

semantics, and measure an operational rate that includes sparse supports, amplitudes, and dictionary cost. That broader rate–distortion view could determine when the additional readout fidelity is worth the representation complexity required to preserve it.

REFERENCES

Kola Ayonrinde, Michael T. Pearce, and Lee Sharkey. Interpretability as compression: Reconsidering SAE explanations of neural activations with MDL-SAEs, October 2024. URL https://arxiv.org/abs/2410.11179.

Usha Bhalla, Alex Oesterling, Claudio Mayrink Verdun, Himabindu Lakkaraju, and Flavio P. Calmon. Temporal sparse autoencoders: Leveraging the sequential nature of language for interpretability. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=bojVI4l9Kn.

Enric Boix-Adserà, Hannah Lawrence, George Stepaniants, and Philippe Rigollet. GULP: A prediction-based metric between representations. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/2f0435cffef91068ced08d7c7d8e643e-Abstract-Conference.html.

Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. Identifying functionally important features with end-to-end sparse dictionary learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3408. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c212c1b88395ab68d5e1671c17883ec6-Abstract-Conference.html.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK sparse autoencoders, 2024. URL https://arxiv.org/abs/2412.06410.

Valérie Costa, Thomas Fel, Ekdeep S. Lubana, Bahareh Tolooshams, and Demba Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/311bd32bcd952f93cf820bacd6955f0a-Abstract-Conference.html.

Grégoire Dhimoïla, Thomas Fel, Victor Boutin, and Agustin Picard. Cross-modal redundancy and the geometry of vision–language embeddings. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10009130.

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 *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/42ef3308c230942d223c411adf182c88-Abstract-Conference.html.

Ruben Härle, Felix Friedrich, Manuel Brack, Björn Deiseroth, Stephan Waeldchen, Patrick Schramowski, and Kristian Kersting. Measuring and guiding monosemanticity. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f4daa773a5bb2d562a9204a7e2225a67-Abstract-Conference.html.

Sarah E. Harvey, David Lipshutz, and Alex H. Williams. What representational similarity measures imply about decodable information. In *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*, volume 285 of *Proceedings of Machine Learning Research*, pp. 140–151. PMLR, 2024. URL https://proceedings.mlr.press/v285/harvey24a.html.

Sai Sumedh R. Hindupur, Ekdeep S. Lubana, Thomas Fel, and Demba Ba. Projecting assumptions: The duality between sparse autoencoders and concept geometry. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/110d919b4a711f25962a7cd5961f4955-Abstract-Conference.html.

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://proceedings.iclr.cc/paper_files/paper/2024/hash/1fa1ab11f4bd5f94b2ec20e794dbfa3b-Abstract-Conference.html.

Adam Karvonen. Revisiting end-to-end sparse autoencoder training: A short finetune is all you need, 2025. URL https://arxiv.org/abs/2503.17272.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability, 2025. URL https://arxiv.org/abs/2503.09532.

Mathis Le Bail, Jérémie Dentan, Davide Buscaldi, and Sonia Vanier. Unveiling decision-making in LLMs for text classification: Extraction of influential and interpretable concepts with sparse autoencoders. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 2477–2504. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-eacl.129. URL https://aclanthology.org/2026.findings-eacl.129/.

Julien Mairal, Francis Bach, and Jean Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804, 2012. doi: 10.1109/TPAMI.2011.156. URL https://arxiv.org/abs/1009.5358.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0970. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/370df50ccfdf8bde18f8f9c2d9151bda-Abstract-Datasets_and_Benchmarks_Track.html.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, Janos Kramar, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders, 2024. URL https://arxiv.org/abs/2407.14435.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1365–1374, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Tung_Similarity-Preserving_Knowledge_Distillation_ICCV_2019_paper.html.

Yifei Yao, Hanrong Zhang, and Mengnan Du. AdaptiveK: Complexity-driven sparse autoencoders for interpretable language model representations. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23702–23728. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.findings-acl.1187/.

A Experimental Details

A.1 Finite-probe estimator and corpus aggregation

For geometry groups $\mathcal{G}$, write $D_g^2 = D^2_{\Sigma_{y,g}}(X_g, \widehat{X}_g)$ and $S_g = S_{\Sigma_{y,g}}(X_g)$. Assuming $\sum_g S_g > 0$ for the first expression and $S_g > 0$ for every group in the second, the two exact corpus summaries are

$$R_{\mathcal{G}}^{\mathrm{sum}} = \frac{\sum_{g\in\mathcal{G}} D_g^2}{\sum_{g\in\mathcal{G}} S_g}, \qquad R_{\mathcal{G}}^{\mathrm{mean}} = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} \frac{D_g^2}{S_g}. \tag{15}$$

For sampled group statistics $\widehat{D}^2_{g,m} = \|\Delta_{X_g}(\widehat{X}_g) Y_{g,m}\|_F^2$ and $\widehat{S}_{g,m} = \|K_{X_g} Y_{g,m}\|_F^2$, the corresponding estimators are

$$\widehat{R}^{\mathrm{sum}}_{\mathcal{G},m} = \frac{\sum_g \widehat{D}^2_{g,m}}{\max\{\sum_g \widehat{S}_{g,m}, \epsilon\}}, \qquad \widehat{R}^{\mathrm{mean}}_{\mathcal{G},m} = \frac{1}{|\mathcal{G}|}\sum_g \frac{\widehat{D}^2_{g,m}}{\max\{\widehat{S}_{g,m}, \epsilon\}}. \tag{16}$$

Language-model training and headline evaluation use the ratio of sums; the controlled synthetic experiments average groupwise ratios. Identity targets compute the isotropic trace exactly for a fixed group, not for a population over activations.

The language-model implementation draws $g_j \sim \mathcal{N}(0, I_n)$ and uses $y_j = \sqrt{n}\, g_j / \|g_j\|_2$. It omits the analytic $m^{-1/2}$ scale, which cancels in the ratio whenever the denominator clamp is inactive; no clamp activation occurs in the reported audit. The normalized numerator is unbiased for D_I^2, but the finite-m self-normalized ratio is not exactly unbiased. Appendix B gives the fixed-radius trace-estimator variance.

For the isotropic ratio-of-sums comparison, let $R_M = R_{\mathcal{G}}^{\mathrm{sum}}(X, \widehat{X}_M)$ and $R_D = R_{\mathcal{G}}^{\mathrm{sum}}(X, \widehat{X}_D)$. Proposition 2 gives

$$R_M - R_D = \frac{\sum_g \operatorname{tr} Q_g}{\sum_g \|K_{X_g}\|_F^2}. \tag{17}$$

We calibrate ridge through the effective decoder degrees of freedom $\operatorname{tr}(K_X)/n$, using a target of 0.25 in the language-model experiments. Batched FP32 Cholesky solves apply the operators without materializing a task covariance.

A.2 Language-model protocol and paired GPT-2 evaluation

Language-model representation and preprocessing. The main GPT-2 experiment uses `openai-community/gpt2`, with 768-dimensional residual-stream activations after block 8. A fixed training-calibration mean vector and one global RMS scalar normalize all later activations. Geometry groups are not centered independently. We calibrate the ridge penalty by a target effective degrees-of-freedom fraction, $\operatorname{tr}(K_X)/n$. The untied nonnegative BatchTopK SAE has 16,384 latents, a learned decoder bias, unit-normalized decoder columns, ReLU preactivations, and a global training budget of $k = 32$ activations per token on average. SAE encoding and decoding use BF16 autocast on CUDA, while reconstructions passed to the decoder objective and all ridge solves use FP32.

Data splits and model selection. Decoder-weight and architecture choices use the designated 180M–185M FineWeb range (Penedo et al., 2024). The clean readout evaluation uses the disjoint 190M–195M range. A frozen-network diagnostic opened 195M–200M once and did not enter model or hyperparameter selection. After the output-KL metric, 1.01 margin, sequence sampler, sample size, bootstrap, and secondary metrics were fixed, the frozen-network evaluation opened the fresh 200M–210M range. Exact readout evaluation uses 16,384 tokens, groups of 128, calibrated ridge, identity targets, and 10,000 paired group-bootstrap resamples.

Paired GPT-2 confirmation. The selected $\gamma = 0.03125$ is frozen before training the three paired runs. Each model sees 100,001,792 training tokens. Initialization seeds are 0, 1, and 2, while token order and fixed-radius isotropic probe streams are matched within each pair. Exact evaluation uses ridge $\lambda = 1.6049035191535947$ and reports the ratio-of-sums distortion $\sum_g \|K(X_g) - K(\widehat{X}_g)\|_F^2 / \sum_g \|K(X_g)\|_F^2$. Seedwise reductions and conditional 95% intervals

are 10.61% [10.15, 11.07], 11.38% [10.95, 11.82], and 10.84% [10.28, 11.40]. The complete run uses clean revision `79607d4`; its run manifest records code, configuration, cache, checkpoint, and evaluation hashes.

Decoder-weight selection sweep. The selection sweep trains paired 25M-token models at $\gamma \in \{0.03125, 0.0625, 0.09375, 0.125, 0.25, 0.5, 1\}$ with matched initialization, token order, probe stream, optimizer, and $k = 32$ budget. The predeclared rule chooses the smallest weight with at least 10% lower exact decoder distortion and NMSE no greater than 1.01 times MSE. This selects $\gamma = 0.03125$ for the paired 100M-token evaluation; the full sweep remains a selection-set trade-off curve.

A.3 Static covariance controls

For calibration covariance C, the whitening operator is a floored inverse covariance square root. The static spectral operator is the square root of $M_\lambda = C(C + \lambda I)^{-2}$, so its residual norm implements the rank-relaxation weighting. Each coefficient is selected from $\{0.125, 0.25, 0.5, 1\}$ under an NMSE cap on fresh activations, then evaluated across three paired 100M-token runs on 185M–190M. These control runs use revision `38b4c3b`; the Exp08 manifest pins their result artifact and configuration hash.

At $\gamma = 0.25$, DPSAE reduces decoder distortion by 24.23–24.39% at a 6.91–7.11% NMSE cost. The selected static spectral loss changes distortion by 0.37–0.91% at a 0.85–0.92% NMSE cost, while selected whitening worsens distortion by 12.16–12.57%. These points were selected separately and are not matched in NMSE, so they do not identify a like-for-like effect comparison (Figure 8, left).

The later matched-quality screen is a separate preregistered feasibility test. Its $[1.06, 1.08]$ band targets the roughly 7% NMSE cost of the earlier high-weight DPSAE runs. A co-trained 25M-token MSE/DPSAE anchor has an NMSE ratio of 1.0285 and a 32.51% decoder-distortion reduction. The five static spectral coefficients $\beta \in \{2, 4, 8, 16, 32\}$ produce NMSE ratios 0.9061, 0.8986, 0.8952, 0.8959, and 0.8974, all outside the specified band (Figure 7). None reaches the regime needed for a matched comparison, leaving the static-baseline question unresolved.

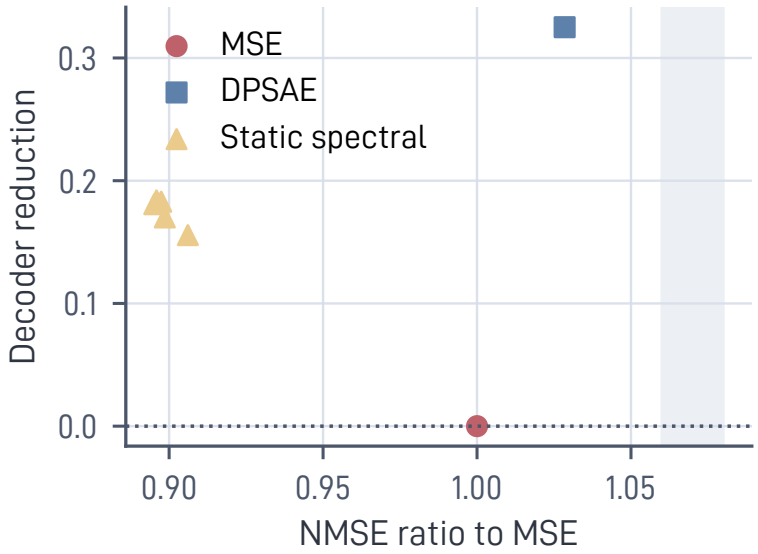

Figure 7: **The tested static spectral coefficients do not reach the matched-NMSE target.** The shaded band marks the predeclared $[1.06, 1.08]$ DPSAE/MSE NMSE-ratio range. All five static candidates fall outside it, so the screen cannot compare a static metric with DPSAE at matched reconstruction quality.

A.4 TASKWISE ADVANTAGE AUDIT

The exact spectrum audit reuses the 16,384 held-out tokens, 128 groups of 128, and the calibrated ridge. For each group we form Q_g in float64, check symmetry and trace reconciliation, and evaluate 4,096 random unit directions drawn once and reused across runs. Materiality is an absolute task advantage above 5% of the group's baseline mean task error. A shared direction materially favors DPSAE in all three runs with probability 32.73%, versus 31.78% under conditional independence; pairwise score correlations are 0.054–0.075. These weak recurrence statistics do not identify a stable semantic subspace. The source summary is `artifacts/exp08_experiment_figure/task_spectrum/advantage_spectrum_summary.json`; compact shares are versioned in `data/fig_lm_task_shares.csv`.

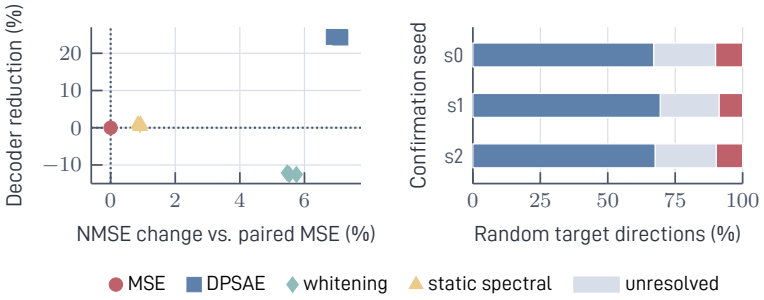

Figure 8: **Static covariance controls remain unmatched, while the average DPSAE advantage is taskwise mixed.** Left: separately selected high-weight DPSAE, whitening, and static spectral points; these controls do not share an NMSE operating point. Right: random-direction shares from the three paired GPT-2 runs. The main text uses the enlarged taskwise panel in Figure 5.

A.5 FROZEN-NETWORK COMPATIBILITY

The evaluation inserts each of the three selected 100M-token GPT-2 BatchTopK pairs at the original block-8 residual-stream site and leaves every downstream parameter fixed. It samples 2,048 nonoverlapping length-256 sequences by a deterministic permutation of aligned blocks from FineWeb positions 200M–210M. The primary metric is the ratio $\sum_t \mathrm{KL}(p_{\mathrm{orig},t} \| p_{\mathrm{DPSAE},t}) / \sum_t \mathrm{KL}(p_{\mathrm{orig},t} \| p_{\mathrm{MSE},t})$ over evaluated tokens. Sequence-level paired bootstraps use 10,000 draws, and every run must have an upper 95% endpoint below 1.01. An identity hook reproduces all logits exactly; the mean-ablation control replaces the normalized activation with zero before denormalization.

All three natural-text pairs meet the noninferiority margin, with slightly better activation NMSE for DPSAE and effectively matched inference sparsity (Table 1). The inference-L_0 differences are $+0.0082$, $+0.0010$, and -0.0106 activations for seeds 0–2. Cross-entropy, top-1 agreement, and next-token accuracy are directionally favorable or unresolved on natural text, but they remain secondary to the frozen KL ratio.

The controlled IOI evaluation uses 2,048 held-out prompts generated before evaluation. The original model is correct on 99.66% of them. DPSAE reconstruction is less accurate than paired MSE reconstruction in every run, and its absolute error in the correct-minus-incorrect name logit difference is higher by 0.1206, 0.0759, and 0.1085, with all prompt-bootstrap intervals above zero. The result shows why average natural-text output-KL compatibility should not be restated as uniform frozen-behavior preservation.

Table 1: **Natural-text output-KL noninferiority passes, while the secondary IOI accuracy endpoint worsens.** The output-KL ratio is $\mathrm{KL}(p_{\mathrm{orig}} \| p_{\mathrm{DPSAE}}) / \mathrm{KL}(p_{\mathrm{orig}} \| p_{\mathrm{MSE}})$; the IOI accuracy difference is DPSAE minus MSE. Intervals resample natural sequences or IOI prompts and are conditional on each trained checkpoint pair.

Seed	Natural KL ratio [95% CI]	Activation NMSE ratio	IOI Δaccuracy (pp) [95% CI]
0	0.9905 [0.9880, 0.9929]	0.9975	-2.39 [-3.22, -1.61]
1	0.9744 [0.9720, 0.9768]	0.9967	-0.88 [-1.56, -0.20]
2	0.9777 [0.9755, 0.9800]	0.9988	-1.51 [-2.20, -0.88]

A.6 Standard-concept concentration pilot

The evaluation reuses one 25M-token Pythia-160M-deduped block-8 BatchTopK pair with width 16,384, target $L_0 = 32$, and decoder weight $\gamma = 0.03125$. Its evaluation-split DPSAE/MSE NMSE ratio is 1.0031; the methods' relative L_0 errors are 0.00117 and 0.00144, and their paired relative L_0 difference is 0.00027. The pair is therefore matched closely enough in reconstruction and sparsity to test concept concentration.

The benchmark contains 113 frozen labeled tasks grouped into 60 concept families. Sparse probes use the pinned `sae-probes` L1 procedure at $k \in \{1, 2, 5\}$ and ten probe seeds. Companion L2 logistic probes evaluate the original residual stream, each full reconstruction, and each full sparse code. The primary analysis first averages probe seeds within each task, then computes the family-macro DPSAE-minus-MSE test-AUROC difference at $k = 5$ and resamples the 60 family blocks 10,000 times. Missing or extra task–k–seed–method cells fail the run rather than being dropped.

The primary difference is -0.003865, with a 95% family-block interval of $[-0.006910, -0.001150]$. All ten probe-seed macro differences are negative, from -0.006643 to -0.000893, while 29 of 60 individual family means are positive. Full-reconstruction and full-code differences are near zero, and the sparse representation ladder in Figure 9 shows no DPSAE concentration gain at $k = 5, 2$, or 1. Because the preregistered primary aggregate did not improve, we did not run additional training seeds, context mining, or API feature labeling.

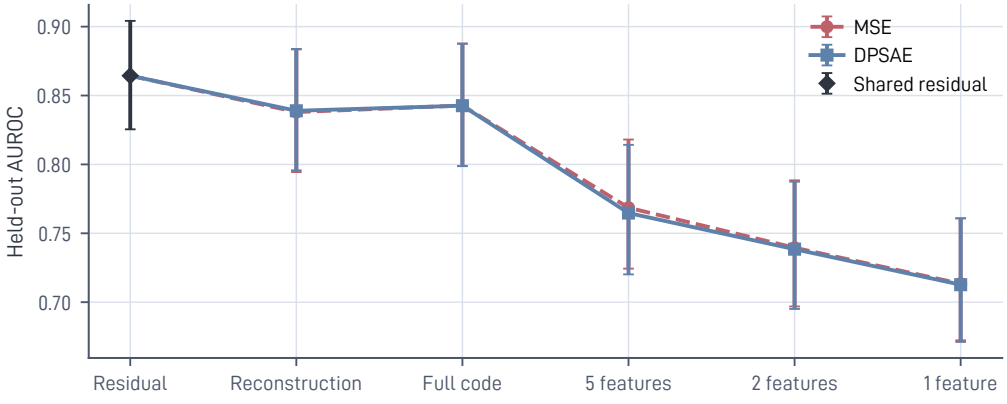

Figure 9: **One matched Pythia pair does not improve standard sparse-concept probes.** The representation ladder separates original-activation recoverability, full-reconstruction and full-code preservation, and concentration into $k = 5, 2, 1$ selected features. Intervals resample the fixed concept-family blocks after averaging the ten probe seeds within each task. Because no additional training pairs were evaluated, the result is conditional on this checkpoint pair.

A.7 Evaluation geometry and computational cost

Geometry audit. The audit reuses the three paired GPT-2 checkpoints and changes one evaluation choice at a time. Ridge penalties target effective-DoF fractions 0.125, 0.25, and 0.5; group-size checks use 64, 128, and 256 tokens; grouping checks compare contiguous, globally shuffled, and document-balanced groups at size 128. Every seed retains a positive paired reduction in every condition. Because group size changes the finite-group operator itself, the resulting magnitudes should not be compared as measurements of one fixed quantity.

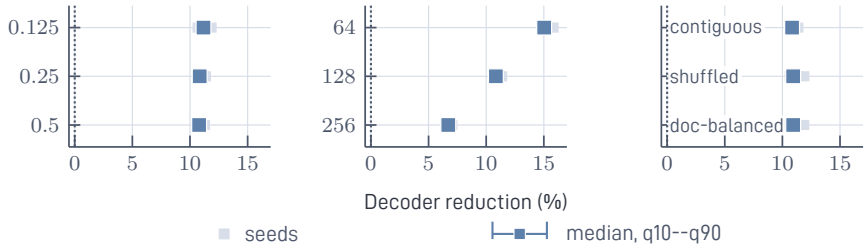

Figure 10: **The sign of the DPSAE advantage survives the tested evaluation geometries.** Points show paired reductions and bars show the median with 10th–90th percentiles. Left: ridge effective-DoF fraction. Center: geometry-group size. Right: contiguous, shuffled, and document-balanced grouping. These are evaluation-only checks on fixed checkpoints; changing group size or ridge changes the measured operator.

Computational cost. The isolated optimization benchmark uses cached normalized activation batches, so it excludes the shared language-model forward pass. Median step time is 6.494 ms for MSE and 7.878 ms for DPSAE, a 21.3% increase. Corresponding throughput is 315,385 and 259,964 activation tokens per second. Peak allocated memory is 1.3422 and 1.3485 GiB, respectively. The reported robustness and optimization suite uses 1.113 active single-GPU hours, including 1.075 hours of SAE training; queue idle time and figure rendering are excluded.

Single-seed generality diagnostics. GPT-2 block 4 clears the 10% decoder-reduction threshold in one paired diagnostic, while Pythia block 8 reaches a 9.13% reduction. These screens use neither three paired runs nor the complete matched-quality protocol, so they do not establish cross-layer or cross-model generality.

A.8 Isotropic sparse control

In the rank-24 numerical construction, truncated SVD matches the decoder-distortion tail in Theorem 3 at every retained rank, with maximum absolute error 2.22×10^{-15}. This verifies the implementation rather than adding empirical evidence for the theorem. Replacing the rank constraint with a TopK bottleneck on an overcomplete, nonorthogonal sparse generator changes the allocation: isotropic DPSAE lowers decoder distortion relative to paired MSE in all ten seeds, with a median reduction of 15.5% and a range of 8.9–21.7%, while median NMSE increases by 12.4%. Thus sparsity permits departures from the rank-relaxed allocation, but this isotropic control does not produce a matched-reconstruction improvement or identify a privileged low-variance subspace.

A.9 Structured-prior weight sweep

The sparse generator uses the crossover from the commuting two-direction rank construction only to set a dimensionless weight scale. Across relative weights 0.25 to 4, median protected-task reductions remain between 23.9% and 26.6%, but neither a crossover nor a monotone response appears. The result shows that the protected-task advantage is stable across this range; it does not confirm that the sparse noncommuting system inherits the rank threshold.

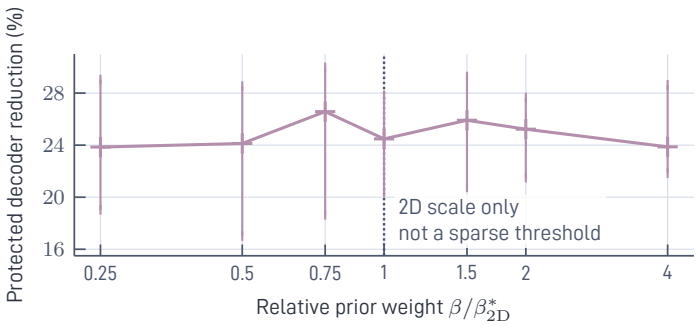

Figure 11: **Structured-prior protection is stable across task weights, without a sparse crossover.** Points show median protected-task distortion reduction relative to paired MSE over ten seeds; whiskers span the 10th–90th percentiles. The vertical reference marks the crossover scale from the separate commuting two-direction rank construction, not a predicted transition for this sparse generator.

A.10 Optimization and architecture audits

Finite-probe gradient audit. For each of the six confirmation checkpoints, the audit compares gradients from 16-probe banks against identity-target reference gradients in both reconstruction and row-Gram coordinates. It uses 256 independent banks per batch and paired self-normalized and fixed-denominator estimates. The conservative columns below take the worse of the two coordinate systems. "Expected UCB" is the familywise paired confidence upper bound on sampled-minus-fixed gradient error; "finite error" and cosine describe the 256-bank mean rather than one 16-probe update.

Table 2: **The finite-probe decoder gradient is expectation-aligned but noisy at finite sample size.** Across all checkpoints, the familywise upper confidence bound on expected gradient error remains below 0.53%, whereas the 256-bank mean retains 13.8–14.1% relative error.

Checkpoint	Median UCB (%)	P90 UCB (%)	Finite error (%)	256-bank mean cosine
MSE s0	0.470	0.528	14.12	0.9902
DPSAE s0	0.464	0.522	13.80	0.9906
MSE s1	0.470	0.529	14.14	0.9902
DPSAE s1	0.464	0.521	13.82	0.9906
MSE s2	0.470	0.528	14.12	0.9902
DPSAE s2	0.464	0.521	13.87	0.9905

The denominator coefficient of variation is 1.143% at the median and 1.216% at the 90th percentile, with no denominator or target-RMS clamp activations. Mean error follows the expected $m^{-1/2}$ Monte Carlo rate. This audit applies before the active-set Jacobian, MSE mixing, γ scaling, gradient clipping, and Adam, so it does not certify the complete optimizer update.

Matched-NMSE and architecture screens. The architecture screen replaces BatchTopK with tokenwise TopK or attempts a gated JumpReLU extension. Tokenwise TopK keeps 32 preactivations per token and uses the selected decoder weight in a single paired 25M-token run.

JumpReLU validity screen. The JumpReLU extension uses $z_j = a_j \mathbf{1}\{a_j > \theta_j\}$ and a shared penalty $(\|z\|_0/k - 1)^2$. We parameterize thresholds in log space and train them with rectangular-kernel pseudo-gradients (Rajamanoharan et al., 2024). Scores and threshold gradients use FP32, and a one-time first-batch quantile sets the initial average $L_0 = k$.

The final JumpReLU validity study trains the paired objectives for 25,001,984 tokens at shared sparsity-loss weights $2, 4, 8$, and 16. Selection uses only held-out L_0, late-window drift, finite values, threshold stability, dead-feature fraction, and run provenance; NMSE, decoder distortion, and language-model loss are not inspected. The respective DPSAE/MSE L_0 values are $37.73/38.55$,

37.38/36.93, 36.10/36.11, and 35.88/35.61. Every setting falls outside the prespecified $[30.4, 33.6]$ band, and maximum dead-feature fractions range from 48.5% to 54.4%, above the prespecified 10% maximum. No weight is selected, so the JumpReLU comparison remains unidentified rather than providing evidence for or against DPSAE.

Reproducibility and provenance. The paired GPT-2 evaluation, gamma sweep, task spectrum, robustness audit, and structured-prior sweep use revision `79607d4`. Their versioned run manifest records code, configuration, input, checkpoint, and result hashes. The static-control runs use revision `38b4c3b`; the later matched-quality screen and frozen-network evaluation use revision `4fe3e91`. The gradient audit uses revision `46eb22a`, and the full-horizon JumpReLU study uses revision `6a2f338`.

The release is bound to core-manifest SHA-256 prefix `db9deb08d28c` and publication-payload SHA-256 prefix `15887dd617c6`. The versioned files `data/arxiv_closure_payload.json`, `data/arxiv_closure_payload_manifest.json`, `data/arxiv_closure_summary.csv`, and `data/arxiv_closure_figure_manifest.json` retain the full digests, experiment decisions, and exact figure hashes. The secondary record `data/exp09_frozen_network_secondary_summary.json` binds the natural-text and IOI values above to their full-result hashes. Generated checkpoints and full per-example artifacts remain outside the manuscript repository.

B Proofs for the Readout-Distortion Results

Proof of Proposition 1. Fix $\lambda > 0$, let $\tau = 2\lambda$, and choose $a, \delta > 0$. Set

$$X = aI_2, \qquad W = I_2, \qquad Z_1 = \begin{pmatrix} a+\delta & 0 \\ 0 & a \end{pmatrix}, \qquad Z_2 = \begin{pmatrix} a & 0 \\ 0 & a+\delta \end{pmatrix}.$$

The reconstructions are $\widehat{X}_j = Z_j W = Z_j$. Each code has one nonzero entry per row, so $L_0(Z_1) = L_0(Z_2) = 1$, and both reconstruction errors equal δ^2. Define

$$\kappa(t) = \frac{t^2}{t^2+\tau}, \qquad \eta = \kappa(a+\delta) - \kappa(a) > 0.$$

Direct substitution into Equation 1 gives

$$K_X = \kappa(a) I_2, \quad \mathcal{A}_X(\widehat{X}_1) = \begin{pmatrix} \eta^2 & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathcal{A}_X(\widehat{X}_2) = \begin{pmatrix} 0 & 0 \\ 0 & \eta^2 \end{pmatrix}.$$

Their traces, and hence their isotropic decoder distances, agree. Their difference is $\operatorname{diag}(\eta^2, -\eta^2)$, which is indefinite. The first reconstruction has lower distortion on e_2, and the second has lower distortion on e_1. □

Proof of Proposition 2. Equation 3 gives

$$d_{\widehat{X}_M}(y) - d_{\widehat{X}_D}(y) = y^\top \Big(\mathcal{A}_X(\widehat{X}_M) - \mathcal{A}_X(\widehat{X}_D) \Big) y = y^\top Q y.$$

Rayleigh–Ritz gives the extrema over unit targets, and nonnegativity for every target is equivalent to $Q \succeq 0$. Taking expectation under $\mathbb{E}[yy^\top] = \Sigma_y$ gives

$$\mathbb{E}[y^\top Q y] = \operatorname{tr}(\Sigma_y Q).$$

If Q is indefinite, unit eigenvectors associated with a positive and a negative eigenvalue define rank-one task priors that prefer opposite reconstructions. □

Proof of Theorem 3. Write $\tau = n\lambda$. The compact SVD of X gives

$$K_X = U \operatorname{diag}\left(\frac{\sigma_1^2}{\sigma_1^2+\tau}, \dots, \frac{\sigma_s^2}{\sigma_s^2+\tau} \right) U^\top = U \operatorname{diag}(q_1, \dots, q_s) U^\top. \tag{18}$$

For any $\widehat{X}$ with rank at most r, $K_{\widehat{X}}$ also has rank at most r. Eckart–Young therefore gives

$$\|K_X - K_{\widehat{X}}\|_F^2 \geq \min_{\operatorname{rank}(M) \leq r} \|K_X - M\|_F^2 = \sum_{i>r} q_i^2.$$

The truncated SVD $X_r = U_r \operatorname{diag}(\sigma_1, \ldots, \sigma_r) V_r^\top$ satisfies

$$K_{X_r} = U_r \operatorname{diag}(q_1, \ldots, q_r) U_r^\top,$$

so it attains this lower bound. □

When $0 < r < s$ and $q_r > q_{r+1}$, the best rank-r approximation to K_X is unique. Every optimizer then satisfies

$$K_{\widehat{X}} = U_r \operatorname{diag}(q_1, \ldots, q_r) U_r^\top, \qquad \widehat{X}\widehat{X}^\top = U_r \operatorname{diag}(\sigma_1^2, \ldots, \sigma_r^2) U_r^\top.$$

With the same feature dimension, this is equivalent to $\widehat{X} = X_r Q$ for some orthogonal Q. At a tied cutoff, any corresponding subspace within the tied left-singular eigenspace is optimal.

Proof of Corollary 4. Because $XX^\top$ and Σ_y are symmetric and commute, they admit a simultaneous orthonormal eigenbasis. In this basis,

$$K_X \Sigma_y^{1/2} = U \operatorname{diag}(q_i \sqrt{\omega_i}) U^\top.$$

For any rank-r reconstruction, $K_{\widehat{X}} \Sigma_y^{1/2}$ has rank at most r, while

$$D^2_{\Sigma_y}(X, \widehat{X}) = \|(K_X - K_{\widehat{X}})\Sigma_y^{1/2}\|_F^2.$$

Eckart–Young lower-bounds this expression by the sum of the omitted squared singular values of $K_X \Sigma_y^{1/2}$, namely $\sum_{i \notin S_r} \omega_i q_i^2$. A reconstruction whose row Gram matrix is

$$\widehat{X}\widehat{X}^\top = \sum_{i \in S_r} \sigma_i^2 u_i u_i^\top$$

has ridge operator $\sum_{i \in S_r} q_i u_i u_i^\top$ and attains the bound. The argument does not require Σ_y to be invertible. □

Proof of Proposition 5. Let $G = XX^\top$ and $\tau = n\lambda > 0$. The dual form of the ridge operator is

$$K_X = G(G + \tau I)^{-1} = I - \tau(G + \tau I)^{-1}.$$

Equality of two ridge operators is therefore equivalent to equality of the corresponding resolvents, and hence to equality of their row Gram matrices. If $\Sigma_y \succ 0$, then

$$D^2_{\Sigma_y}(X, \widehat{X}) = \|\Delta_X(\widehat{X})\Sigma_y^{1/2}\|_F^2 = 0$$

is equivalent to $\Delta_X(\widehat{X}) = 0$. If $X, \widehat{X} \in \mathbb{R}^{n \times d}$ have equal row Grams, take thin singular-value decompositions with the same left singular vectors and singular values, then extend their right singular bases to orthogonal bases of $\mathbb{R}^d$. The resulting change of right basis gives $\widehat{X} = XQ$ for an orthogonal Q. □

For singular Σ_y, zero weighted distance only requires $\Delta_X(\widehat{X})\Sigma_y^{1/2} = 0$. The objective can therefore ignore changes on the nullspace of its task prior.

Proof of the task-ellipsoid bound. For $y \in \operatorname{range}(\Sigma_y)$, write $y = \Sigma_y^{1/2} z$ with the minimum-norm choice $z = \Sigma_y^{\dagger/2} y$. Then

$$\|\Delta_X(\widehat{X})y\|_2^2 \le \|\Delta_X(\widehat{X})\Sigma_y^{1/2}\|_F^2 \|z\|_2^2 = D^2_{\Sigma_y}(X, \widehat{X}) y^\top \Sigma_y^\dagger y.$$

Equivalently,

$$\sup_{\substack{y \in \operatorname{range}(\Sigma_y) \\ y^\top \Sigma_y^\dagger y \le 1}} \|\Delta_X(\widehat{X})y\|_2^2 = \|\Delta_X(\widehat{X})\Sigma_y^{1/2}\|_{\mathrm{op}}^2 \le D^2_{\Sigma_y}(X, \widehat{X}).$$

MSE controls decoder distance under bounded norms. Let $G = XX^\top$ and $\widehat{G} = \widehat{X}\widehat{X}^\top$. The resolvent identity and $G + \tau I, \widehat{G} + \tau I \succeq \tau I$ give

$$\|K_X - K_{\widehat{X}}\|_F \le \frac{1}{\tau}\|G - \widehat{G}\|_F \le \frac{\|X\|_{\mathrm{op}} + \|\widehat{X}\|_{\mathrm{op}}}{\tau}\|X - \widehat{X}\|_F.$$

Consequently,

$$D^2_{\Sigma_y}(X, \widehat{X}) \le \frac{\|\Sigma_y\|_{\mathrm{op}}(\|X\|_{\mathrm{op}} + \|\widehat{X}\|_{\mathrm{op}})^2}{(n\lambda)^2}\|X - \widehat{X}\|_F^2.$$

Thus norm-controlled Euclidean reconstruction bounds decoder distortion at fixed nonzero ridge. Proposition 5 supplies the nonconverse.

Ridge nondegeneracy. If X has full row rank and smallest singular value $\sigma_{\min}(X) > 0$, Equation 18 yields

$$\|I - K_X\|_{\mathrm{op}} = \frac{n\lambda}{\sigma_{\min}(X)^2 + n\lambda}.$$

Hence $K_X \to I$ as $\lambda \to 0$. In this regime, any two fixed full-row-rank representations have nearly identical ridge operators, so calibrating $\mathrm{tr}(K_X)/n$ away from one prevents a degenerate decoder distance.

Fixed-radius trace-estimator variance. Let $s_j = \sqrt{n}\, g_j/\|g_j\|_2$ for independent $g_j \sim \mathcal{N}(0, I_n)$, let C be fixed, and define

$$\widehat{T}_m = \frac{1}{m}\sum_{j=1}^{m}\|Cs_j\|_2^2, \qquad B = C^\top C.$$

Since s_j is uniform on the sphere of radius $\sqrt{n}$,

$$\mathbb{E}[s_{j,a}s_{j,b}s_{j,c}s_{j,d}] = \frac{n}{n+2}(\delta_{ab}\delta_{cd} + \delta_{ac}\delta_{bd} + \delta_{ad}\delta_{bc}).$$

It follows that

$$\mathbb{E}[\widehat{T}_m] = \mathrm{tr}\, B = \|C\|_F^2, \tag{19}$$

$$\mathrm{Var}(\widehat{T}_m) = \frac{2}{m(n+2)}\left[n\|B\|_F^2 - (\mathrm{tr}\, B)^2\right]. \tag{20}$$

For $C \neq 0$, with $r_{\mathrm{eff}}(C) = \|C\|_F^4/\|C^\top C\|_F^2$, the relative standard deviation is

$$\frac{\mathrm{sd}(\widehat{T}_m)}{\|C\|_F^2} = \sqrt{\frac{2}{m(n+2)}\left(\frac{n}{r_{\mathrm{eff}}(C)} - 1\right)}.$$

Taking $C = \Delta_X(\widehat{X})$ gives the isotropic numerator used by the language-model DPSAE before its target-RMS safety clamp, which did not activate in the reported audit. Taking $C = \Delta_X(\widehat{X})\Sigma_y^{1/2}$ gives a transformed-sphere estimator, but it is not the concatenated structured-prior estimator used in the synthetic experiment. This result does not make the finite-probe ratio in Equation 13 unbiased.

Finite-corpus trace reconciliation. For isotropic group g, let $D^2_{m,g} = \|\Delta_X(\widehat{X}_m)\|_F^2$ and $S_g = \|K_{X_g}\|_F^2$. Proposition 2 gives $\mathrm{tr}\, Q_g = D^2_{M,g} - D^2_{D,g}$. Summing and dividing by $\sum_g S_g$ proves Equation 17. For a mean of groupwise ratios, the corresponding difference is

$$\frac{1}{|\mathcal{G}|}\sum_g \frac{\mathrm{tr}\, Q_g}{S_g},$$

which generally differs from the ratio-of-sums identity.