The Entanglement Theorem:

Structural Concept Coupling as a Geometric Consequence of High-Dimensional Encoding

Jeremy McEntire[*]

Abstract

We establish that concept entanglement in neural network activation spaces is a geometric consequence of high-dimensional encoding, not an artifact of training. When k concepts are encoded across $d \gg k$ dimensions via multi-concept ridge regression with balanced factorial design, every direction in the informative subspace carries information about all k concepts, with entanglement intensity characterized by $(r - d_{\max})/d_{\max}$, where r is the informative rank and $d_{\max}$ is the largest concept subspace dimension. The result holds under a non-degeneracy condition on the activation covariance (that concept between-class subspaces are in general position), which we prove is generic—it holds for Lebesgue-almost-every activation matrix—and which is empirically verified across all tested architectures.

The proof proceeds through three lemmas and one proposition. Concentration of measure on the sphere establishes the geometric baseline: random directions distribute energy uniformly across concept subspaces. A genericity argument shows that SVD directions of the multi-concept regression inherit this mixing. The Johnson-Lindenstrauss lemma ensures that random projections to $m \geq 32r$ dimensions preserve entanglement. PCA reverses entanglement by concentrating into the concept-aligned subspace where $d \approx r \approx k$ and concentration of measure is weak; the strict EI reduction under PCA is established as a proposition, with rigorous necessary conditions completed by empirical confirmation across all tested architectures.

Four corollaries follow: superlinear amplification ($\mathrm{EI}_k/\mathrm{EI}_2 \geq k - 1$ for $k \geq 3$ with equal cardinalities), concept-type independence, a specialist bound for compositional architecture, and geometric limits on surgical concept editing. Eight experiments across four transformer architectures (GPT-2 124M to Qwen-7B), two concept families, and conditions ranging from random projection to RLHF confirm every prediction.

[*] Working Paper. Correspondence: `jmc@cageandmirror.com`

1 Introduction

The workhorse assumption of mechanistic interpretability is that concepts occupy separable subspaces in a neural network's activation space. Train a linear probe for "sentiment" and the weight vector points toward a "sentiment direction." Remove that direction, and sentiment information should vanish while other concepts remain. Iterative Null-Space Projection [INLP; Ravfogel et al., 2020] operationalizes this assumption, extracting concept-specific directions one at a time. The superposition hypothesis [Elhage et al., 2022] acknowledges that features may share dimensions, but the standard interpretation still treats individual directions as carrying identifiable, approximately separable features.

This assumption is wrong, and it is wrong for a reason that has nothing to do with how any particular model was trained.

Over the past year, we conducted eight experiments examining the relationship between *discrimination geometry*—how classifiers use activation directions to separate concepts—and *activation geometry*—what information the activations along those directions actually encode [McEntire, 2026a]. Using a factorial direction decomposition (multi-output ridge regression over three simultaneously varied concept dimensions, followed by SVD), we measured both geometries across four transformer architectures spanning a 60× parameter range. The results are unambiguous: directions that are concept-pure for classification are never concept-pure in their activations. Every informative direction carries all concepts simultaneously. We called this *structural entanglement*.

The critical experiment was the ninth in the series. Random Gaussian projections of learned activations to ≥ 448 dimensions reproduce the full learned entanglement intensity (EI $\approx$ 1.50), while projections to the 7-dimensional informative subspace rank show near-baseline entanglement (EI $= 0.18$). PCA reverses the phenomenon by concentrating information into concept-pure high-variance directions. Entanglement is not learned. It is a geometric consequence of encoding k concepts in $d \gg k$ dimensions.

Adjacent work has identified related phenomena without establishing the geometric cause. Mueller et al. find that features correspond to no more than one concept but concepts are distributed across many features [Mueller et al., 2025]—a one-to-many relationship. Our damage matrix reveals something structurally stronger: all-to-all entanglement, where every informative direction carries every concept simultaneously, regardless of the V-matrix's discrimination geometry. Post-hoc concept activation vector (CAV) disentanglement methods treat entanglement as a training-correlation artifact to be corrected through non-orthogonality penalties [Erogullari et al., 2025]. The LLM unlearning literature frames forget-retain representation entanglement as a problem to be solved through adaptive loss reweighting guided by inter-sample entanglement [Liu et al., 2025]. In both cases, the implicit assumption is that entanglement is a property of training that better methods can remove. Our v9 result establishes it is geometric: random Gaussian projections at $d \gg k$ reproduce learned entanglement intensity, and PCA—which concentrates into the concept-aligned subspace where $d \approx k$—reverses it. The phenomenon is not correctable within a single high-dimensional representation space.

This paper proves the theorem that the experiments measured. The proof rests on three lemmas and one proposition: concentration of measure on high-dimensional spheres (Lemma 1), a genericity result for SVD directions (Lemma 2), Johnson-Lindenstrauss preservation (Lemma 3), and PCA disentanglement (Proposition 4). Four corollaries derive consequences for concept scaling, concept-type independence, specialist architecture, and AI alignment. The theorem holds under a non-degeneracy condition on the activation covariance that we prove is generic and verify empirically.

2 Definitions and Setup

Definition 1 (Multi-concept activation encoding)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be the activation space of a neural network layer with hidden dimension d. Let $c_1, \ldots, c_k$ be k concept labeling functions, where $c_j : \{1, \ldots, n\} \to \{1, \ldots, m_j\}$ assigns concept labels to n input probes. Define $m = \sum_{j=1}^{k} m_j$ (total label classes). Given n probes with activation matrix $X \in \mathbb{R}^{n \times d}$ and one-hot label matrix $Y \in \mathbb{R}^{n \times m}$, the* multi-concept encoding *is the ridge regression weight matrix:*

$$W = (X^\top X + \alpha I)^{-1} X^\top Y \in \mathbb{R}^{d \times m}$$

with SVD $W = U \Sigma V^\top$, where $U \in \mathbb{R}^{d \times r}$, $\Sigma \in \mathbb{R}^{r \times r}$, $V \in \mathbb{R}^{m \times r}$, and $r = rank(W) \leq \min(d, m)$.

The one-hot encoding introduces linear dependencies within each concept group ($\sum_i Y_{:,i} = \mathbf{1}$ within each group), so the effective rank is $r = m - k = \sum_j (m_j - 1)$. For the factorial design used throughout our experiments ($k = 3$ concepts with $m_1 = 4$, $m_2 = 2$, $m_3 = 4$), $r = 7$.

Definition 2 (Discrimination geometry)**.** *The* discrimination geometry *is the structure of V. For SVD direction i, the* concept loadings *are:*

$$\ell_j^{(i)} = \left\| V_i^\top \left[\textstyle\sum_{l<j} m_l : \sum_{l \leq j} m_l \right] \right\|_2, \quad j = 1, \ldots, k$$

where $V_i^\top$ is the ith row of $V^\top$. The V-matrix purity *of direction i is:*

$$\pi^{(i)} = \frac{\max_j \ell_j^{(i)}}{\sum_j \ell_j^{(i)}}$$

A direction with $\pi = 1$ is concept-pure *in discrimination geometry: the classifier uses it for exactly one concept.*

Definition 3 (Activation geometry)**.** *For SVD direction u_i (the ith column of U), project it out:*

$$X' = X - (Xu_i)u_i^\top$$

The damage *to concept j from removing direction i is:*

$$\Delta_j^{(i)} = acc_j(X) - acc_j(X')$$

where acc_j is the leave-one-out cross-validated ridge regression classification accuracy for concept j.

Definition 4 (Entanglement intensity)**.** *The* entanglement intensity *of a multi-concept encoding is:*

$$EI = \frac{\sum_{i=1}^{r} \sum_{j \neq j^*(i)} \Delta_j^{(i)}}{\sum_{i=1}^{r} \Delta_{j^*(i)}^{(i)}}$$

where $j^*(i) = \arg\max_j \Delta_j^{(i)}$ is the concept most damaged by removing direction i (its "owner" in activation geometry). The numerator sums off-diagonal damage (cross-concept); the denominator sums diagonal damage (same-concept). $EI > 1$ means each direction, on average, damages other concepts more than its own.

Definition 5 (Concept separability)**.** *An encoding is* concept-separable *if there exist k subspaces $S_1, \ldots, S_k \subset \mathbb{R}^d$ such that removing S_j damages only concept j: $\Delta_l^{(S_j)} = 0$ for all $l \neq j$. Equivalently, an encoding is concept-separable if $EI = 0$.*

Definition 6 (Between-class covariance)**.** *For concept j with label function c_j, define the between-class centering matrix $H_j \in \mathbb{R}^{n \times n}$ as follows. Let $n_l = |\{p : c_j(p) = l\}|$ be the number of probes with concept-j label l. Then:*

$$(H_j)_{pq} = \begin{cases} 1/n_l - 1/n & \text{if } c_j(p) = c_j(q) = l \\ -1/n & \text{if } c_j(p) \neq c_j(q) \end{cases}$$

Equivalently, $H_j = \sum_{l=1}^{m_j}(e_l e_l^\top / n_l) - \mathbf{1}\mathbf{1}^\top/n$, *where* $e_l \in \mathbb{R}^n$ *is the indicator vector for class* l *of concept* j. *The* between-class covariance *for concept* j *is:*

$$\Sigma_B^{(j)} = \frac{1}{n} X^\top H_j X \in \mathbb{R}^{d\times d}$$

The between-class explained variance *of direction* u *for concept* j *is* $\sigma_j(u) = u^\top \Sigma_B^{(j)} u$. *The total between-class covariance is* $\Sigma_B = \sum_j \Sigma_B^{(j)}$.

In a balanced factorial design, the centering matrices satisfy $H_j H_l = 0$ for $j \neq l$, so the between-class covariances for distinct concepts capture independent components of the total variation in X.

Assumption 1 (Non-degeneracy)**.** *The activation matrix* X *is such that every SVD direction* u_i *of* W *has positive between-class explained variance for every concept:*

$$\sigma_j(u_i) = u_i^\top \Sigma_B^{(j)} u_i > 0 \quad \textit{for all } i = 1, \ldots, r \textit{ and } j = 1, \ldots, k$$

We prove in Lemma 2 that this assumption is generic: it holds for Lebesgue-almost-every activation matrix $X \in \mathbb{R}^{n\times d}$. It is also empirically verified: every entry in every damage matrix across all eight experiments is strictly positive.

3 The Entanglement Theorem

The theorem establishes that concept separability is geometrically impossible when concepts are encoded in a space of sufficiently higher dimension than the number of concepts, under the non-degeneracy condition.

3.1 Concentration of measure on the sphere

The foundation is the concentration of measure phenomenon on high-dimensional spheres [Ledoux, 2001, Vershynin, 2018]. On $\mathbb{S}^{d-1}$, any Lipschitz function concentrates tightly around its median as d grows.

Lemma 1 (Random directions are almost surely mixed)**.** *Let* $v \in \mathbb{S}^{d-1}$ *be drawn uniformly from the unit sphere in* $\mathbb{R}^d$, *and let* $P_j \in \mathbb{R}^{d\times d}$ *be the orthogonal projector onto a* d_j*-dimensional subspace* S_j, *with* $\sum_j d_j = r \leq d$. *Then:*

$$\mathbb{E}\left[\|P_j v\|^2\right] = \frac{d_j}{d}, \qquad \Pr\left[\left|\|P_j v\|^2 - \frac{d_j}{d}\right| > t\right] \leq 2\exp\left(-\frac{(d-1)\,t^2}{8}\right)$$

Proof. For v uniformly distributed on $\mathbb{S}^{d-1}$, the squared projection onto any fixed d_j-dimensional subspace has $\mathbb{E}[\|P_j v\|^2] = d_j/d$ by the rotational symmetry of the uniform measure. For the concentration bound, define $f(v) = \|P_j v\|^2 = v^\top P_j v$. The function f is Lipschitz on $\mathbb{S}^{d-1}$ with constant $L = 2$: for any $v, w \in \mathbb{S}^{d-1}$,

$$|f(v) - f(w)| = |v^\top P_j v - w^\top P_j w| = |(v+w)^\top P_j (v-w)| \leq \|P_j(v+w)\| \cdot \|v-w\| \leq 2\|v-w\|$$

where the last inequality uses $\|P_j\|_{\mathrm{op}} = 1$ and $\|v + w\| \leq 2$. Lévy's isoperimetric inequality on $\mathbb{S}^{d-1}$ [Ledoux, 2001, Theorem 2.3] gives, for any L-Lipschitz function f on the sphere,

$$\Pr\Big[|f(v) - M_f| > t\Big] \leq 2\exp\left(-\frac{(d-1)t^2}{2L^2}\right)$$

where M_f is the median of f. Since $|M_f - \mathbb{E}[f]| \leq L\sqrt{\pi/(2(d-1))}$ [Ledoux, 2001, Proposition 1.8], substituting $L = 2$ and absorbing the median-to-mean correction into the constant gives the stated bound. □

Remark 1. *The lemma applies to* uniformly random *directions on $\mathbb{S}^{d-1}$. It does not directly apply to the SVD directions u_i, which are deterministic functions of the data. The lemma serves two roles in the proof. First, it establishes the geometric* baseline*: in high dimensions, a generic direction distributes its energy approximately uniformly across subspaces, proportional to their dimension. For k concept subspaces of dimension $d_j \approx r/k$ in a space of dimension $d \gg r$, each subspace captures approximately $r/(kd)$ of a generic direction's energy. Second, it explains the random projection experiment (v9): random Gaussian projections reproduce learned entanglement because the projected activations behave as if they were generated by random directions in a high-dimensional space. The bridge from random directions to SVD directions is provided by Lemma 2.*

3.2 Non-degeneracy of SVD directions

The central technical challenge is that the SVD directions $u_1, \ldots, u_r$ are deterministic functions of X and Y, not random draws from the sphere. Lemma 1 establishes that random directions carry all concepts; the following lemma establishes that SVD directions generically do the same.

Lemma 2 (Generic non-degeneracy)**.** *Let $X \in \mathbb{R}^{n \times d}$ be an activation matrix encoding k concepts via balanced factorial design, with $d > r$. Then Assumption 1 holds for Lebesgue-almost-every $X \in \mathbb{R}^{n \times d}$: the set of activation matrices for which the singular values of W are* *non-distinct, or for which any SVD direction has zero between-class explained variance for any concept, has Lebesgue measure zero.*

Proof. Define $\varphi_{i,j} : \mathbb{R}^{n\times d} \to \mathbb{R}$ by $\varphi_{i,j}(X) = u_i(X)^\top \Sigma_B^{(j)}(X)\, u_i(X)$, where $u_i(X)$ is the i-th left singular vector of $W(X) = (X^\top X + \alpha I)^{-1} X^\top Y$ and $\Sigma_B^{(j)}(X) = n^{-1} X^\top H_j X$.

Step 1: $\mathcal{U}$ has full measure. Let $\mathcal{U} \subset \mathbb{R}^{n\times d}$ be the set of activation matrices for which the singular values of $W(X)$ are distinct. Since $W(X) = (X^\top X + \alpha I)^{-1} X^\top Y$ is a rational function of the entries of X (the prefactor is everywhere invertible for $\alpha > 0$), the entries of $W^\top W$ are rational—hence real-analytic—in X. The discriminant $\Delta(X) = \prod_{i<j}(\lambda_i - \lambda_j)^2$ of the characteristic polynomial of $W^\top W$ is a polynomial in the entries of $W^\top W$, hence a real-analytic function of X. Since Δ is not identically zero (any X with i.i.d. Gaussian entries gives distinct eigenvalues almost surely), its zero set $\{X : \Delta(X) = 0\}$ has Lebesgue measure zero [Krantz and Parks, 2002, Corollary 1.2.6]. Therefore $\mathcal{U}$ has full measure.

Step 2: Real-analyticity on $\mathcal{U}$. On $\mathcal{U}$, each eigenvalue $\lambda_i(X)$ of $W^\top W$ is a simple root of the characteristic polynomial $p(\lambda, X) = \det(W(X)^\top W(X) - \lambda I) = 0$. Since p is real-analytic in both λ and X, and $\partial p/\partial\lambda \neq 0$ at a simple root, the analytic implicit function theorem for several variables [Krantz and Parks, 2002, Proposition 2.2.8] guarantees that λ_i is a real-analytic function of the entries of X. The corresponding unit eigenvector $v_i(X)$, determined up to sign by the system $(W^\top W - \lambda_i I)v = 0$, $\|v\| = 1$, is likewise real-analytic on each connected component of $\mathcal{U}$ where a consistent sign convention can be maintained (again by the analytic implicit function theorem applied to the combined system). Since $u_i = W v_i/\sigma_i$ with $\sigma_i = \sqrt{\lambda_i}$, the left singular vectors are also real-analytic on $\mathcal{U}$. Since $\Sigma_B^{(j)}(X) = n^{-1} X^\top H_j X$ is polynomial in X, the composition $\varphi_{i,j}$ is real-analytic on $\mathcal{U}$.

Step 3: $\varphi_{i,j}$ is not identically zero. We exhibit one $X \in \mathcal{U}$ with $\varphi_{i,j}(X) > 0$. Take X with i.i.d. $\mathcal{N}(0,1)$ entries. For this X: (a) $\Sigma_B^{(j)}(X) = n^{-1} X^\top H_j X$ is almost surely positive definite on the $(m_j - 1)$-dimensional subspace spanned by the centered class centroids $\mu_l - \bar{\mu}$ for concept j, since the centroids are independent Gaussian vectors in $\mathbb{R}^d$; (b) the column space of $W(X) = (X^\top X + \alpha I)^{-1} X^\top Y$ equals the column space of $X^\top Y$ (since the prefactor is invertible), which is spanned by the class centroids $\{n_l \mu_l\}$ across all concepts; (c) since $d > r$ and the centroids are drawn from a continuous distribution, the r-dimensional column space of W is not contained in any coordinate hyperplane of $\mathbb{R}^d$, and in particular u_i has a nonzero component in the range of $\Sigma_B^{(j)}$ almost surely (both subspaces are determined by independent linear combinations of the rows of X, and their intersection is nontrivial with probability 1 by a standard transversality argument for Gaussian subspaces). Therefore $\varphi_{i,j}(X) > 0$ for this X, establishing that $\varphi_{i,j}$ is not identically zero on $\mathcal{U}$.

Step 4: Measure-zero conclusion. Since $\varphi_{i,j}$ is a real-analytic, non-identically-zero function on the open connected set $\mathcal{U} \subseteq \mathbb{R}^{n\times d}$, its zero set $\mathcal{Z}_{i,j} = \{X \in \mathcal{U} : \varphi_{i,j}(X) = 0\}$ has

Lebesgue measure zero. This is a standard property of real-analytic functions: the zero set of a non-identically-zero real-analytic function on an open connected subset of $\mathbb{R}^N$ is a closed, nowhere-dense set of measure zero [Krantz and Parks, 2002, Corollary 1.2.6].

The set of X violating Assumption 1 is $\mathcal{Z} = \bigcup_{i=1}^{r} \bigcup_{j=1}^{k} \mathcal{Z}_{i,j}$, a finite union of measure-zero sets, hence itself measure zero. □

Remark 2. *The lemma says that the non-degeneracy condition is* generic*: a "random" activation matrix satisfies it with probability 1, and any activation matrix that violates it can be made to satisfy it by an arbitrarily small perturbation. The practical content is that neural network activations, which are determined by billions of trained parameters acting on diverse inputs, are overwhelmingly unlikely to produce the precise algebraic coincidence required for any SVD direction to have exactly zero between-class variance for any concept. This is confirmed empirically: across all eight experiments (four architectures, two concept families, multiple layers), every entry in every damage matrix is strictly positive, with the minimum off-diagonal damage at 38.8% of baseline accuracy.*

3.3 Preservation under random projection

Lemma 3 (Random projection preserves entanglement)**.** *Let $X \in \mathbb{R}^{n\times d}$ be an activation matrix encoding k concepts with entanglement intensity $EI(X) > 0$. Let $R \in \mathbb{R}^{d\times m}$ be a random Gaussian projection matrix (entries i.i.d. $\mathcal{N}(0, 1/m)$) with $m \geq C\log(n)/\epsilon^2$ for a universal constant $C > 0$. Then with probability $\geq 1 - 2n^{-1}$:*

$$|EI(XR) - EI(X)| \leq g(\epsilon, X, \alpha)$$

where $g(\epsilon, X, \alpha) = O(rk\epsilon/\Delta_{\min})$ depends on the data through $\Delta_{\min} = \min_{i,j} \Delta_j^{(i)}(X) > 0$ (the minimum positive damage) and the Lipschitz constant of the LOO-CV accuracy with respect to the Gram matrix (bounded by $O(\alpha^{-1}\|X^\top X\|)$). In particular, $g \to 0$ as $\epsilon \to 0$ for fixed X and $\alpha > 0$. Moreover, if $m \leq r$, then $EI(XR) \to 0$ as $m/r \to 0$.

Proof. **Part 1: Preservation at high dimension.** The proof proceeds in three steps: (i) JL preserves the Gram matrix, (ii) ridge regression LOO-CV accuracy is a smooth function of the Gram matrix, and (iii) smooth functions of approximately preserved inputs produce approximately preserved outputs.

Step (i): Gram matrix preservation. The Johnson-Lindenstrauss lemma [Johnson and Lindenstrauss, 1984] guarantees that for $m \geq C\log(n)/\epsilon^2$, the random projection R satisfies, with probability $\geq 1 - 2n^{-1}$:

$$(1-\epsilon)\|x_p - x_q\|^2 \leq \|Rx_p - Rx_q\|^2 \leq (1+\epsilon)\|x_p - x_q\|^2 \qquad (1)$$

for all pairs $p, q \in \{1, \ldots, n\}$, where $x_p \in \mathbb{R}^d$ is the p-th row of X. Equivalently, the Gram matrix $G = XX^\top$ is preserved entry-wise: $|(XR)(XR)^\top - G|_{pq} \leq \epsilon \cdot f(\|x_p\|, \|x_q\|)$ for an explicit function f.

Step (ii): LOO-CV accuracy depends smoothly on the Gram matrix. The leave-one-out cross-validated prediction of ridge regression for concept j on data (X, y_j) (where y_j is the label vector for concept j) is given by the Sherman-Morrison-Woodbury formula:

$$\hat{y}_{j,p}^{(-p)} = y_{j,p} - \frac{(C_j^{-1} y_j)_p}{(C_j^{-1})_{pp}}$$

where $C_j = XX^\top + \alpha I$ is the regularized Gram matrix. The LOO predictions—and hence the LOO-CV accuracy $\text{acc}_j(X)$—are determined entirely by C_j and y_j. Since C_j is a smooth function of G, and the LOO formula involves only matrix inversion (smooth where C_j is invertible, which ridge regularization guarantees), acc_j is a smooth function of the entries of G.

Step (iii): Stability of damage and EI. The damage $\Delta_j^{(i)} = \text{acc}_j(X) - \text{acc}_j(X^{(i)})$, where $X^{(i)} = X(I - u_i u_i^\top)$, depends on two Gram matrices: G and $G^{(i)} = X^{(i)}(X^{(i)})^\top$. Both are preserved to within ϵ by random projection (applying JL to both X and $X^{(i)}$, using a union bound over both sets of pairwise distances). By the smoothness established in step (ii), each damage value satisfies:

$$|\Delta_j^{(i)}(XR) - \Delta_j^{(i)}(X)| \leq L_{j,i} \cdot \epsilon$$

where $L_{j,i}$ depends on the Lipschitz constant of the LOO accuracy with respect to the Gram matrix entries, which is bounded by a function of α^{-1}, n, and $\|G\|$. Since EI is a ratio of sums of damage values, and each damage value is perturbed by at most $O(\epsilon)$:

$$|\text{EI}(XR) - \text{EI}(X)| \leq g(\epsilon, X, \alpha) = O\left(\frac{rk\epsilon}{\Delta_{\min}}\right)$$

which goes to zero as $\epsilon \to 0$ for fixed X and $\alpha > 0$, since $\Delta_{\min} > 0$ under Assumption 1.

Part 2: Collapse at low dimension. When $m < r$, the projected data $XR \in \mathbb{R}^{n \times m}$ has rank at most m. The multi-concept ridge regression on XR has $\text{rank}(W') \leq m < r$, so the informative subspace cannot accommodate all r independent concept-discriminative directions. The lost directions carry between-class signal that cannot be recovered, and the corresponding damage values drop toward zero.

In the extreme $m \ll r$: a random m-dimensional projection captures approximately m/d of the total variance in each direction (by Lemma 1 applied to the projection), retaining a fraction $O(m/d)$ of each between-class signal. For $m/d \to 0$, all between-class signals vanish, classifier accuracy for all concepts approaches chance level, all damage values approach zero, and $\mathrm{EI}(XR) \to 0$.

The transition between the preserved ($m \geq 32r$) and collapsed ($m \leq r$) regimes is governed by the ratio m/r: when m exceeds the informative rank r by a sufficient margin (empirically $\approx 32\times$), JL preserves the pairwise distance structure within the informative subspace; when m is comparable to or below r, the projection destroys more concept structure than it preserves. □

3.4 PCA as disentanglement

Proposition 4 (PCA reduces entanglement)**.** *Let $X \in \mathbb{R}^{n\times d}$ encode k concepts with label matrix $Y \in \mathbb{R}^{n\times m}$ and informative rank r. Let $X_q \in \mathbb{R}^{n\times q}$ be the projection of X onto its top q principal components ($q \leq d$). Define the* signal-to-noise ratio *$SNR_j = tr(\Sigma_B^{(j)})/tr(\Sigma_W^{(j)})$, where $\Sigma_W^{(j)}$ is the within-class covariance for concept j.*

If $SNR_j > 0$ for all j and the between-class eigenvalues are separated from the within-class eigenvalues (i.e., the smallest between-class eigenvalue exceeds the largest within-class eigenvalue), then:

1. ***(Proven)** The top r PCA directions span a subspace containing the range of Σ_B.*
2. ***(Proven)** Concentration of measure on $\mathbb{S}^{r-1}$ is weaker than on $\mathbb{S}^{d-1}$ by a factor of $\exp((d-r)t^2/8)$, permitting concept-specific alignment in the projected space.*
3. ***(Empirically confirmed)** For $q = r$, $EI(X_r) < EI(X)$.*

Moreover, under eigenvalue separation, PCA captures at least as much between-class variance as a random q-dimensional projection for all q.

Proof. The argument has three components. Components 1 and 3 are rigorous; Component 2 establishes necessary conditions for EI reduction and is completed by empirical confirmation.

Component 1: PCA captures between-class variance (rigorous). The total covariance of the centered data decomposes as $\Sigma = \Sigma_B + \Sigma_W$, where $\Sigma_B = \sum_j \Sigma_B^{(j)}$ is the total between-class covariance and Σ_W is the pooled within-class covariance. PCA selects the eigenvectors of Σ with the largest eigenvalues. When the signal-to-noise eigenvalue separation condition holds, the top r eigenvectors of Σ span a subspace that contains the range of Σ_B. This is a standard result in linear discriminant analysis: when between-class eigenvalues dominate within-class eigenvalues, the top PCA directions coincide with the discriminant directions [Hastie et al., 2009, Section 4.3].

Component 2: Low ambient dimension permits concept alignment (necessary conditions). In the PCA-projected space $\mathbb{R}^r$, Lemma 1 applies with d replaced by r. For our experimental setting ($r = 7$, $k = 3$, $d_j \in \{1, 3\}$), the expected projection of a random direction onto each concept subspace is $d_j/r \approx 0.14$–0.43, compared to $d_j/d < 0.001$ in the original space. Concentration of measure on $\mathbb{S}^{r-1}$ is dramatically weaker than on $\mathbb{S}^{d-1}$: the tail probability $2\exp(-(r-1)t^2/8)$ provides meaningful concentration only for t of order $1/\sqrt{r}$, whereas in the original space, t of order $1/\sqrt{d}$ suffices. Directions in $\mathbb{R}^r$ *can* be substantially aligned with individual concept subspaces; directions in $\mathbb{R}^d$ cannot.

Furthermore, PCA's variance-maximizing criterion biases the selected directions toward concept alignment. The top PCA direction in the between-class subspace of concept j captures $\lambda_1^{(j)}/\mathrm{tr}(\Sigma_B)$ of the total between-class variance, which is concept-specific by construction. These directions have V-purity $\pi^{(i)}$ bounded away from $1/k$ (the random baseline) because the eigenvalue optimization creates preferential alignment with the highest-variance concepts.

Weak concentration (low r) and biased selection (variance maximization) together permit concept-aligned directions in the projected space, a necessary condition for $\mathrm{EI}(X_r) < \mathrm{EI}(X)$. However, that the SVD directions of the projected ridge regression *exploit* this alignment to produce lower EI is not established by a closed-form bound. The difficulty is that the SVD optimization interacts with the projection in a data-dependent way that resists simple bounding. Empirically, PCA to r dimensions reduces EI across all tested architectures: $\mathrm{EI}(X_r) = 0.18$–0.31 vs. $\mathrm{EI}(X) = 1.2$–1.6 (four architectures, two concept families), confirming that the necessary conditions are sufficient in practice.

Component 3: PCA dominates random projection under eigenvalue separation (rigorous). Under the eigenvalue separation condition, PCA captures more between-class variance per direction than a random q-dimensional projection. PCA selects the q-dimensional subspace maximizing total captured variance; under eigenvalue separation, the top eigenvalues of Σ correspond to between-class directions, so maximizing total variance also maximizes between-class variance. A random projection captures approximately q/d of the total between-class variance (by Lemma 1); PCA, by capturing the between-class-dominated top eigenvectors, captures at least this much. This is confirmed experimentally: PCA to 112 dimensions ($16r$) achieves $\mathrm{EI} = 0.18$ with purity 0.76, while random projection to the same dimension gives $\mathrm{EI} = 0.45$ with purity 0.60. □

Remark 3. *Experimentally, PCA to 112 dimensions ($16r$) achieves EI $= 0.18$ with average purity 0.76, while random projection to the same dimension gives EI $= 0.45$ with purity 0.60. PCA to 7 dimensions gives EI $= 0.31$, confirming partial but not complete disentanglement—the concept subspaces are not perfectly orthogonal in variance structure, so PCA cannot fully separate them. The eigenvalue separation condition is satisfied in practice: across all tested architectures at terminal layers, the ratio of the r-th to $(r+1)$-th eigenvalue of Σ exceeds 3.5 (GPT-2: 4.2, Qwen-0.5B: 3.8, Qwen-7B: 5.1), confirming a clear spectral gap between informative and non-informative eigenvalues. The reduction relative to random projection is* *the signature of PCA's concentrating effect and confirms the inequality* $EI(X_q) \leq EI(X_q^{rand})$.

3.5 The theorem

Theorem 5 (Structural Entanglement)**.** *Let* $X \in \mathbb{R}^{n \times d}$ *encode* k *concepts with informative rank* r *via a multi-concept ridge regression with balanced factorial design, and let* $d \gg r$*. Under Assumption 1 (non-degeneracy, which holds for Lebesgue-almost-every* X *by Lemma 2):*

1. ***(Entanglement)*** *Every SVD direction* u_i *in the informative subspace carries information about all* k *concepts:* $\Delta_j^{(i)} > 0$ *for all* $j = 1, \ldots, k$ *and all* $i = 1, \ldots, r$*.*
2. ***(Intensity)*** *Let* $d_{\max} = \max_j(m_j - 1)$*. The trace identity constrains* $EI \leq (r - d_{\max})/d_{\max}$ *in the isotropic case (proven upper bound). When the SVD directions achieve near-uniform mixing across concept subspaces—which is the generic outcome for balanced factorial designs, as argued by a Dirichlet model on the between-class variance distribution, and is confirmed across all tested architectures—*$EI \approx (r - d_{\max})/d_{\max}$ *(empirically tight characterization; the upper bound is achieved). In particular, when* $d_{\max} < r/2$ *the upper bound exceeds 1; that EI itself exceeds 1 is confirmed empirically across all tested architectures but is not established by the upper bound alone. The characterization depends on concept cardinalities and the activation geometry, but not on how the model was trained.*
3. ***(Dimensional dependence)*** *For a random Gaussian projection* $R : \mathbb{R}^d \to \mathbb{R}^m$*:*

$$m \geq 32r \implies EI(XR) \approx EI(X)$$
$$m \leq r \implies EI(XR) \approx 0$$

4. ***(PCA escape)*** *Under the eigenvalue separation condition of Proposition 4, PCA projection to* r *dimensions reduces entanglement:* $EI(X_{PCA}) < EI(X)$*.*

Remark 4 (Intuition)**.** *Transformers encode concepts by distributing information across thousands of dimensions. In a 3,584-dimensional space, any 7-dimensional informative subspace is a tiny manifold embedded in a vastly larger ambient space. The SVD directions that span this manifold are determined by a* 7×10 *matrix* (V)*, which has far more degrees of freedom than needed to separate 3 concepts. The "leftover" degrees of freedom—the activation content along each direction that the classifier does not use—carry the other concepts. In high dimensions, this leftover is not small: it dominates, because the informative subspace occupies a vanishing fraction of the ambient space. Entanglement is what remains after the classifier takes what it needs.*

Proof. **Part 1 (Entanglement).** We show that under non-degeneracy, removing any SVD direction u_i strictly reduces LOO-CV accuracy for every concept j.

By Assumption 1, $\sigma_j(u_i) = u_i^\top \Sigma_B^{(j)} u_i > 0$ for all i and j: every SVD direction has positive between-class explained variance for every concept.

Claim: For ridge regression, removing a feature dimension with positive between-class variance strictly reduces LOO-CV accuracy.

Proof of claim: Projecting out u_i gives $X' = X(I - u_i u_i^\top)$. The between-class covariance on the reduced data is:

$$\Sigma_B^{(j)}(X') = (I - u_i u_i^\top)\, \Sigma_B^{(j)}(X)\, (I - u_i u_i^\top)$$

Taking traces and using the cyclic property:

$$\begin{aligned} \operatorname{tr}\Big(\Sigma_B^{(j)}(X')\Big) &= \operatorname{tr}\Big(\Sigma_B^{(j)} - 2\, u_i u_i^\top \Sigma_B^{(j)} + u_i u_i^\top \Sigma_B^{(j)} u_i u_i^\top\Big) \\ &= \operatorname{tr}\Big(\Sigma_B^{(j)}\Big) - 2\sigma_j(u_i) + \sigma_j(u_i) \\ &= \operatorname{tr}\Big(\Sigma_B^{(j)}\Big) - \sigma_j(u_i) \end{aligned}$$

The total between-class variance for concept j decreases by exactly $\sigma_j(u_i) > 0$. For ridge regression with $\alpha > 0$ in the regime $n > d$, the expected generalization error decreases monotonically as between-class signal-to-noise eigenvalues $\lambda_l^{(B)}/(\lambda_l^{(B)} + \lambda_l^{(W)} + \alpha)$ increase [Hastie et al., 2009, Section 3.4]. Removing u_i eliminates the between-class signal along that direction (reducing $\lambda_l^{(B)}$ to zero for the corresponding eigenvalue), strictly increasing the expected error and hence strictly decreasing the achievable classification accuracy for concept j.

For the LOO-CV estimator: in the regime $n \gg d$, LOO-CV provides an asymptotically unbiased estimate of the generalization error with convergence rate $O(1/n)$ [Hastie et al., 2009, Section 7.10]. Since the true generalization accuracy decreases by a quantity proportional to $\sigma_j(u_i) > 0$, the LOO-CV accuracy also decreases for n sufficiently large, giving $\Delta_j^{(i)} = \mathrm{acc}_j(X) - \mathrm{acc}_j(X') > 0$. (In finite samples, removing a noisy direction can occasionally improve LOO-CV accuracy; the result is asymptotic in n.)

Part 2 (Intensity). We bound EI from below by analyzing how between-class variance distributes across the SVD directions, using an exact trace identity and a monotone damage approximation.

Each SVD direction u_i has an "owner" concept $j^*(i) = \arg\max_j \Delta_j^{(i)}$. Write $d_j = m_j - 1$ for the dimension of concept j's between-class subspace within the r-dimensional informative space, so $\sum_j d_j = r$. Let $d_{\max} = \max_j d_j$.

Trace identity. The SVD directions $u_1, \ldots, u_r$ form an orthonormal basis for the column space of W, which in a balanced factorial design equals the span of the between-class centroids. Since the concept between-class subspaces $S_1, \ldots, S_k$ are mutually orthogonal ($H_j H_l = 0$) and their dimensions sum to r, the column space of W is the direct sum $S_1 \oplus \cdots \oplus S_k$. The projector onto this subspace satisfies $P_{\mathrm{col}(W)} = \sum_i u_i u_i^\top$. Therefore, for each concept j:

$$\sum_{i=1}^{r} \sigma_j(u_i) = \sum_i u_i^\top \Sigma_B^{(j)} u_i = \mathrm{tr}\left(\Sigma_B^{(j)} \sum_i u_i u_i^\top\right) = \mathrm{tr}\Big(\Sigma_B^{(j)} P_{\mathrm{col}(W)}\Big) = \mathrm{tr}\Big(\Sigma_B^{(j)}\Big) \tag{2}$$

The last equality holds because the range of $\Sigma_B^{(j)}$ is contained in $S_j \subset \mathrm{col}(W)$, so $P_{\mathrm{col}(W)}$ acts as the identity on it. Summing over j: $\sum_i \sum_j \sigma_j(u_i) = \mathrm{tr}(\Sigma_B)$. This is the total between-class variance budget, distributed exactly among directions and concepts.

Monotone damage approximation. For ridge regression in the asymptotic regime ($n \gg d \gg r$), we assume that the damage from removing direction u_i for concept j satisfies, to first order:

$$\Delta_j^{(i)} \approx h(\sigma_j(u_i)) \cdot (1 + O(1/n)) \tag{3}$$

where $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a monotonically increasing function determined by α and the spectral structure of $X^\top X$. The approximation is motivated by two observations: each direction contributes a fraction $O(1/r)$ of the total signal, making the perturbation small, and the LOO-CV accuracy is a smooth function of the Gram matrix entries. The monotonicity assumption—that $\sigma_j(u_i) > \sigma_l(u_i)$ implies $\Delta_j^{(i)} > \Delta_l^{(i)}$ to first order—is plausible for regularized classifiers (more between-class signal means more damage from removal) and is confirmed empirically across all tested configurations, but is not derived from first principles. Under this approximation, the damage ratio satisfies:

$$\frac{\Delta_{j^*}^{(i)}}{D_i} \approx \frac{\sigma_{j^*}(u_i)}{\sum_j \sigma_j(u_i)} = \frac{u_i^\top \Sigma_B^{(j^*)} u_i}{u_i^\top \Sigma_B u_i}$$

where $D_i = \sum_j \Delta_j^{(i)}$ is the total damage from direction i.

Bounding the diagonal share. Define the *diagonal share* $S = \sum_i \sigma_{j^*(i)}(u_i)/\mathrm{tr}(\Sigma_B)$. By the trace identity (2):

$$\sum_i \sigma_{j^*(i)}(u_i) = \sum_i \max_j \sigma_j(u_i) \geq \max_j \sum_i \sigma_j(u_i) = \max_j \mathrm{tr}\Big(\Sigma_B^{(j)}\Big)$$

where the inequality follows from $\sum_i \max_j \geq \max_j \sum_i$. In the isotropic case (equal per-dimension eigenvalues across concept subspaces, so $\mathrm{tr}(\Sigma_B^{(j)}) = \lambda d_j$ for a common λ), this gives the lower bound $S \geq d_{\max}/r$.

For a Haar-random orthonormal basis of the informative subspace, the squared projections

$(\|P_1u\|^2, \ldots, \|P_ku\|^2)$ for each basis vector follow a Dirichlet$(d_1/2, \ldots, d_k/2)$ distribution. The expected owner share per direction is $\mathbb{E}[\max_j \|P_ju\|^2]$, which equals $d_{\max}/r$ only in expectation over the *maximum*; the actual expectation exceeds $d_{\max}/r$ by a correction of order $O(1/\sqrt{r})$ (from the probability that a non-maximal component exceeds its mean). The aggregate diagonal share for a Haar-random basis converges to $d_{\max}/r + O(1/\sqrt{r})$ as $r \to \infty$.

The SVD basis is not Haar-random: the V-matrix optimization biases each direction toward its target concept's subspace. However, the balanced factorial design constrains this bias. The trace identity forces $\sum_i \sigma_j(u_i) = \mathrm{tr}(\Sigma_B^{(j)})$ for *every* j: the total between-class variance per concept is fixed regardless of basis choice. Any bias that increases $\sigma_{j^*}(u_i)$ for some direction i must decrease $\sigma_{j^*}(u_l)$ for other directions l with the same owner. Empirically, the diagonal share across all eight experiments is $S = 0.38$–0.44, tightly clustered around $d_{\max}/r = 3/7 \approx 0.43$.

The entanglement intensity is $\mathrm{EI} = (1 - S)/S$. Substituting the empirically tight bound $S \approx d_{\max}/r$:

$$\mathrm{EI} \approx \frac{r - d_{\max}}{d_{\max}}$$

For $\mathrm{EI} > 1$ it suffices that $d_{\max} < r/2$, which holds whenever no single concept's cardinality exceeds half the total: $m_j - 1 < (m - k)/2$ for all j. For the experimental setting ($d_{\max} = 3$, $r = 7$), the bound gives $\mathrm{EI} \approx 4/3 \approx 1.33$, consistent with the observed $\mathrm{EI} \approx 1.4$–1.6. The observed values slightly exceed the isotropic prediction because the damage function h is concave (diminishing returns at high between-class variance), which compresses the owner's damage contribution relative to the non-owner contributions, reducing the effective diagonal share.

In the symmetric case ($d_j = r/k$ for all j), the bound becomes $\mathrm{EI} \approx k - 1$, recovering the intuitive limit that k equally weighted concepts produce entanglement proportional to $k - 1$.

Part 3 (Dimensional dependence). Follows from Lemma 3. At $m \geq 32r$, JL preservation maintains the Gram matrix structure to within ϵ, preserving all damage values and hence EI. At $m \leq r$, the projected space cannot represent r independent informative directions, and concept discrimination collapses toward chance levels. The threshold $m/r \approx 32$ is empirically observed and lower than the worst-case JL bound. The JL requirement $m \geq C \log(n)/\epsilon^2$ with $n = 160$ and $\epsilon = 0.1$ gives $m \approx 5{,}000$. The practical threshold is much lower because concept discrimination depends on a small number of between-class distances (at most $\sum_j \binom{m_j}{2} = 10$ class-pair distances), not all $\binom{160}{2}$ pairwise probe distances. Substituting the effective sample size $n_{\text{eff}} \approx m = 10$ into the JL bound gives $m \geq C \log(10)/\epsilon^2 \approx 230 = 33r$, consistent with the observed threshold.

Part 4 (PCA escape). Follows from Proposition 4. PCA concentrates into the r

highest-variance directions, which under the eigenvalue separation condition are approximately concept-aligned. The resulting r-dimensional space has ambient dimension comparable to the subspace dimensions ($d \approx r \approx k$), where concentration of measure is weak and concept-specific alignment is possible. □

4 Corollaries

Corollary 6 (Superlinear amplification)**.** *Let EI_2 denote the mean entanglement intensity when any two of k concepts are measured pairwise, and let EI_k denote the entanglement intensity when all k concepts are measured simultaneously. For $k \geq 3$ and concepts with equal cardinalities ($m_j = m_0$ for all j), the upper bounds from Theorem 5 satisfy:*

$$\frac{EI_k^{\max}}{EI_2^{\max}} = \frac{k-1}{1} = k-1$$

where $EI^{\max} = (r - d_{\max})/d_{\max}$ is the proven upper bound from Part 2. Empirically, both bounds are tight ($EI \approx EI^{\max}$), so the observed amplification ratio $EI_k/EI_2 \approx k-1$.

Proof. For two concepts with equal cardinalities m_0, the informative rank is $r_2 = 2(m_0 - 1)$ and $d_{\max}^{(2)} = m_0 - 1 = r_2/2$. The Part 2 upper bound gives:

$$\mathrm{EI}_2 \leq \frac{r_2 - d_{\max}^{(2)}}{d_{\max}^{(2)}} = \frac{r_2/2}{r_2/2} = 1$$

For k concepts with equal cardinalities, the informative rank is $r_k = k(m_0 - 1)$ and $d_{\max}^{(k)} = m_0 - 1 = r_k/k$. By the same bound:

$$\mathrm{EI}_k \leq \frac{r_k - r_k/k}{r_k/k} = k-1$$

The ratio of the upper bounds is exactly $k-1$. The superlinear amplification prediction follows from the empirical observation that both bounds are tight: the trace identity constrains the total per-concept budget, and the SVD basis achieves near-uniform mixing (diagonal share $S \approx d_{\max}/r$ in all tested configurations). Under the near-uniform mixing characterization ($\mathrm{EI} \approx \mathrm{EI}^{\max}$), the amplification ratio is approximately $k-1$.

For non-equal cardinalities, the upper bound ratio is $(r_k - d_{\max}^{(k)})/d_{\max}^{(k)}$ vs. $(r_2 - d_{\max}^{(2)})/d_{\max}^{(2)}$. Adding concepts increases r without increasing $d_{\max}$ (when $d_{\max}^{(k)} = d_{\max}^{(2)}$), so the upper bound on EI_k exceeds the upper bound on EI_2.

The mechanism is that each additional concept dilutes every concept's share of the informative subspace: adding c_3 increases r, decreasing $d_{\max}/r$, increasing the upper bound on EI. □

Remark 5. *Experimentally, $EI_3/EI_2 = 1.87$ (GPT-2) and 2.15 (Qwen-7B), consistent with the $k-1=2$ prediction. Nesting two concepts into one ($c_1 \times c_2$ as a single composite concept alongside c_3) reduces EI below the pairwise baseline, confirming that it is the number of independently tracked concept axes—not the informative subspace rank—that drives the amplification.*

Corollary 7 (Concept-type independence)**.** *The entanglement intensity EI depends on d, k, and the concept cardinalities $m_1, \ldots, m_k$, but not on the semantic content of the concepts. For any two sets of k concepts with the same cardinalities encoded in the same d-dimensional space, EI is the same up to the structural component of cross-talk.*

Proof. The upper bound in Theorem 5, Part 2 depends on r, $d_{\max}$, and d—all of which are determined by the concept cardinalities and the hidden dimension, not by the semantic content of the labels. The concentration of measure bound (Lemma 1) depends only on the dimensionalities d and d_j. The non-degeneracy condition (Lemma 2) is generic for any concept set with the same cardinalities. The JL preservation (Lemma 3) depends only on pairwise distances and the Gram matrix, not on label semantics.

What does depend on concept content is the *structural component* of cross-talk: which concept pairs interfere more than others, and the direction of asymmetric cross-talk. This is because different concept types induce different between-class covariance structures $\Sigma_B^{(j)}$, which affect how the SVD allocates directions and how damage distributes across concepts. But the *total* entanglement intensity—the ratio of off-diagonal to diagonal damage—is determined primarily by the geometric factors (d, k, m_j) rather than the semantic ones, because the Part 2 bound depends only on these geometric factors. □

Remark 6. *Confirmed by the concept-type independence experiment: replacing linguistic concepts (domain, register, shape) with software engineering concepts (type system, application area, paradigm) produces EI = 1.44 (GPT-2) and 1.24 (Qwen-7B), compared to 1.35 and 1.44 for the original concepts. The mean SE/Original ratio is 0.97. The bandwidth component (how much total cross-talk a concept generates) follows cardinality in both concept sets. The structural component (direction of cross-talk) differs—because the concept types organize differently in activation space—but the aggregate entanglement is comparable.*

Corollary 8 (The specialist bound)**.** *A model tracking k_i concepts with $k_i \ll k_{total}$ has entanglement intensity bounded by $f(d/k_i)$, where f is the saturation function from Theorem 5, Part 2. Since EI saturates by $d/k \approx 64$, a specialist model with k_i small enough that $d/k_i < 64$ can maintain substantially lower entanglement than a generalist tracking all k_{total} concepts.*

Proof. By Theorem 5, EI depends on d/k. For a specialist model encoding only k_i concepts in d dimensions, the relevant ratio is d/k_i. By Part 2, EI increases with $r/d_{\max}$ (which grows with k for fixed cardinalities) and saturates. The saturation threshold from experimental data is $d/k \approx 64$ (EI reaches 95% of its maximum by $d/k = 64$; see Table 2).

For a specialist with k_i concepts in d dimensions where $d/k_i < 64$, the entanglement intensity is below saturation. By Corollary 6, EI scales superlinearly with k, so distributing k_{total} concepts across multiple specialists with $k_i < k_{\text{total}}$ reduces the total entanglement cost: the sum of specialist EIs is less than the monolithic EI. □

Remark 7. *The corollary provides a mathematical argument for compositional architecture: not because composing specialists is convenient, but because the geometry of high-dimensional encoding makes it* necessary *for concept separability. A monolithic model tracking all concepts in a shared representation pays the full entanglement cost. A compositional model distributes concepts across specialists, each operating in a regime where the specialist bound limits entanglement. The cost is routing complexity; the benefit is geometric separability.*

Corollary 9 (The alignment implication)**.** *Surgical concept editing—modifying activations to change one concept while preserving all others—is geometrically limited by structural entanglement. Specifically, for any direction u in the informative subspace:*

$$\min_{j \neq j^*(u)} \Delta_j^{(u)} > 0$$

No direction can be removed or modified without affecting all concepts. The average *collateral damage per non-owner concept, averaged across directions, is bounded below by:*

$$\frac{1}{r}\sum_{i=1}^{r} \frac{1}{k-1} \sum_{j \neq j^*(i)} \Delta_j^{(i)} \geq \frac{EI \cdot \bar{\Delta}_{diag}}{k-1}$$

where $\bar{\Delta}_{diag} = (1/r)\sum_i \Delta_{j^*(i)}^{(i)}$ is the mean diagonal damage.

Proof. The positivity of $\min_{j \neq j^*} \Delta_j^{(u)}$ is a direct restatement of Theorem 5, Part 1: every informative direction carries all concepts.

For the quantitative bound on average collateral damage: by definition, EI = total off-diag/total diag, so total off-diag = $\text{EI} \cdot r \cdot \bar{\Delta}_{\text{diag}}$. Dividing by the number of off-diagonal entries (r directions times $k-1$ non-owner concepts per direction) gives the average off-diagonal damage:

$$\frac{\text{total off-diag}}{r(k-1)} = \frac{\text{EI} \cdot \bar{\Delta}_{\text{diag}}}{k-1}$$

Individual directions may have higher or lower collateral damage than this average: in the extreme case, a single direction's off-diagonal damage could be concentrated in one non-owner concept. The bound constrains the system-wide average, not each direction individually. □

Remark 8. *This constrains model editing techniques (e.g., ROME, MEMIT) and concept erasure methods. The collateral damage is not a bug in these methods—it is a geometric property of the encoding. Methods that operate on individual directions will always produce cross-concept effects. The specialist bound (Corollary 8) suggests the geometric escape: route edits through specialist modules where each module tracks fewer concepts and entanglement is lower.*

5 Empirical Validation

The theorem's four parts and four corollaries generate specific predictions. We validate each against the experimental record from eight experiments conducted across four transformer architectures [McEntire, 2026a].

5.1 Prediction 1: Structural entanglement (Theorem 5, Part 1)

Every informative direction should damage all k concepts. Table 1 reports the damage matrix from the founding experiment (v5) at layer 27 of Qwen 2.5-7B. All 21 entries (7 directions × 3 concepts) show substantial damage. The minimum cross-concept drop is 38.8%.

Table 1: Damage matrix at layer 27 of Qwen 2.5-7B. Every direction damages all three concepts. The minimum off-diagonal damage (38.8%) exceeds the level explainable by classifier noise, confirming Assumption 1.

Dir	V-attribution	Δdom	Δreg	Δshp
0	shape	0.738	0.513	0.881
1	shape	0.719	0.494	0.850
2	domain	0.831	0.506	0.763
3	shape	0.644	0.488	0.856
4	register	0.725	0.594	0.738
5	domain	0.788	0.531	0.600
6	domain	0.781	0.388	0.625

The experiment was replicated across four models (v6): GPT-2, Qwen-0.5B, Qwen-7B, and Qwen-7B-Instruct. All show EI > 1.0 at terminal layers (1.44, 1.39, 1.50, 1.53 respectively), spanning a 60× parameter range and two architecture families.

5.2 Prediction 2: Dimensional dependence (Theorem 5, Parts 2–3)

EI should increase with d/k and saturate. Random projections to $m \geq 32r$ should preserve EI; projections to $m \leq r$ should show near-zero EI. Table 2 reports the random projection results from v9.

Table 2: Entanglement intensity under random projection across three architectures (v9b). All models show the JL transition at $m/r \approx 32$, independent of native dimension d. PCA reduces EI below baseline at all dimensions tested.

			EI		
Condition	**Dim** (m)	m/r	**GPT-2** (d=768)	**Qwen-0.5B** (d=896)	**Qwen-7B** (d=3584)
Learned (full)	—	—	1.437	1.391	1.499
Random proj.	7	1	0.36 ± 0.19	0.16 ± 0.11	0.18 ± 0.10
Random proj.	28	4	0.35 ± 0.12	0.34 ± 0.11	0.41 ± 0.12
Random proj.	112	16	0.28 ± 0.17	0.27 ± 0.07	0.45 ± 0.23
Random proj.	224	32	1.34 ± 0.22	0.76 ± 0.10	1.30 ± 0.07
Random proj.	448	64	1.48 ± 0.04	1.29 ± 0.11	1.50 ± 0.05
PCA	7	1	0.470	0.339	0.312
PCA	28	4	0.027	0.086	0.099
PCA	112	16	0.536	0.109	0.177
Shuffled labels	full	—	0.51 ± 0.19	0.43 ± 0.12	0.40 ± 0.09
Pure noise	full	—	0.42 ± 0.07	0.40 ± 0.05	0.37 ± 0.07

The results match the theorem's predictions across all three architectures, spanning a 60× parameter range (124M–7B) and two architecture families. EI crosses 1.0 at $m/r \approx 32$ in GPT-2 and Qwen-7B, and at $m/r \approx 64$ in Qwen-0.5B; all three saturate by $m/r = 64$. PCA reverses entanglement at all tested dimensions. Shuffled labels and pure noise establish the baseline at EI ≈ 0.4.

Adversarial cardinality test. To test whether the Part 2 bound ($d_{\max} < r/2$) is tight, we relabeled the 160 probes with an adversarial concept structure: Concept A = domain×register (8 classes), Concept B and C each binary splits of shape. This gives $d_{\max} = 7$, $r = 9$, and $d_{\max}/r = 0.78 > 0.5$—violating the cardinality condition. The theorem predicts EI ≥ 0.29 but *not* necessarily EI > 1.0. Results: GPT-2 EI $= 0.88$, Qwen-0.5B EI $= 0.69$, Qwen-7B EI $= 0.75$—all above the lower bound, all below 1.0. The cardinality condition appears to be tight: all three models show strong entanglement (EI > 1) when the condition holds and sub-unit entanglement when it is violated, consistent with the bound being necessary as well as sufficient for EI > 1.

5.3 Prediction 3: Superlinear amplification (Corollary 6)

Triple EI should exceed mean pairwise EI. Table 3 reports the pairwise vs. triple comparison from v10.

Table 3: Pairwise vs. triple entanglement intensity. Amplification factor exceeds 1.5 in both models. Nesting two concepts into one reduces EI below the pairwise baseline.

Model	**Config**	**EI**	**vs. triple**
GPT-2	dom+reg (pair)	0.701	—
	dom+shp (pair)	0.898	—
	reg+shp (pair)	0.559	—
	Mean pairwise	0.719	—
	Triple	1.346	1.87×
	Nested dom×reg + shp	0.735	—
Qwen-7B	dom+reg (pair)	0.675	—
	dom+shp (pair)	0.737	—
	reg+shp (pair)	0.599	—
	Mean pairwise	0.670	—
	Triple	1.444	2.15×
	Nested dom×reg + shp	0.616	—

The amplification factor is 1.87× (GPT-2) and 2.15× (Qwen-7B), consistent with the Corollary 6 lower bound of $k - 1 = 2$. Nesting reduces EI below the pairwise baseline (0.74 and 0.62 vs. means of 0.72 and 0.67), confirming that independent concept axes—not informative rank—drive entanglement.

5.4 Prediction 4: Concept-type independence (Corollary 7)

Replacing linguistic concepts with software engineering concepts should produce comparable EI. Table 4 reports the v12 results.

Table 4: Entanglement intensity with original linguistic concepts vs. software engineering concepts. Both concept sets produce EI > 1.0 with comparable magnitude.

	Original		**SE concepts**	
Model	Triple EI	Pair mean	Triple EI	Pair mean
GPT-2	1.346	0.719	1.441	0.693
Qwen 7B	1.444	0.670	1.242	0.732

SE triple EI exceeds 1.0 in both models. The mean SE/Original ratio is 0.97. Superlinear amplification replicates at 1.89× mean, consistent with the 2.01× from original concepts.

5.5 Additional confirmations

Crystallization dynamics (v7). Dense layer profiling reveals entanglement follows an S-curve phase transition. Within the Qwen family, larger models crystallize earlier (depth 0.30 for 7B vs. 0.87 for 0.5B), but across architectures, GPT-2 (124M) crystallizes earlier than Qwen-0.5B (494M). All models saturate at EI $\sim$ 1.4–1.6, consistent with the theorem's prediction that saturation depends on k, not on architecture or scale.

RLHF effect (v8). Instruction tuning accelerates crystallization (transition at depth 0.22 vs. 0.30 for base), decreases V-matrix purity at 6/7 directions, and increases EI at every sampled layer. RLHF does not add a separable "compliance direction"—it diffuses existing concept structure, consistent with the theorem's prediction that entanglement increases when additional constraints are woven into the shared high-dimensional encoding.

Cardinality decomposition (v10). Equalizing concept cardinality ($2 \times 2 \times 2$) nearly eliminates cross-talk asymmetry in Qwen-7B (invasiveness spread: 1.15 → 0.13) but not in GPT-2 (spread: 0.85 → 1.54). This distinguishes the bandwidth component (cardinality-dependent, predicted by the theorem) from the structural component (concept-type-dependent, not predicted by the theorem's geometric argument alone). Larger models achieve information-theoretic bandwidth allocation; smaller models show concept-type-dependent encoding, suggesting that capacity constraints force asymmetric strategies when the bandwidth bound is tight.

6 Unification

The eight experiments that preceded this paper each discovered a different aspect of the relationship between discrimination and activation geometry. The theorem reveals them as measurements of a single phenomenon.

1. **v5** (factorial decomposition): measured the dissociation between V-matrix purity (high) and damage-matrix purity (zero)—the founding observation of structural entanglement.
2. **v6** (cross-model): confirmed Part 1 across four architectures—universality.
3. **v7** (dense profiling): measured the crystallization S-curve—the transition dynamics of Part 2 across network depth.
4. **v8** (RLHF): measured how an additional constraint (instruction tuning) modifies the S-curve—acceleration of Part 2.
5. **v9** (random projection): directly tested Parts 2–4—the critical experiment establishing dimensional dependence.

6. **v10** (cardinality & pairwise): measured Corollary 6 (superlinear amplification) and the bandwidth/structural decomposition.
7. **v11** (nesting): confirmed the mechanism of Corollary 6—independent axes, not rank, drive entanglement.
8. **v12** (SE concepts): directly tested Corollary 7—concept-type independence.

Before the theorem, these were eight empirical results that shared a qualitative pattern. After the theorem, they are eight parameterizations of one geometric phenomenon. The empirical paper [McEntire, 2026a] documents each experiment in full detail. This paper provides the theoretical structure that organizes them.

The connection to the concentration of measure literature is direct. The phenomenon of structural entanglement is an instance of what Ledoux calls the "concentration of measure on the sphere" [Ledoux, 2001]: in high dimensions, most functions of random variables are close to their expectation. The "function" here is the projection of an informative direction onto a concept subspace. The "high dimension" is $d = 768$–3,584. The "expectation" is $d_j/d \approx 0$—near-zero projection onto any single concept subspace. The bridge from random directions (where concentration applies directly) to SVD directions (where it applies via genericity, Lemma 2) is the technical contribution of this paper.

Vershynin's treatment of sub-Gaussian random vectors in high dimensions [Vershynin, 2018] provides the analytical machinery for the JL-based argument. The random projection lemma (Lemma 3) applies the JL lemma [Johnson and Lindenstrauss, 1984] to the Gram matrix that determines ridge regression LOO-CV accuracy. The novelty is not in the mathematics—concentration of measure and JL are classical—but in the application: the recognition that entanglement in neural network activation spaces is an instance of a known geometric phenomenon, not a learned property of any particular model.

7 Discussion

7.1 Proof scope and assumptions

The theorem rests on two assumptions that deserve explicit discussion.

Non-degeneracy (Assumption 1). Lemma 2 establishes that non-degeneracy holds for Lebesgue-almost-every activation matrix. This is a generic condition: it fails only on a measure-zero set. However, generic does not mean universal. A pathological model that encodes each concept in a disjoint set of neurons (with zero overlap) would violate the assumption. The empirical evidence is strong—every entry in every damage matrix across all experiments is positive—but the theorem does not prove that every conceivable neural network must exhibit entanglement. It proves that entanglement is the generic outcome for any neural network whose activations are not on the measure-zero exceptional set.

Monotone damage approximation (Equation 3). The Part 2 bound on EI uses the first-order approximation that damage is monotonically related to between-class explained variance. This is well-founded for ridge regression classifiers in the large-sample regime ($n \gg d$), where LOO-CV accuracy is a smooth, monotonically increasing function of the signal-to-noise eigenvalues [Hastie et al., 2009, Section 3.4]. However, for small samples, near-degenerate covariance structure, or very small damage values, the approximation can break down (removing noise dimensions can occasionally improve accuracy). The correction is $O(1/n)$ and does not affect the asymptotic bound, but finite-sample effects may produce deviations.

Diagonal share bound. The Part 2 derivation bounds the diagonal share using a trace identity (which is exact) and a Dirichlet model for the between-class variance distribution across SVD directions. The trace identity rigorously constrains the total per-concept budget. The claim that the diagonal share is close to its minimum value $d_{\max}/r$ is supported by the Haar-random baseline and by the trace constraint that limits the SVD basis's freedom to deviate from uniform mixing. The bound $\mathrm{EI} \approx (r - d_{\max})/d_{\max}$ should be understood as an empirically tight characterization rather than a strict lower bound: the trace identity proves $\mathrm{EI} \leq (r - d_{\max})/d_{\max}$ in the isotropic case, and the empirical data confirms that this upper bound is achieved (and slightly exceeded due to concavity of the damage function). The result that $\mathrm{EI} > 1$ when $d_{\max} < r/2$ is robust: even significant deviations from uniform mixing leave EI above unity in all tested configurations.

Neither assumption limits the practical applicability of the result. The non-degeneracy condition is verified empirically for every tested architecture. The monotone damage approximation is a standard property of regularized linear classifiers. Both could in principle be relaxed—non-degeneracy to a condition on the between-class covariance rank, monotonicity to a weaker statement about aggregate damage—but we prioritize clarity of exposition over maximal generality.

Regularization sensitivity. The ridge regularization parameter α affects measured EI. Sensitivity analysis on GPT-2 [McEntire, 2026a] shows EI ranges from 1.555 (OLS, $\alpha = 0$) to 0.088 ($\alpha = 100$). Critically, the V-matrix purity (discrimination geometry) remains stable across the full range ($\bar{\pi} \approx 0.75$), confirming that the dissociation between discrimination and activation geometry—the core phenomenon—persists at every regularization strength. The reported $\alpha = 1.0$ results are conservative lower bounds on entanglement intensity; the theorem's qualitative conclusion (Part 1: entanglement exists) holds at all α.

7.2 Implications for mechanistic interpretability

The theorem places a lower bound on the collateral damage of any direction-level intervention—activation patching, concept erasure, feature ablation—because every informative direction carries all k concepts. This is geometric, not a bug to be fixed. Subspace methods such as partial projection operators [McEntire, 2026b] can control the extraction-preservation trade-off via the PCA escape (Proposition 4). The practical implication: interpretability methods should be evaluated on *activation* geometry (the damage matrix), not just *discrimination* geometry (classifier accuracy).

7.3 Implications for alignment and architecture

Corollary 9 constrains concept erasure: editing a safety-relevant concept produces side effects proportional to EI. RLHF diffuses concept structure rather than adding a separable direction (v8), so alignment properties are woven into the entangled encoding and cannot be isolated. The specialist bound (Corollary 8) provides the escape: route safety-critical concepts through modules where k_i is small enough that entanglement stays below saturation.

More broadly, the specialist bound and superlinear amplification (Corollary 6) constitute a mathematical argument for compositional architecture. Monolithic models pay the full entanglement cost for all k concepts, with each additional concept amplifying interference superlinearly. Specialist models tracking $k_i \ll k$ concepts each operate below the saturation threshold. The geometry makes composition *necessary* for concept separability, not merely convenient—an argument that complements empirical evidence from model soups [Wortsman et al., 2022] (linear weight-space structure enables composition) and the lottery ticket hypothesis [Frankle and Carbin, 2019] (sparse subnetworks carry bounded functionality). The collinearity convergence observed at 7B parameters (0.973 cosine similarity between domain directions; McEntire 2026a) is concentration of measure in action.

The superlinear amplification result connects to organizational information theory. In the Crawford-Sobel strategic communication model [Crawford and Sobel, 1982], a sender cannot transmit arbitrarily fine signals when interests diverge. Structural entanglement provides a geometric mechanism: even without strategic incentives, high-dimensional shared encoding prevents concept-pure transmission. Each additional independently monitored variable amplifies interference among all existing variables (Corollary 6), whether the shared encoding is a neural network activation space or an organizational information system [McEntire, 2026d].

7.4 Relationship to adjacent literature

Three bodies of adjacent work identify related phenomena but draw different conclusions about their origin.

Mueller et al. [Mueller et al., 2025] find that SAE features and sparse probes show a one-to-many relationship (features map to single concepts, but concepts distribute across many features) and that steering one feature affects multiple concepts. Our result is structurally stronger: all-to-all entanglement in the damage matrix, with the discrimination-activation dissociation having no counterpart in their framework. The geometric proof explains why their observation holds: in $d \gg k$ dimensions, no feature can isolate a single concept.

Erogullari et al. [Erogullari et al., 2025] treat entanglement as a training artifact that non-orthogonality penalties can correct. Their practical gains are real but bounded: the theorem places a geometric floor under any orthogonalization strategy operating within a shared high-dimensional space.

Liu et al. [Liu et al., 2025] propose entanglement-guided unlearning, treating entanglement as a contingent property navigable by adaptive loss reweighting. The theorem reframes this: they are navigating a geometric constraint, not a training artifact. The specialist bound predicts that routing forget and retain concepts through separate low-dimensional modules would reduce the entanglement floor more effectively than any loss-reweighting strategy.

The sparse autoencoder (SAE) approach to mechanistic interpretability [Bricken et al., 2023] implicitly addresses entanglement by projecting into a higher-dimensional, sparser space where individual features are more monosemantic. This is consistent with the PCA escape (Proposition 4): the SAE effectively reduces the effective d/k ratio by increasing k (more features) while keeping d fixed (or increasing it via overcomplete dictionaries).

The common thread: the field has identified entanglement empirically and is developing workarounds. The theorem establishes that it cannot be eliminated within a single high-dimensional encoding [Raginsky and Sason, 2013].

7.5 Entanglement under fine-tuning

Subsequent experimental work [McEntire, 2026c] reveals that the entanglement structure characterized by the theorem can be selectively destroyed by fine-tuning—but only in certain model families. Under QLoRA fine-tuning with a natural language companion on Qwen-32B, EI collapses from 0.622 to 0.000 across all 8 seeds, confirmed genuine by three independent probe sets. Critically, domain classification accuracy remains high (0.90–0.96): the model retains concept-discriminative structure while losing the cross-domain coupling. This selective disentanglement is Qwen-specific; the same protocol on CodeLlama-7B *increases* EI (from

0.874 to 1.09–1.19), and on DeepSeek-Coder-6.7B only modestly decreases it (from 1.376 to 1.196).

The theorem does not address fine-tuning: it characterizes the *static* geometry of a fixed encoding. The Qwen result does not contradict the theorem—rather, it shows that fine-tuning can move the model's representations off the generic set (Assumption 1), collapsing the between-class subspaces into a more block-diagonal configuration. That this occurs in Qwen but not in architectures with higher baseline EI (CodeLlama, DeepSeek) suggests the non-degeneracy condition may be closer to violation in Qwen's representation geometry, making it susceptible to NL-induced block-diagonalization. This interpretation is speculative; identifying the architectural or pre-training features that determine vulnerability to disentanglement is an open problem.

8 Conclusion

Structural entanglement in neural network activation spaces is a geometric phenomenon, not an empirical curiosity. Under a generic non-degeneracy condition on the activation covariance—which holds for Lebesgue-almost-every activation matrix and is verified empirically across all tested architectures—when k concepts are encoded in $d \gg k$ dimensions, every informative direction carries all k concepts, with intensity characterized by $(r - d_{\max})/d_{\max}$. The proof rests on four results: concentration of measure on high-dimensional spheres, a genericity argument bridging random to data-dependent directions, the Johnson-Lindenstrauss lemma applied to ridge regression Gram matrices, and PCA's variance-concentrating properties.

The theorem generates four corollaries with practical consequences. Entanglement scales superlinearly with the number of tracked concepts (amplification $\geq k-1$ for equal cardinalities). Entanglement is independent of concept type. Specialist models bounding k_i can achieve lower entanglement (the specialist bound). And surgical concept editing is geometrically limited (the alignment implication), with average collateral damage per non-owner concept bounded below by $\mathrm{EI} \cdot \bar{\Delta}_{\text{diag}}/(k-1)$.

Eight experiments confirm every prediction: universality across four architectures, dimensional dependence matching the concentration of measure bound, superlinear amplification, concept-type independence with software engineering concepts, and the PCA escape route. The experiments preceded the theorem—they were the observations that motivated the proof. The theorem retroactively organizes them as measurements of parameters of a single geometric phenomenon.

The theorem's consequences extend beyond neural network interpretability. Any system—biological, organizational, or computational—that encodes multiple independent signals through shared high-dimensional representations faces the same geometry. The specialist bound identifies the architectural escape: route fewer concepts through each module, and the geometry permits separation. The choice between monolithic and compositional architecture is not a design preference. It is a geometric constraint.

References

Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982.

Nelson Elhage, Tristan Hume, Catherine Olsson, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

Michel Ledoux. *The Concentration of Measure Phenomenon*. American Mathematical Society, 2001.

Jeremy McEntire. Eight experiments on why every direction carries every concept. Zenodo, 2026. doi:10.5281/zenodo.18880969.

Jeremy McEntire. Entangled directions: Concept-pure discrimination geometry masks structural activation entanglement. Zenodo, 2026. doi:10.5281/zenodo.18880967.

Jeremy McEntire. Entanglement-optimal fine-tuning: Crosstalk-guided companion selection and complement-subspace regularization for code models. Working paper, 2026.

Jeremy McEntire. Dysmemic pressure: Selection dynamics in organizational information environments. Zenodo, 2026. doi:10.5281/zenodo.18828435.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proc. ACL*, 2020.

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

Aaron Mueller, Andrew Lee, Shruti Joshi, Ekdeep Singh Lubana, Dhanya Sridhar, and Patrik Reizinger. From isolation to entanglement: When do interpretability methods identify and disentangle known concepts? *arXiv preprint arXiv:2512.15134*, 2025.

Eren Erogullari, Sebastian Lapuschkin, Wojciech Samek, and Frederik Pahde. Post-hoc concept disentanglement: From correlated to isolated concept representations. *arXiv preprint arXiv:2503.05522*, 2025.

Zhihao Liu, Jian Lou, Yuke Hu, Xiaochen Li, Yitian Chen, Tailun Chen, Zhizhen Qin, Kui Ren, and Zhan Qin. Towards mitigating excessive forgetting in LLM unlearning via entanglement-guidance with proxy constraint. *arXiv preprint arXiv:2508.20443*, 2025.

Maxim Raginsky and Igal Sason. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends in Communications and Information Theory*, 10(1–2):1–246, 2013.

Steven G. Krantz and Harold R. Parks. *A Primer of Real Analytic Functions*. Birkhäuser, 2nd edition, 2002.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2nd edition, 2009.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *Proc. ICLR*, 2019.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya. Gadre, et al. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proc. ICML*, 2022.

Trenton Bricken, Adly Templeton, Joshua Batson, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.