Structural Entanglement in the Informative Subspace:

Eight Experiments on Why Every Direction
Carries Every Concept

Jeremy McEntire*

Abstract

Linear probing assumes that concept-specific information occupies separable subspaces in transformer activations. We show this assumption fails. Using factorial direction decomposition—multi-output ridge regression over three simultaneously varied concept dimensions, followed by SVD—we decompose the activation space into directions ordered by total label information.

The SVD's right singular vectors (V-matrix) show clean concept separation: each direction loads primarily on one concept. But the *damage matrix*—accuracy loss when each direction is removed—reveals structural entanglement: every informative direction causes 39–88% accuracy drops across *all three* concepts. Directions that are concept-pure in *discrimination* geometry (how the classifier uses them) are concept-entangled in *activation* geometry (what information they carry).

Eight experiments establish the phenomenon's scope. Cross-model replication in GPT-2, Qwen-0.5B, Qwen-7B, and Qwen-7B-Instruct shows EI > 1.0 at all terminal layers. Dense layer profiling reveals entanglement follows an S-curve phase transition with architecture-specific dynamics. Random projection experiments show entanglement is dimension-dependent: projections to $\geq$448 dimensions reproduce the full learned EI across a 60× parameter range, while PCA concentrates information into concept-pure directions, reducing EI below baseline. Entanglement scales superlinearly with concept count (2× amplification from pairwise to triple), and the phenomenon replicates with software engineering concepts unrelated to the original linguistic dimensions.

The discrimination–activation dissociation has methodological implications: any interpretability method claiming to isolate a concept in activation space should demonstrate isolation in the damage matrix, not just in classifier weights.

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

1 Introduction

Linear probing is the workhorse of mechanistic interpretability. Train a classifier to predict concept labels from a model's internal activations, and the classifier's weight vector identifies a "concept direction" in activation space. Iterative Null-Space Projection [INLP; Ravfogel et al., 2020] extends this to find complete concept subspaces by iteratively removing each discovered direction and re-probing. The implicit assumption is that concepts occupy separable subspaces—that a "domain direction" carries domain information and not much else.

Over the past year, we have published seven papers examining this assumption in Qwen 2.5-7B. Each targeted a different aspect of the relationship between classifier geometry and activation geometry:

1. **Double dissociation** [McEntire, 2026a]: INLP erases domain to chance while shape classification holds at $\geq$95.6%. Conclusion: domain and shape occupy separable subspaces.
2. **Spectral isotropy** [McEntire, 2026b]: The layer Jacobian amplifies INLP directions and random directions identically (ratio 0.99 ± 0.05). Conclusion: the forward pass treats domain directions as generic.
3. **Classification-intervention gap** [McEntire, 2026c]: Directions optimized for classification produce the best interventions; causally derived directions are worse than random. Conclusion: the two tasks operate under different constraints.
4. **Causal anti-selectivity** [McEntire, 2026d]: INLP is the only basis with positive selectivity (+0.618); patching (-0.630), contrastive (-0.777), and gradient (-0.116) bases are all anti-selective. Conclusion: causal directions target a geometry where selectivity is impossible.
5. **Proportional compression** [McEntire, 2026e]: The natural language interface drops $\sim$80% of activation information, proportionally across concepts. Conclusion: domain-specific information is too small a fraction ($\sim$1.4%) for any lossy channel to preserve selectively.
6. **Asymmetric entanglement** [McEntire, 2026f]: A single register INLP direction drops domain accuracy 52.5% despite lying 82% outside domain INLP's 36-dimensional subspace. Conclusion: INLP has a blind spot for cross-concept directions.

Each of these findings is correct at the level it claims. But they share a pattern that none of them explained: classifiers succeed where interventions fail, and interventions always produce collateral damage. Why?

This paper provides the answer. Using a factorial direction decomposition—multi-output ridge regression over three simultaneously varied concept dimensions, followed by SVD—we measure both the *discrimination geometry* (how a classifier uses each direction) and the *activation geometry* (what information the activations along each direction carry). These are fundamentally dissociated. Directions that are concept-pure for classification are never concept-pure in their activations. Every informative direction in the 7-dimensional label subspace carries all three concepts simultaneously.

Contributions.

1. We introduce **factorial direction decomposition**: multi-output ridge regression + SVD over balanced factorial probes, producing directions with both structural (V-matrix) and functional (damage matrix) characterizations.
2. We demonstrate the **discrimination–activation dissociation**: the V-matrix shows concept-pure directions; the damage matrix shows those same directions carry universal cross-concept information.
3. We confirm **cross-model universality**: entanglement intensity > 1.0 in GPT-2, Qwen-0.5B, Qwen-7B, and Qwen-7B-Instruct, with crystallization phase transitions whose dynamics are architecture-specific.
4. We establish that entanglement is **dimension-dependent**: random projections to $\geq$448 dimensions reproduce the full learned EI across a 60× parameter range, while PCA concentrates information into concept-pure directions, reducing EI below baseline.
5. We demonstrate **superlinear amplification**: triple entanglement exceeds mean pairwise by 2×, confirmed with both linguistic and software engineering concept types.

Additional results on RLHF effects, cross-talk decomposition, cardinality bandwidth, fractional factorial efficiency, and concept-type independence are presented in the body.

2 Background and Prior Results

2.1 INLP and the concept-purity assumption

INLP [Ravfogel et al., 2020] finds concept-specific directions by training a linear classifier to predict concept labels from activations, extracting the weight vector as a direction, projecting the activations onto the null space of that direction, and repeating until classification accuracy falls below threshold. The procedure assumes that the directions it finds are concept-pure—that removing a domain direction removes domain information without affecting other concepts.

This assumption is testable. If a domain INLP direction is truly domain-pure, removing it should decrease domain classification accuracy and leave other concept accuracies unchanged. We call this the **concept-purity test**: remove a direction and measure all concepts.

2.2 Summary of prior findings

Table 1 summarizes the seven prior results. The rightmost column identifies what each study actually measured—a distinction that becomes critical in light of the present findings.

Table 1: Prior results from the activation geometry program. All findings are in Qwen 2.5-7B.

Paper	Method	Key finding	What it measured
Shape	INLP + shape probes	Domain erasure preserves shape ($\geq$95.6%)	Classifier robustness
Spectral	Jacobian amplification	INLP/random ratio = 0.99	Forward-pass geometry
Synthesis	Multi-basis comparison	Classification $\neq$ intervention	Task dissociation
Causal	Selectivity scoring	Only INLP is selective (+0.618)	Basis quality
Compression	NL interface analysis	80% drop, proportional	Channel capacity
Entangled	Sigma sweep	Register dir → 52.5% domain drop	Cross-concept coupling

2.3 The shared pattern

Across these studies, a consistent pattern emerges: classifiers trained on single concepts achieve high accuracy, but interventions based on the directions those classifiers identify produce cross-concept damage. The Shape paper found domain erasure preserves shape—but only measured shape classification accuracy after domain removal, not whether the activations along domain directions carry shape information. The Spectral paper found the forward pass treats domain directions as generic—but did not ask whether "generic" means "carrying all concepts." The Causal paper found INLP is the only selective basis—but did not ask whether the selectivity is a property of the classifier or of the activations.

Each study measured *classifier geometry*. None measured *activation geometry* directly.

3 Method

3.1 Factorial probe design

We construct 160 text probes by independently varying three concept dimensions:

- **Domain** (4 classes): medical, legal, code, science
- **Register** (2 classes): formal, informal
- **Shape** (4 classes): hierarchical, causal, constraint, evidence

The full factorial design is $4 \times 2 \times 4 = 32$ cells with 5 replications per cell ($32 \times 5 = 160$ probes). Each probe independently instantiates all three dimensions: a medical-formal-hierarchical probe is a formally written medical text with hierarchical reasoning structure. The balanced design ensures that each concept dimension varies orthogonally to the others in the label space.

We extract last-token activations from layers 7, 14, 21, and 27 of Qwen 2.5-7B (3,584-dimensional hidden states, FP16 capture, float64 analysis). Activations are centered (mean-subtracted) per layer.

3.2 Multi-output ridge regression and SVD

We construct a one-hot target matrix $Y \in \mathbb{R}^{160\times 10}$ encoding all three concept dimensions (4 domain + 2 register + 4 shape columns). We fit a multi-output ridge regression:

$$W = (X^\top X + \alpha I)^{-1} X^\top Y, \qquad W \in \mathbb{R}^{3584\times 10}$$

with $\alpha = 1.0$ (matching all prior work in this program). We test sensitivity to α in §6.5. The SVD of W gives:

$$W = U\Sigma V^\top$$

where $U \in \mathbb{R}^{3584\times 10}$ contains directions in activation space, Σ is a diagonal matrix of singular values, and $V \in \mathbb{R}^{10\times 10}$ contains concept loadings. Since Y has effective rank 7 (one-hot columns within each concept group sum to 1, removing 3 degrees of freedom), W has at most 7 non-trivial singular values.

3.3 V-matrix structural loadings (discrimination geometry)

The rows of $V^\top$ reveal how each direction's weight vector decomposes across concept labels. For direction i, we compute:

$$\begin{aligned}
\text{dom\_V}_i &= \|V_i^\top[0\!:\!4]\|_2 \quad \text{(domain loading)}\\
\text{reg\_V}_i &= \|V_i^\top[4\!:\!6]\|_2 \quad \text{(register loading)}\\
\text{shp\_V}_i &= \|V_i^\top[6\!:\!10]\|_2 \quad \text{(shape loading)}
\end{aligned}$$

A direction with $\text{dom\_V} \approx 1$ and $\text{reg\_V}, \text{shp\_V} \approx 0$ is used by the regression exclusively for domain discrimination. We call this the **discrimination geometry**: how the classifier allocates each direction to concepts.

3.4 Removal damage matrix (activation geometry)

For each direction u_i (column of U), we project it out of the activation matrix:

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

and measure leave-one-out cross-validated classification accuracy on all three concepts. The **damage** for concept c from removing direction i is:

$$\Delta_c^{(i)} = \mathrm{acc}_c(\mathrm{baseline}) - \mathrm{acc}_c(\mathrm{after\ removal})$$

If a direction is truly concept-pure (carries only domain information), removing it should damage domain accuracy but leave register and shape unchanged. The damage matrix reveals the **activation geometry**: what information the activations along each direction actually encode, regardless of how the classifier uses them.

3.5 INLP comparison

We run single-concept INLP independently for each concept dimension. For multi-class concepts (domain, shape), we extract the first right singular vector of the classifier's weight matrix ($\mathrm{coef}_{n_{\mathrm{classes}} \times d}$) as the INLP direction—the direction of maximum class separation in feature space. We compute:

1. **Coverage**: the fraction of each SVD direction's variance that lies in each INLP subspace, measured as $\|Q_c^\top u_i\|^2$ where Q_c is the orthonormal basis of the INLP subspace for concept c.
2. **Cross-concept damage**: remove each INLP direction and measure all three concept accuracies.

3.6 Fractional factorial efficiency

We re-run the full SVD analysis on reduced probe sets: half (3 per cell, 96 probes), quarter (2 per cell, 64 probes), and eighth (1 per cell, 32 probes). Loading matrices are compared via greedy direction matching with cosine similarity.

3.7 Bootstrap confidence intervals

To quantify uncertainty in entanglement intensity, we use a stratified bootstrap procedure with 2,000 iterations. Each iteration resamples probes with replacement within each factorial cell (preserving the balanced design), re-fits the multi-output ridge regression and SVD, recomputes the full removal damage pipeline, and yields a bootstrap EI estimate. We report percentile-based 95% confidence intervals $[\hat{\theta}_{0.025}, \hat{\theta}_{0.975}]$. Stratification ensures that every bootstrap sample maintains the factorial structure—without it, some cells could be absent from a resample, invalidating the concept-orthogonality guarantee.

4 Results

All results are from layer 27 of Qwen 2.5-7B unless otherwise noted. Baselines: domain accuracy 96.3%, register 100.0%, shape 95.6%.

4.1 V-matrix: clean concept separation

Table 2 shows the V-matrix structural loadings for all 7 non-trivial SVD directions. The decomposition is strikingly clean: three directions load primarily on shape (shp_V > 0.97), three on domain (dom_V > 0.98), and one on register (reg_V = 0.961). No direction has high loadings on two concepts simultaneously.

Table 2: V-matrix structural loadings at layer 27. Each direction loads primarily on a single concept.

Dir	σ	**dom_V**	**reg_V**	**shp_V**	**Attribution**
0	0.0190	0.123	0.170	0.978	shape
1	0.0173	0.095	0.114	0.989	shape
2	0.0163	0.991	0.046	0.123	domain
3	0.0140	0.105	0.046	0.993	shape
4	0.0117	0.166	0.961	0.220	register
5	0.0105	0.980	0.169	0.106	domain
6	0.0083	0.997	0.034	0.070	domain

At the discrimination level, the factorial SVD has successfully decomposed the activation space into concept-pure directions. A standard interpretation would conclude that the model represents domain, register, and shape in separable subspaces.

4.2 Damage matrix: structural entanglement

Table 3 tells a different story. Removing *any* single SVD direction causes massive accuracy drops across *all three* concepts—including concepts that the V-matrix says the direction does not serve.

Direction 0 is attributed to shape by the V-matrix (shp_V = 0.978). Removing it drops domain accuracy by 73.8% and register accuracy by 51.3%. Direction 2 is attributed to domain

Table 3: Damage matrix at layer 27. Accuracy drop (Δ) when each SVD direction is individually removed. Every direction damages all three concepts. Drops > 0.40 in bold.

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

(dom_V = 0.991). Removing it drops shape accuracy by 76.3%. No direction is concept-pure in the activation sense. The minimum cross-concept drop across all 21 direction-concept pairs (7 directions × 3 concepts) is 38.8%.

4.3 The dissociation

Tables 2 and 3 measure the same directions but reveal different geometries:

- The V-matrix measures how the regression *allocates* each direction to concept labels. This is the **discrimination geometry**—the structure the classifier imposes on the space.
- The damage matrix measures what the *activations* along each direction actually *encode*. This is the **activation geometry**—the structure the model's representations actually have.

The discrimination geometry shows clean separation. The activation geometry shows structural entanglement. These are not contradictory—they measure different things. A direction can be used exclusively for domain discrimination while carrying domain, register, and shape information simultaneously in its activation values. The classifier selects *which aspect* of the activation to attend to; it does not determine *what information* the activation carries.

4.4 INLP coverage: primary directions captured, secondary missed

Table 4 shows how much of each SVD direction lies within the INLP subspaces. Single-concept INLP captures the primary direction for each concept (Dir 0 by shape INLP, Dir 2 by domain INLP, Dir 4 by register INLP) but completely misses four of seven directions.

INLP finds one direction per concept (accuracy drops below threshold after the first iteration for all three concepts). It captures the highest-sigma direction for each—but misses the second and third shape directions (Dirs 1, 3), the second and third domain directions

Table 4: INLP coverage of SVD directions at layer 27. Coverage = $\|Q_c^\top u_i\|^2$.

Dir	V-attr.	dom cov.	reg cov.	shp cov.	union cov.
0	shape	0.020	0.072	**0.974**	0.998
1	shape	0.001	0.027	0.005	0.038
2	domain	**0.978**	0.004	0.010	0.998
3	shape	0.001	0.003	0.000	0.004
4	register	0.000	**0.872**	0.010	0.938
5	domain	0.000	0.022	0.001	0.024
6	domain	0.000	0.001	0.000	0.001

(Dirs 5, 6). Four of seven informative directions have union coverage below 4%.

4.5 INLP cross-concept damage

Table 5 applies the concept-purity test directly to INLP directions. If INLP directions were concept-pure, removing the domain direction should damage only domain, and similarly for register and shape.

Table 5: Cross-concept damage from removing each INLP direction at layer 27. Drops > 0.30 in bold.

Direction removed	Δdom	Δreg	Δshp
Domain INLP	**0.825**	0.044	**0.394**
Register INLP	**0.525**	**0.600**	**0.694**
Shape INLP	**0.488**	**0.500**	**0.863**

No INLP direction passes the concept-purity test. Domain INLP removal drops shape by 39.4%. Shape INLP removal drops domain by 48.8% and register by 50.0%. Register INLP removal drops domain by 52.5% (replicating the Paper 41 finding exactly) and shape by 69.4%.

Methodological note. The damage in Table 5 is measured using the multi-output ridge classifier (same as Table 3), not the single-concept INLP classifiers themselves. We project out the INLP direction from the activations, then evaluate all three concept accuracies via the ridge classifier. This measures whether the *activations* along the INLP direction carry cross-concept information, using the ridge classifier as a readout instrument. The finding is that concept-pure *classifiers* (INLP) operate on representations that are never concept-pure.

4.6 Paper 41 validation

The register-dominant SVD direction (Dir 4) has cosine similarity 0.934 with the register INLP direction from the independent Paper 41 analysis, and only 1.5% of its variance lies within the standard 36-dimensional domain INLP subspace. This replicates the asymmetric entanglement finding and confirms that the factorial decomposition recovers the same geometric structure as independent single-concept analysis.

4.7 Fractional factorial efficiency

Table 6: Fractional factorial efficiency at layer 27. Similarity = mean cosine between matched direction loading vectors.

Design	Probes	Per cell	Similarity
Full factorial	160	5	1.000
Half factorial	96	3	0.971
Quarter factorial	64	2	0.835
Eighth factorial	32	1	0.831

Half the probes recover the full structure with 97.1% fidelity. Even a single probe per cell (32 probes total) achieves 83.1% similarity. The entanglement structure is robust—it is not an artifact of overfitting to 160 probes.

4.8 Cross-model universality

We apply the factorial decomposition to four models: GPT-2 (124M, 12 layers), Qwen 2.5-0.5B (24 layers), Qwen 2.5-7B (28 layers), and Qwen 2.5-7B-Instruct (28 layers). Table 7 reports entanglement intensity and V-matrix purity at each model's terminal layer.

Table 7: Cross-model entanglement at terminal layers. Entanglement intensity (EI) = off-diagonal damage / diagonal damage; EI > 1.0 means more cross-concept than same-concept damage. Average V-matrix purity across top 7 directions (1.0 = concept-pure, 0.33 = fully mixed). 95% CIs from 2,000 stratified bootstrap iterations.

Model	Params	Hidden	Terminal	EI	Purity	95% CI
GPT-2	124M	768	L11	1.437	0.638	[1.273, 1.594]
Qwen 2.5-0.5B	494M	896	L23	1.391	0.622	[1.141, 1.549]
Qwen 2.5-7B	7.6B	3,584	L27	1.499	0.691	[0.965, 1.537]
Qwen 2.5-7B-Inst	7.6B	3,584	L27	1.527	0.604	[1.032, 1.581]

All four models show EI > 1.0—more cross-concept damage than same-concept damage when any informative direction is removed. Bootstrap 95% CIs confirm that entanglement is statistically robust: three of four models have CIs entirely above 1.0, and the fourth (Qwen-7B, CI lower bound 0.965) is consistent with EI $\geq$ 1.0 at the 93rd percentile. The phenomenon spans a 60× parameter range and two distinct architectures (GPT-2's architecture vs. Qwen's). Structural entanglement is not specific to one model.

Procrustes alignment of V-matrices in label space reveals that the concept structure is architecture-family specific: within the Qwen family, alignment ranges from 0.74 (0.5B vs. 7B) to 0.91 (7B vs. 7B-Instruct). GPT-2 aligns with Qwen models at only 0.27–0.52. The *phenomenon* is universal; the *geometry* is architecture-dependent.

4.9 Crystallization phase transitions

Dense layer profiling (every layer for GPT-2 and Qwen-0.5B; every 2 layers for Qwen-7B and 7B-Instruct) reveals that entanglement follows an S-curve phase transition with depth. Table 8 summarizes the crystallization profiles across all four models.

Table 8: Crystallization profiles from dense layer sampling. Transition = depth fraction where EI first exceeds 1.0. Peak = maximum EI observed. Steepest = depth of maximum EI gradient.

Model	**Params**	**EI at L0**	**Transition**	**Peak EI**	**Steepest**
GPT-2	124M	0.537	0.545	1.497 @ d=0.91	0.500
Qwen 2.5-0.5B	494M	0.167	0.870	1.391 @ d=1.00	0.587
Qwen 2.5-7B	7.6B	0.280	0.296	1.599 @ d=0.81	0.111
Qwen 2.5-7B-Instruct	7.6B	0.202	0.222	1.620 @ d=0.81	0.111

Four findings emerge. First, within the Qwen architecture family, larger models crystallize dramatically earlier: Qwen-7B transitions at depth 0.30 versus Qwen-0.5B at 0.87—a 3× difference in transition depth for a 15× parameter increase. Second, all four models saturate at similar EI levels ($\sim$1.4–1.6) regardless of scale or architecture, suggesting a universal ceiling on entanglement intensity. Third, entanglement is present from layer 0 in all models (EI = 0.17–0.54)—it begins with the embedding and intensifies through the network. Fourth, the relationship between model size and crystallization depth is *architecture-specific*: GPT-2 (124M) crystallizes earlier than Qwen-0.5B (494M) despite having fewer parameters, indicating that architectural topology—not raw parameter count—determines crystallization dynamics.

GPT-2 shows a smooth S-curve with the steepest change at mid-depth (0.50). Qwen-0.5B shows a gradual ramp that only crosses the entanglement threshold in the final 13% of layers, with peak EI at the terminal layer—suggesting the 0.5B model barely reaches the crystallized regime. Qwen-7B shows a rapid early ramp (steepest at depth 0.11) followed by a plateau with mild oscillation between 1.3–1.6. The plateau suggests a saturation effect: once entanglement reaches a critical intensity, deeper layers maintain but do not further increase it.

4.10 RLHF and concept diffusion

Comparing Qwen 2.5-7B (base) and Qwen 2.5-7B-Instruct reveals that RLHF systematically increases entanglement. V-matrix purity decreases at 6 of 7 directions (-0.05 to -0.17), and entanglement intensity increases at every sampled layer ($+0.03$ to $+0.09$). The V-matrix alignment between base and Instruct is 0.91—high, but the systematic purity decrease indicates that RLHF does not add a new "compliance direction" to the basis. Instead, it diffuses existing concept structure, redistributing activation energy across directions and making the representation more entangled.

Dense layer profiling sharpens this picture. RLHF accelerates crystallization: the Instruct model transitions to EI > 1.0 at depth fraction 0.222, versus 0.296 for the base model. The Instruct model also reaches a higher peak EI (1.620 vs. 1.599). RLHF does not merely increase entanglement at each layer—it shifts the entire crystallization curve leftward, causing the model to develop entangled representations earlier in the forward pass.

This is consistent with RLHF operating as a fine-tuning procedure that redistributes activation energy across concept dimensions without restructuring the underlying geometry. The compliance behavior learned through RLHF is not encoded as a separable concept—it is woven into the existing entangled representation, and the weaving begins earlier.

4.11 Cross-talk decomposition

The aggregate entanglement intensity treats all concept pairs equally. Decomposing the damage matrix into a directed 3×3 cross-talk matrix—where entry C_{AB} measures how much removing directions "owned by" concept A (highest V-matrix loading) damages concept B—reveals asymmetric structure. Table 9 shows the terminal-layer cross-talk matrices for all three models profiled at dense resolution.

Three patterns are consistent across all models. First, **shape is the most invasive concept**: shape-owned directions cause the largest cross-talk damage to both domain and register (1.93–2.24 to domain, 1.44–1.53 to register). This is consistent across all three architectures and scales. Second, **register is the least invasive**: register-owned directions cause the smallest cross-talk (0.52–0.78). In Qwen-7B, register removal causes *zero* damage to domain and shape at multiple intermediate layers, indicating that register information becomes fully redundant—entirely recoverable from domain and shape directions. Third, **cross-talk is asymmetric**: shape→domain consistently exceeds domain→shape, with mean

Table 9: Terminal-layer cross-talk matrices. Each row shows how much removing that concept's directions damages each column concept. Diagonal = self-damage, off-diagonal = cross-talk. Shape is systematically the most invasive concept (largest off-diagonal); register is the least.

Model	**Owner**	**→Domain**	**→Register**	**→Shape**
GPT-2	Domain	—	1.14	2.09
	Register	0.74	—	0.62
	Shape	1.93	1.44	—
Qwen 0.5B	Domain	—	0.97	1.37
	Register	0.66	—	0.52
	Shape	2.17	1.51	—
Qwen 7B	Domain	—	1.54	2.06
	Register	0.77	—	0.78
	Shape	2.24	1.53	—

asymmetry of 0.34–0.45 across models.

These patterns admit an information-theoretic interpretation. Register is binary (2 classes, 1 bit), while domain and shape are quaternary (4 classes, 2 bits each). The concept with lower information content entangles less and becomes redundant earlier—its information can be reconstructed from the higher-entropy concept directions. Shape, as the concept capturing reasoning structure (hierarchical, causal, constraint, evidence), appears to be the primary organizing axis of the representation: the model's internal geometry is structured around *how* information is organized rather than *what domain* it belongs to.

Cross-talk sequencing across layers confirms this hierarchy. Domain–shape entanglement appears from layer 0, while register entanglement emerges later (depth fraction 0.26–0.43 depending on model). The structural concepts entangle first; the stylistic concept entangles later.

4.12 Cardinality controls bandwidth, not direction

The information-theoretic interpretation predicts that equalizing concept cardinalities should equalize cross-talk. To test this, we apply four labeling schemes to identical terminal-layer activations: the original 4×2×4, a collapsed 2×2×2 (domain: professional/technical; shape: structural/evaluative), and two asymmetric swaps (2×2×4 and 4×2×2). Because only the labels change—not the activations—differences in cross-talk are purely attributable to label structure and the resulting SVD decomposition.

The results distinguish two effects. First, **cardinality controls cross-talk bandwidth**: in both models, the concept with 4 classes always occupies the most SVD directions and

Table 10: Cross-talk invasiveness ratio (off-diagonal / self-talk) by labeling condition. At equal cardinality (2×2×2), Qwen-7B invasiveness spread drops from 1.15 to 0.13, but the 4-class concept does not always rank highest in either model.

Model	**Condition**	**Dom**	**Reg**	**Shp**	**Spread**
GPT-2	4×2×4 (orig.)	0.93	1.78	1.52	0.85
	2×2×2	2.01	0.47	1.24	1.54
	2×2×4	2.13	1.26	1.24	0.89
	4×2×2	1.21	1.08	1.80	0.72
Qwen 7B	4×2×4 (orig.)	1.36	2.52	1.56	1.15
	2×2×2	1.42	1.45	1.56	0.13
	2×2×4	1.91	1.82	1.25	0.66
	4×2×2	1.18	2.00	1.75	0.82

generates the largest *absolute* cross-talk (e.g., in Qwen-7B swapped 2×2×4, shape-owned directions generate 3.01 total off-diagonal damage vs. 1.22 for domain and 1.07 for register). Second, **cardinality does not determine direction**: the 4-class concept is not always the most *invasive by ratio*. In both swapped conditions, the concept with only 2 classes sometimes produces a higher invasiveness ratio than the 4-class concept (e.g., GPT-2 swapped 2×2×4: domain has ratio 2.13 despite having 2 classes, vs. shape's 1.24 with 4 classes).

The critical test is equalization. In Qwen-7B, collapsing to 2×2×2 reduces the invasiveness spread from 1.15 to 0.13—near-perfect equalization. Cross-talk asymmetry between concept pairs drops correspondingly (domain↔register: 0.498 → 0.024). GPT-2, however, shows the opposite: spread *increases* from 0.85 to 1.54 at equal cardinality, suggesting its smaller representation lacks the capacity to distribute concepts symmetrically when all are binary.

The interpretation is that cross-talk has two components: a **bandwidth component** that scales with information content (bits per concept, controlled by cardinality) and a **structural component** that depends on how the model internally organizes different concept types. Larger models with richer representations show cardinality-driven bandwidth allocation—equalize the bits and the cross-talk equalizes. Smaller models show more concept-type dependence, possibly because their limited capacity forces asymmetric encoding strategies.

4.13 Superlinear amplification: pairwise vs. triple entanglement

A critical question for both information theory and organizational design is whether entanglement scales linearly or superlinearly with the number of simultaneously tracked concepts. We test this by running the full SVD + damage pipeline on each concept pair (domain+register, domain+shape, register+shape), on all three together (control), and on a nested configuration where domain×register is treated as a single 8-class concept alongside shape.

Table 11: Entanglement intensity by concept configuration. Pairwise EIs are all < 1.0; triple EI exceeds 1.0. The amplification factor (triple / mean pairwise) is $\sim 2\times$, showing superlinear scaling. Nesting two concepts into one reduces EI below the corresponding pair. 95% CIs from 2,000 stratified bootstrap iterations.

Model	**Configuration**	**Rank**	**EI**	**Ratio**	**95% CI**
GPT-2	dom+reg	4	0.701	0.52	[0.483, 0.815]
	dom+shp	6	0.898	0.67	[0.717, 0.924]
	reg+shp	4	0.559	0.42	[0.455, 0.750]
	dom+reg+shp (triple)	7	1.346	1.00	[1.273, 1.594]
	nested+shp	10	0.735	0.55	—
Qwen 7B	dom+reg	4	0.675	0.47	[0.215, 0.857]
	dom+shp	6	0.737	0.51	[0.510, 0.839]
	reg+shp	4	0.599	0.41	[0.241, 0.832]
	dom+reg+shp (triple)	7	1.444	1.00	[0.965, 1.537]
	nested+shp	10	0.616	0.43	—

The results reveal **superlinear amplification**: triple entanglement exceeds the mean pairwise entanglement by a factor of 1.87× (GPT-2) and 2.15× (Qwen-7B). Every pairwise EI is below 1.0—two concepts alone produce only moderate cross-contamination. But the triple EI exceeds 1.0 in both models, meaning the combined encoding crosses a qualitative threshold where off-diagonal damage exceeds diagonal damage. Bootstrap CIs confirm the separation: all six pairwise CIs have upper bounds below 1.0 (GPT-2) or well below the triple CI lower bound (Qwen-7B). The gap between pairwise and triple is not a sampling artifact. The third concept does not merely add its own cross-talk; it amplifies interference between the existing two.

The nesting result provides the mechanism: when domain×register is treated as a single 8-class concept (rank 10 vs. rank 7 for the triple), EI *drops* to 0.62–0.74, below even the pairwise average. Bundling two concepts into one eliminates the cross-talk between them, because the SVD can no longer assign ownership of directions to separate concepts—they share a single concept identity. Nesting *reduces* the number of independent concept axes even as it increases the informative subspace rank, and it is the number of axes, not the rank, that drives entanglement.

The implication is precise: **entanglement scales superlinearly with the number of independently measured concept dimensions.** This is not merely an artifact of having more directions to damage—the pairwise configurations with rank 4–6 already have enough directions for entanglement to appear if it were rank-driven. Instead, the superlinearity emerges because each concept's directions carry information about all other concepts, and each additional concept axis adds a new source of cross-contamination to every existing direction.

4.14 Concept-type independence: software engineering replication

Every experiment so far uses the same three concept types: domain (content topic), register (stylistic formality), and reasoning shape (logical structure). A skeptic could argue that entanglement is a property of *these particular concepts*—perhaps content, style, and structure are inherently entangled in natural language, and the phenomenon would vanish with different concept types. We test this by constructing an entirely new factorial probe set drawn from software engineering, with concept dimensions that have no semantic overlap with the original three.

The new dimensions are:

- **Type system** (2 classes): statically typed vs. dynamically typed programming languages
- **Application area** (4 classes): web development, systems programming, data science, infrastructure
- **Programming paradigm** (4 classes): imperative, functional, concurrent, declarative

This yields the same $2 \times 4 \times 4 = 32$ cell structure with 5 repetitions per cell (160 probes). Each probe describes a concrete programming scenario grounded in specific languages and frameworks (e.g., Rust async runtimes, Python pandas pipelines, Haskell monadic IO), ensuring the model processes genuine technical content rather than abstract category labels.

Table 12: Entanglement intensity with software engineering concepts vs. original linguistic concepts. Both concept sets produce EI > 1.0 at terminal layers, and superlinear amplification replicates with the SE concepts. 95% CIs from 2,000 stratified bootstrap iterations.

	Original (dom/reg/shp)			**SE (type/area/paradigm)**		
Model	Triple EI	95% CI	Pair mean	Triple EI [95% CI]	Pair mean	SE/Orig
GPT-2	1.346	[1.27, 1.59]	0.719	1.441 [1.16, 1.60]	0.693	1.07
Qwen 7B	1.444	[0.97, 1.54]	0.670	1.242 [1.22, 1.57]	0.732	0.86

Both models show EI > 1.0 with the software engineering concepts (Table 12), with all four triple CIs entirely above 1.0 (or, for Qwen-7B original, above 0.97). GPT-2 actually shows *higher* SE entanglement than original (1.441 vs. 1.346), while Qwen-7B shows moderately lower (1.242 vs. 1.444). The mean SE/Original ratio is 0.97, indicating that entanglement intensity is comparable regardless of concept type.

Superlinear amplification replicates: SE triple EI exceeds mean pairwise EI by 2.08× (GPT-2) and 1.70× (Qwen-7B), with a mean of 1.89×—consistent with the 2.01× mean from the original concepts. The cross-talk structure also mirrors the original pattern: higher-cardinality concepts (application area and paradigm, each with 4 classes) generate more cross-talk than the 2-class concept (type system), consistent with the bandwidth-cardinality relationship established in the cardinality experiment.

The baselines confirm that both models encode the SE concepts well: GPT-2 achieves LOO-CV accuracy of 85.0% (type system), 78.7% (area), and 80.6% (paradigm); Qwen-7B achieves 96.3%, 76.2%, and 92.5% respectively. These are comparable to the original concept baselines, ruling out the possibility that low discriminability could artificially inflate entanglement.

Conclusion: Structural entanglement is concept-type independent. It appears with the same qualitative character—EI > 1.0, superlinear amplification, cardinality-dependent cross-talk bandwidth—whether the three concept dimensions describe linguistic properties (domain, register, shape) or software engineering properties (type system, application area, paradigm). The phenomenon is a property of how transformers encode *any* simultaneously present concept structure, not of the specific concepts involved.

5 The Resolution: Structural Entanglement

5.1 Why the V-matrix and damage matrix dissociate

The V-matrix decomposes the weight matrix $W = U\Sigma V^\top$. Each row of $V^\top$ records how that direction's regression weight distributes across concept labels. The V-matrix is a property of the *mapping from activations to labels*—the classifier's strategy.

The damage matrix measures something different: the information content of the *activations* projected onto each direction. When we remove direction u_i and re-classify, we are asking: does the *remaining* activation space still contain enough information for each concept? The answer is no, for all directions and all concepts. This means the 7-dimensional informative subspace (spanned by the columns of U) encodes all three concepts jointly in every direction.

The dissociation arises because a linear regression can impose concept-pure weights on concept-entangled activations. The regression assigns direction u_0 to shape by giving it large weights on shape labels and small weights on domain/register labels. But the activation values Xu_0 carry domain and register information that the regression simply ignores. Removing u_0 destroys that information, damaging all concepts.

5.2 How structural entanglement explains each prior finding

Double dissociation (Paper 4). Domain INLP removal preserves shape classification at $\geq$95.6%. This is real at the classifier level: the shape classifier uses different directions than the domain classifier. But the activations along domain directions *do* carry shape information—removing the domain INLP direction drops shape by 39.4% in the factorial experiment. Paper 4's probe design did not test for this because it measured shape *classification accuracy*, not shape *information content* along domain directions. The dissociation is between classifiers, not between representations.

Spectral isotropy (Paper 26). The forward pass amplifies INLP and random directions identically (ratio 0.99). This now has a mechanistic explanation: INLP directions are "generic" in the activation sense because they carry information about *all* concepts, not just the concept they were found for. The forward pass treats them as generic because they *are* generic—they encode the full joint concept distribution, just like any other direction in the informative subspace. The isotropy is a consequence of structural entanglement, not a coincidence.

Classification-intervention gap (Paper 29). Classification succeeds because classifiers can impose concept-pure strategies on concept-entangled representations. Intervention fails because it operates on the activations directly, where no direction is concept-pure. The "different constraints" identified in Paper 29 are precisely the difference between discrimination geometry and activation geometry.

Causal anti-selectivity (Paper 30). INLP is the only basis with positive selectivity because INLP targets discrimination geometry, where concept-pure directions exist. Patching, contrastive, and gradient bases target activation geometry, where every direction is structurally entangled. Causal intervention in a structurally entangled space is anti-selective by construction—you cannot surgically remove one concept from activations that encode all concepts in every direction.

Proportional compression (Paper 31). The natural language interface drops ∼80% of information proportionally across concepts. If all informative directions encode all concepts jointly, then any lossy channel that drops directions will drop all concepts proportionally. The proportionality is not coincidental—it is a direct consequence of entangled encoding. The concentration barrier ($k/d_{\text{eff}} \to 0$) is the *right bound*, but the mechanism is entanglement: every direction lost takes all concepts with it.

Asymmetric entanglement (Paper 41). The register direction causing 52.5% domain damage is a special case of structural entanglement. Paper 41 discovered the phenomenon in one direction pair; the factorial decomposition shows it holds universally across all informative directions and all concept pairs. The asymmetry (register removal damages domain more than domain removal damages register) reflects the asymmetric singular value structure: register information concentrates in fewer directions, so each is more informationally dense.

5.3 The concentration barrier as consequence

Previous work identified the concentration barrier—selectivity bounded by k/d_{eff}, with mean-pooled d_{eff} collapsing to $\sim$1.0—as the fundamental limit on interpretability interventions. The factorial decomposition reveals this is a consequence, not a cause.

The effective dimensionality collapses because the informative subspace has rank 7 (set by the label structure, not by the ambient dimensionality), and within this 7-dimensional space, all directions carry all concepts. The concentration barrier formalizes this as a ratio, but the underlying mechanism is simpler: the model stores concepts in jointly entangled directions. There is no 3,584-dimensional obstacle—there is a 7-dimensional space where everything is entangled.

5.4 Entanglement as a dimension-dependent mathematical property

A key question is whether entanglement is learned by the model or is an intrinsic consequence of projecting multi-concept information through high-dimensional space. To distinguish these, we compare entanglement intensity across five conditions using Qwen 2.5-7B's terminal-layer activations as the source (3,584 dimensions):

1. **Learned**: the model's actual activations (EI = 1.499)
2. **Random projection**: project learned activations through random Gaussian matrices to target dimensions 7–1,792
3. **PCA**: project learned activations via PCA to the same target dimensions
4. **Shuffled labels**: learned activations with randomly permuted concept labels (EI = 0.401 ± 0.089)
5. **Pure noise**: Gaussian random vectors replacing activations entirely (EI = 0.366 ± 0.067)

Table 13 summarizes the results.

Three findings emerge, now replicated across three architectures spanning a 60× parameter range. First, random projections at 448+ dimensions match the learned representation's entanglement in all three models (GPT-2: 1.48 ± 0.04; Qwen-0.5B: 1.29 ± 0.11; Qwen-7B: 1.50 ± 0.05). At the informative subspace rank (dim=7), random projections show near-

Table 13: Entanglement intensity under random projection across three architectures (v9b). The JL transition occurs at $m/r \approx 32$ in all models regardless of native dimension d. PCA concentrates information into concept-pure directions, reducing EI well below the random projection baseline at all tested dimensions.

		EI		
Condition	**Dim**	**GPT-2** ($d{=}768$)	**Qwen-0.5B** ($d{=}896$)	**Qwen-7B** ($d{=}3584$)
Learned (full)	—	1.437	1.391	1.499
Random proj.	7	0.36 ± 0.19	0.16 ± 0.11	0.18 ± 0.10
Random proj.	28	0.35 ± 0.12	0.34 ± 0.11	0.41 ± 0.12
Random proj.	112	0.28 ± 0.17	0.27 ± 0.07	0.45 ± 0.23
Random proj.	224	1.34 ± 0.22	0.76 ± 0.10	1.30 ± 0.07
Random proj.	448	1.48 ± 0.04	1.29 ± 0.11	1.50 ± 0.05
PCA	7	0.470	0.339	0.312
PCA	28	0.027	0.086	0.099
PCA	112	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

baseline EI (0.16–0.36). The transition occurs at $m/r \approx 32$ (dim=224) in GPT-2 and Qwen-7B, and at $m/r \approx 64$ (dim=448) in Qwen-0.5B.

Second, PCA *reduces* entanglement in all three models by concentrating information into high-variance, concept-pure directions. At dim=28, PCA achieves EI < 0.10 in all models, while random projections at the same dimension produce EI ≈ 0.35–0.41.

Third, shuffled labels and pure noise produce EI ≈ 0.37–0.51 across all models, establishing the null baseline. The learned representations exceed this by 2.7–4.1×, and random projections at dim=448+ reproduce this ratio, confirming entanglement is not a LOO-CV artifact.

The interpretation is precise: **entanglement is a mathematical consequence of high-dimensional distributed encoding, independent of architecture, parameter count, or hidden dimension.** When concept information is spread across $d \gg k$ dimensions, any random m-dimensional view with $m > 32k$ will find concepts inextricably co-encoded. PCA can partially reverse this by concentrating into the k most concept-relevant directions, but the full d-dimensional representation is entangled by construction.

6 Discussion

6.1 Implications for linear probing

The concept-purity assumption—that a direction found by probing for concept c carries primarily concept c—fails for every informative direction we tested. This does not mean linear probing is useless. It correctly identifies *which directions discriminate* each concept. But it is silent on whether those directions carry only that concept's information. The distinction matters for any application that moves beyond classification to intervention: erasure, steering, causal analysis, or feature circuit analysis.

6.2 Multi-concept probing

Factorial direction decomposition recovers both the discrimination structure (V-matrix) and the activation structure (damage matrix) in a single analysis. By probing multiple concepts simultaneously, it avoids the concept-purity assumption entirely. The 97% fidelity of the half-factorial design (96 probes) suggests this approach is practical for routine use. A balanced factorial design over k concept dimensions with n replications per cell requires $n \cdot \prod_i |c_i|$ probes—manageable for the 3–5 concept dimensions typical of interpretability studies.

6.3 Implications for mechanistic interpretability

Structural entanglement does not mean interpretability is impossible. It means interventions on *individual* directions will always produce collateral damage, because no individual direction is concept-pure. Methods that operate on *subspaces* rather than individual directions can mitigate this by controlling the trade-off between extraction and preservation.

The superposition hypothesis [Elhage et al., 2022] predicts that models encode more features than they have dimensions, using distributed representations. Sparse autoencoders [Bricken et al., 2023, Cunningham et al., 2023] attempt to recover monosemantic features from these distributed representations. Our findings raise a question for the SAE program: if entanglement is universal within the informative subspace, do SAE features achieve genuine monosemanticity, or do they impose concept-pure *discrimination* structure on concept-entangled activations—the same dissociation we observe with ridge regression and INLP? Applying the damage matrix test (remove an SAE feature direction and measure all concept accuracies) would answer this directly.

The linear representation hypothesis [Park et al., 2024] posits that concepts are encoded as linear directions in activation space. Our results are consistent with this hypothesis at the classifier level (the V-matrix shows clean linear structure) but reveal that the *information* *content* of those linear directions is richer than the hypothesis assumes. Directions are linear; they are not concept-pure.

6.4 Communication-theoretic interpretation

The experimental results map onto a communication-theoretic framework. A transformer layer is a shared channel carrying multiple concept signals simultaneously:

1. **Bandwidth allocation**: Concepts with higher information content (more classes) claim proportionally more channel capacity (more SVD directions).
2. **Asymmetric cross-talk**: Information leakage direction depends on concept type, not just information content.
3. **Superlinear complexity**: Adding a third concept dimension amplifies cross-talk between the existing two by $\sim 2\times$.
4. **Nesting as compression**: Bundling two concepts into a composite reduces entanglement below the pairwise baseline, at the cost of independent sensitivity.

6.5 Limitations

Regularization sensitivity. Ridge regression with $\alpha = 1.0$ distributes weight across directions via the L2 penalty, which could in principle amplify or suppress measured entanglement relative to an unregularized classifier. We conducted a systematic α sensitivity analysis on GPT-2 (layer 11) across $\alpha \in \{0, 0.001, 0.01, 0.1, 1.0, 10.0, 100.0\}$, where $\alpha = 0$ is ordinary least squares (OLS) with no regularization.

Table 14: Entanglement intensity as a function of ridge regularization strength α in GPT-2 (layer 11, 160 factorial probes). OLS ($\alpha = 0$) yields the highest EI. V-matrix purity is stable across α.

α	**EI**	**Avg V-Purity**	**Dom base**	**Shp base**
0 (OLS)	1.555	0.747	0.956	0.838
0.001	1.553	0.747	0.956	0.838
0.01	1.519	0.747	0.956	0.838
0.1	1.175	0.748	0.956	0.831
1.0	0.407	0.757	0.963	0.831
10.0	0.094	0.763	0.969	0.850
100.0	0.088	0.658	0.981	0.812

The result is the opposite of the concern: regularization *suppresses* measured entanglement, not amplifies it. OLS produces EI = 1.555—the highest value in the sweep—while $\alpha = 1.0$ gives EI = 0.407, a conservative underestimate. V-matrix purity remains stable ($\sim$0.75) across all α values, confirming that the discrimination–activation dissociation exists at every regularization strength.

The mechanism is that higher α makes the LOO-CV damage classifier more robust to individual direction removal: the L2 penalty distributes information more broadly, so removing any single direction causes less damage. At OLS, the classifier is more fragile and direction removal causes large cross-concept drops (0.44–0.78 per direction). A control experiment on GPT-2 (varying the SVD α while holding the damage classifier α fixed at 1.0) shows only modest EI variation (0.38–0.45 for SVD $\alpha \in [0.001, 10]$), confirming that in GPT-2 the EI sensitivity is primarily driven by the damage classifier's regularization, not the SVD decomposition. The same control at 7B shows a wider range: EI is near zero at SVD $\alpha \leq 0.01$ (where the SVD is numerically unstable) but stabilizes to 1.20–1.38 for SVD $\alpha \in [0.1, 100]$ with fixed damage $\alpha = 1.0$, confirming that the phenomenon is robust once the SVD is well-conditioned.

We replicated this analysis on Qwen 2.5-7B (layer 27, d = 3,584). At 7B scale, OLS ($\alpha = 0$) is numerically ill-conditioned—the first singular value of W exceeds the second by a factor of 3×10^6, and the leading direction has zero damage because it captures the mean signal rather than concept-specific variation. Within the well-conditioned range ($\alpha \in [0.1, 10]$), EI ranges from 0.898 ($\alpha = 10$) to 1.310 ($\alpha = 1.0$), and V-matrix purity is stable at 0.82–0.83. The discrimination–activation dissociation replicates fully at 7B.

The practical consequence is that our reported EI values at $\alpha = 1.0$ are *lower bounds* on the well-conditioned entanglement at GPT-2 scale, and are in the center of the well-conditioned range at 7B scale. The core qualitative finding—every direction damages all concepts when removed—holds across both models at all α values where the SVD is numerically stable.

Rank-dimensionality confound. The weight matrix W has rank 7, producing exactly 7 non-trivial SVD directions. Removing 1 of 7 directions from a rank-7 space necessarily reduces the representational capacity. The question is whether the *universal* pattern (all concepts damaged by any removal) is a mathematical consequence of the rank-7 constraint, or whether it reflects genuine co-encoding in the activations. The random projection experiment provides evidence for the latter: entanglement reproduces in random subspaces of dimension $\geq$448 (rank $\gg$ 7), suggesting the phenomenon is not an artifact of minimal rank. The pairwise experiments (§4.10) provide further evidence: rank-4 and rank-6 subspaces show EI $<$ 1.0, ruling out the hypothesis that any rank-constrained removal produces universal damage. Still, we emphasize that the damage matrix measures a classifier-relative quantity, not a direct activation-space property, and this distinction should inform interpretation.

PCA disentanglement. PCA to 28 dimensions reduces EI below 0.10 in all models (Table 13). This means concept-pure directions *do exist*—PCA finds them by concentrating variance into the highest-variance, most concept-aligned directions. Entanglement is therefore a property of *distributed* representations (where concept information is spread across many directions), not an inescapable property of the activations themselves. The claim is that transformer representations are naturally distributed in high dimensions, and that any method operating in the full activation space will encounter entanglement—not that no basis transformation could disentangle them. PCA-based probing may satisfy concept-purity, but at the cost of discarding the distributed structure that the model actually uses.

Additional limitations. All experiments cover four models across two architecture families (GPT-2 and Qwen); whether the phenomenon replicates in non-autoregressive architectures (e.g., encoder-only or diffusion-based models) or in modalities beyond language is untested. Both probe sets share a $2 \times 4 \times 4$ factorial structure; whether the phenomenon holds for higher-dimensional factorial designs or non-factorial concept relationships remains open. The LOO-CV damage metric is well-calibrated but computationally expensive. Dense layer profiling shows entanglement is present from layer 0 in all four models (EI = 0.17–0.54), suggesting it may originate in the tokenizer-embedding interaction. Crystallization dynamics are architecture-specific (GPT-2 crystallizes earlier than the larger Qwen-0.5B), but which architectural features determine the transition depth—hidden dimension, attention head count, or layer normalization—remains open.

Stability under fine-tuning. Subsequent work [McEntire, 2026f] shows that entanglement intensity responds differently to fine-tuning across model families. Under QLoRA fine-tuning with a natural language companion (code+NL), Qwen-32B's EI collapses to zero across all 8 seeds—a complete destruction of the measured entanglement structure, confirmed genuine by three independent probe sets (including 132 probes from real datasets). However, the same training protocol *increases* EI on CodeLlama-7B (from 0.874 to 1.093–1.192) and only modestly decreases it on DeepSeek-Coder-6.7B (from 1.376 to 1.196). Baseline EI also varies across families: DeepSeek-6.7B starts at 1.376, CodeLlama-7B at 0.874, and Qwen-32B at 0.622. These results confirm that entanglement is present across model families (consistent with the universality claim) but show that its response to perturbation is architecture-dependent. The Qwen-specific collapse raises the question of what architectural or pre-training features make certain models vulnerable to NL-induced disentanglement—a direction for future investigation.

7 Conclusion

Linear probing finds concept-pure *classifiers*. It does not find concept-pure *representations*. The directions a classifier uses for one concept carry all concepts in their activations. Within the low-rank informative subspace, every direction is entangled across every concept we measured, in every model we tested, with both linguistic and software engineering concept types.

The dissociation between discrimination geometry and activation geometry explains why classifiers succeed where interventions fail, why the forward pass treats concept directions as generic, and why lossy channels compress concepts proportionally.

Cross-model experiments show EI > 1.0 in GPT-2, Qwen-0.5B, Qwen-7B, and Qwen-7B-Instruct. Entanglement follows a phase transition whose dynamics are architecture-specific, with all models saturating at EI $\sim$1.4–1.6. Random projections to $\geq$448 dimensions reproduce the full learned EI across a 60× parameter range, while PCA concentrates information into concept-pure directions. Entanglement is a property of distributed high-dimensional encoding; PCA can partially reverse it by concentrating into the k most concept-aligned directions, but the full-dimensional representation is entangled by the geometry of the space.

Entanglement scales superlinearly with concept count (2× amplification from pairwise to triple), and nesting two concepts into one reduces entanglement below the pairwise baseline—confirming that it is the number of independently tracked concept axes, not the informative subspace rank, that drives cross-contamination.

An important methodological caveat: the damage matrix measures a classifier-relative quantity (accuracy loss under direction removal in a ridge regression), not a direct activation-space property. A systematic α sensitivity analysis (§6.5) confirms that regularization *suppresses* rather than amplifies measured entanglement: OLS ($\alpha = 0$) gives the highest EI, and the reported $\alpha = 1.0$ values are conservative. V-matrix purity remains stable across the full α range in both GPT-2 and Qwen-7B, confirming that the discrimination–activation dissociation is not an artifact of the regularization choice. EI measurements have non-trivial run-to-run variance ($\sim$14% across different probe instantiations at 7B scale). Stratified bootstrap confidence intervals (2,000 iterations, Tables 7–12) confirm that the core findings—EI > 1.0 for triple configurations and EI < 1.0 for all pairwise configurations—hold at the 95% confidence level across all four models and both concept types.

The practical consequence is that any interpretability method claiming to isolate a concept in activation space—including sparse autoencoders [Bricken et al., 2023]—should demonstrate isolation in the damage matrix, not just in classifier weights. Multi-concept factorial decomposition provides a direct way to make this measurement.

References

T. Bricken, A. Templeton, J. Batson, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey. Sparse autoencoders find highly interpretable features in language models. In *ICLR*, 2024.

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

N. Elhage, T. Hume, C. Olsson, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

K. Park, Y. J. Choe, and V. Veitch. The linear representation hypothesis and the geometry of large language models. In *ICML*, 2024.

S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proc. ACL*, 2020.

J. McEntire. The shape of the problem: Domain-invariant structural signatures in activation space. Working paper, 2026.

J. McEntire. Spectral geometry of the forward pass: How INLP directions interact with layer Jacobians. Working paper, 2026.

J. McEntire. The activation geometry program: Twelve papers on the mathematical structure of neural network representations. Working paper, 2026.

J. McEntire. Causal basis discovery for domain-selective noise injection. Working paper, 2026.

J. McEntire. The inter-instance compression barrier: Domain-specific information loss at the natural language interface. Working paper, 2026.

J. McEntire. Entangled directions in transformer activation space: INLP blind spots, asymmetric feature coupling, and partial extraction efficiency. Working paper, 2026.