Universal Entanglement in Transformer Activation Space:

Discovery, Replication, and the Discrimination–Activation Dissociation

Jeremy McEntire*

Abstract

When you remove a concept from a neural network, other unrelated concepts break. This is not a bug—it reveals fundamental structure. We discover, formalize, and replicate *universal entanglement*: concept directions in transformer activation space are asymmetrically coupled, so that removing one direction damages classification of concepts the direction was never trained to represent.

The discovery comes from a single model. In Qwen 2.5-7B, a register direction (formal vs. informal) causes a 52.5% drop in domain classification accuracy when removed, despite lying 82% outside the 36-dimensional domain subspace found by Iterative Null-Space Projection (INLP). Thirty-six iterations of domain-focused INLP missed this direction entirely. The entanglement is asymmetric: removing domain directions causes zero register damage. It is learned: register directions from domain-neutral prompts carry no domain information and have low cosine similarity (≤ 0.28) with the entangled direction.

The replication uses factorial direction decomposition—multi-output ridge regression over three simultaneously varied concept dimensions, followed by SVD—to measure both *discrimination geometry* (how classifiers use each direction) and *activation geometry* (what information each direction carries). The SVD's V-matrix shows clean concept separation: each direction loads primarily on one concept. The damage matrix tells a different story: every informative direction causes 39–88% accuracy drops across *all three* concepts. Directions that are concept-pure in discrimination geometry are concept-entangled in activation geometry.

Eight experiments establish scope. Cross-model replication in GPT-2, Qwen-0.5B, Qwen-7B, and Qwen-7B-Instruct shows entanglement intensity (EI) > 1.0 at all terminal

*Correspondence: jmc@cageandmirror.com

layers. Entanglement follows an S-curve phase transition with architecture-specific dynamics. Random projections to ≥448 dimensions reproduce the full learned EI across a 60× parameter range, while PCA concentrates information into concept-pure directions, reducing EI below 0.1 at 28 dimensions—demonstrating that entanglement is a property of the informative subspace, not the full representation. Entanglement scales superlinearly with concept count (2× amplification from pairwise to triple) and 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.

1 Introduction

When you remove a concept from a neural network, other concepts break. Not because the removal was careless, and not because the concepts are related. A direction found by probing for linguistic register—formal versus informal—destroys the network's ability to distinguish medical text from legal text, code from science. The domain classifier never used that direction. The register classifier did. Remove it, and domain collapses.

This happens because the direction carries both concepts simultaneously, encoded in a single vector by a training process that found joint representation efficient. The register direction is a hub in the model's representational geometry: one dimension doing double duty. Standard interpretability tools, designed to find single-concept directions, cannot see it.

We call this *universal entanglement*: when removing one piece of learned information—such as the distinction between formal and informal language—unexpectedly damages the model's ability to classify completely different information, such as whether text is medical or legal.

This paper reports the discovery, formalization, and cross-architecture replication of *universal entanglement* in transformer activation spaces. We use "universal" to describe the phenomenon's consistency across all tested base models in their pre-trained state, not to claim it holds in all possible neural networks. As shown in companion work [McEntire, 2026i], fine-tuning with sufficient capacity can reduce entanglement to zero, demonstrating that it reflects the model's representational strategy rather than a geometric inevitability. The phenomenon is straightforward: concept directions found by linear probing are concept-pure in their classifier weights but concept-entangled in their activation values. A direction used exclusively for domain classification carries domain, register, and shape information simultaneously. The classifier selects which aspect to attend to; it does not determine what

information the activation carries.

The story has three parts. First, the discovery: a single register direction in Qwen 2.5-7B drops domain accuracy 52.5% while 36 iterations of domain-focused INLP [Ravfogel et al., 2020] miss it entirely (Section 3). The z-score against random controls exceeds 5000. The entanglement is asymmetric—register removal destroys domain, but domain removal leaves register untouched—and learned rather than architectural.

Second, the formalization: we introduce factorial direction decomposition (multi-output ridge regression + SVD) and the *Entanglement Index* (EI), which quantifies the ratio of cross-concept to same-concept damage when directions are removed (Section 4). The V-matrix captures discrimination geometry; the damage matrix captures activation geometry. Their dissociation is the central finding.

Third, the replication: eight experiments across four architectures spanning a 60× parameter range establish that entanglement is universal, dimension-dependent, and superlinear in concept count (Section 5). PCA disentanglement to 28 dimensions reduces EI below 0.1—a key finding that localizes entanglement to the informative subspace rather than the full representation.

1.1 Prior work on concept directions

The linear probing paradigm treats transformer activations as a space where concepts are encoded along separable directions [Gurnee and Tegmark, 2024, Burns et al., 2023, Park et al., 2024]. INLP [Ravfogel et al., 2020] operationalizes this: given labeled data for a target concept, it iteratively finds discriminative hyperplanes and projects them out, producing directions that capture the concept's linear footprint. The assumption is that these directions constitute the concept's representation—that single-concept search suffices to recover single-concept structure.

This assumption underlies every major concept-erasure and concept-isolation technique in the literature. Distributed Alignment Search [DAS; Geiger et al., 2024] searches for low-dimensional subspaces that mediate causal effects on target concepts, assuming such subspaces can be found independently for each concept. LEACE [Belrose et al., 2023b] provides the optimal linear erasure of a target concept, explicitly designed for single-concept removal. RLACE [Ravfogel et al., 2022] casts concept erasure as an adversarial game, again targeting one concept at a time. ROME [Meng et al., 2022] edits factual associations by modifying individual MLP layers, treating each fact as independently localized. Activation patching [Conmy et al., 2023] and representation engineering [Marks et al., 2024] both operate on the premise that concept-specific directions can be identified and manipulated in isolation.

All of these methods search for individual concept directions. This paper characterizes the

degree and structure of the entanglement that single-concept search leaves unquantified—a structural limitation worth quantifying.

1.2 Contributions

1. **INLP blind spot (discovery).** We demonstrate that single-concept INLP systematically misses the most informative entangled directions in activation space. Thirty-six iterations of domain INLP failed to find a direction that carries more domain-destructive power than all 36 combined.

2. **Factorial direction decomposition.** We introduce multi-output ridge regression + SVD over balanced factorial probes, producing directions with both structural (V-matrix) and functional (damage matrix) characterizations.

3. **Discrimination–activation dissociation.** The V-matrix shows concept-pure directions; the damage matrix shows those same directions carry universal cross-concept information.

4. **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.

5. **Dimension dependence and PCA disentanglement.** Random projections to $\geq$448 dimensions reproduce the full learned EI. PCA to 28 dimensions reduces EI below 0.1—demonstrating that entanglement is a property of the informative subspace, not the full representation.

6. **Superlinear amplification.** Triple entanglement exceeds mean pairwise by $2\times$, confirmed with both linguistic and software engineering concept types.

2 Background and Related Work

Mechanistic interpretability aims to understand neural networks by identifying which internal components encode which concepts. The dominant paradigm assumes that individual concepts occupy identifiable, separable directions or subspaces in activation space. This section surveys the methods built on that assumption.

2.1 Linear probing and concept directions

The simplest approach trains a linear classifier to predict concept labels from a model's internal activations. The classifier's weight vector identifies a "concept direction" [Alain and Bengio, 2017]. If the classifier succeeds, the concept is said to be linearly represented. This technique has been applied to parts of speech, sentiment, factual knowledge, spatial relations, and temporal order [Gurnee and Tegmark, 2024, Burns et al., 2023, Nanda et al., 2023]. The implicit assumption is that each concept's direction is primarily that concept's—that the weight vector identifies where the concept lives, not merely where a classifier can find it.

2.2 Concept erasure methods

INLP [Ravfogel et al., 2020] extends linear probing to concept erasure. It iteratively trains classifiers, extracts weight vectors, and projects activations onto the null space of each discovered direction. After k iterations, the resulting subspace is claimed to capture the concept's full linear footprint. The procedure is inherently single-concept: it takes one set of labels and searches for directions informative about those labels only.

LEACE [Belrose et al., 2023b] provides the theoretically optimal linear erasure for a target concept, minimizing the mutual information between the projected activations and the concept labels. RLACE [Ravfogel et al., 2022] frames erasure as an adversarial minimax game between an eraser and a probe. Both methods target one concept at a time.

2.3 Causal and editing methods

DAS [Geiger et al., 2024] searches for low-dimensional subspaces that mediate causal effects on specific behaviors, using interchange interventions to verify that patching a subspace transfers the target behavior. ROME [Meng et al., 2022] localizes factual associations to specific MLP layers and edits them by rank-one updates to the value matrix. Activation patching [Conmy et al., 2023] identifies which components contribute to a behavior by replacing activations from a clean run with those from a corrupted run. Representation engineering [Marks et al., 2024] identifies concept-relevant directions by contrasting activations across paired inputs and uses those directions for reading and controlling model behavior.

2.4 The shared assumption

Every method above searches for *individual* concept directions or subspaces. The assumption is that a direction found for concept A carries primarily concept A's information. This is the **concept-purity assumption**. It is testable: remove a direction found for concept

A and measure whether concepts B and C are damaged. If they are, the direction is not concept-pure.

The superposition hypothesis [Elhage et al., 2022] already suggests that models encode more features than they have dimensions. Sparse autoencoders [Bricken et al., 2023, Cunningham et al., 2023] attempt to recover monosemantic features from superposed representations. Our findings raise a question for these approaches: 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 superposition hypothesis already moves the field toward acknowledging that concept representations overlap. This paper takes the next step: systematically quantifying the degree and structure of concept entanglement across multiple architectures and concept types.

Recent work by Mueller et al. (2025) found that individual SAE features map to single concepts but single concepts distribute across many features—a one-to-many mapping [Mueller et al., 2025]. Our result is structurally stronger: the damage matrix reveals all-to-all entanglement where removing *any* concept direction damages *every* other concept. The discrimination–activation dissociation has no counterpart in their framework: their feature-level analysis operates entirely within discrimination geometry and does not measure activation-level collateral damage.

3 Discovery: The INLP Blind Spot

The phenomenon was first observed in a single model. This section reports the original finding: a register direction in Qwen 2.5-7B that carries more domain-destructive power than 36 iterations of domain-focused INLP, despite lying almost entirely outside the domain subspace.

3.1 Model and probes

We study Qwen 2.5-7B (28 transformer layers, hidden dimension 3584) at layers 7, 14, 21, and 27 (quarter-points plus terminal).

We construct 80 probes with balanced domain×register factorial structure: 4 domains (medical, legal, code, science) × 2 registers (formal, informal) × 10 probes per cell. Each probe is 1–3 sentences. Register varies independently of domain: each domain has exactly 10 formal and 10 informal probes. Any detected entanglement is from the model's representation, not the probe distribution.

As a held-out control, we use 32 register prompts from prior work [McEntire, 2026b]: 16 formal and 16 informal, all domain-neutral (institutional/conversational language without

domain-specific content).

3.2 Direction sets

Domain INLP directions. We use 36 pre-computed domain INLP directions, orthonormalized via QR decomposition to produce $Q_{\text{domain}} \in \mathbb{R}^{3584\times 36}$. These directions classify domain at $> 97\%$ leave-one-out cross-validated (LOO-CV) accuracy across all layers.

Register INLP directions. We compute register INLP separately at each layer by iterative ridge classification of formal vs. informal labels, projecting out each discriminative direction and repeating until LOO-CV accuracy falls below 55%. At every layer, a single direction achieves 100% accuracy and the second iteration drops below threshold, yielding $Q^{(\ell)}_{\text{register}} \in \mathbb{R}^{3584\times 1}$ per layer ℓ.

Domain-neutral register directions. From McEntire [2026b], 8 register directions computed from the 32 domain-neutral prompts via logistic regression and SVD, orthonormalized to $Q_{\text{p78}} \in \mathbb{R}^{3584\times 8}$.

3.3 Asymmetric entanglement

Table 1 shows the core finding. Removing the single register direction causes 52–70% domain accuracy drops. Removing 36 domain directions causes 0% register damage.

Table 1: Asymmetric entanglement: accuracy after feature removal at $\sigma = 1$.

Layer	Register Removal → Domain			Domain Removal → Register		
	Raw	After	Drop	Raw	After	Drop
7	0.975	0.412	56.2%	1.000	1.000	0.0%
14	0.988	0.287	70.0%	1.000	1.000	0.0%
21	0.988	0.375	61.3%	1.000	1.000	0.0%
27	1.000	0.475	52.5%	1.000	1.000	0.0%

The register direction is not merely correlated with domain. It is the most domain-informative single direction in 3584-dimensional space: removing any of 20 random unit vectors causes 0.0% domain damage (standard deviation: 0.0%), while the register direction causes 52.5% damage. The z-score exceeds 5000.

3.4 Where the register direction lives

At layer 27, the register direction projects 17.8% into the INLP domain subspace and 82.2% into the complement. Yet domain classification accuracy in the INLP complement is 100%—the complement carries full domain information.

Table 2: INLP coverage gap: domain accuracy in the 36-dimensional INLP subspace vs. its complement. Register direction overlap with INLP subspace.

Layer	Domain acc (INLP)	Domain acc (complement)	Register in INLP
7	0.875	0.975	11.9%
14	0.913	0.988	13.0%
21	0.938	0.988	15.5%
27	0.975	1.000	17.8%

The INLP subspace captures domain information, but the complement carries even more. The register direction found domain-informative variance that 36 iterations of domain INLP completely missed, because that variance was entangled with register and INLP was not looking for cross-concept structure.

3.5 Partial extraction efficiency

We define the partial projection operator:

$$P'(\sigma) = I - \sigma\, QQ^\top, \quad \sigma \in [0, 1] \tag{1}$$

where Q is the orthonormal basis of directions to remove. At $\sigma = 0$, no removal occurs. At $\sigma = 1$, full removal. The residual is $x_{\text{res}} = P'(\sigma)x$ and the extracted component is $x_{\text{ext}} = \sigma\, QQ^\top x$.

We introduce a dual-metric evaluation. Prior work used a conditional metric:

$$\text{joint}_{\text{key}}(\sigma) = \text{acc}_{\text{remove}}(x_{\text{ext}}) + \text{acc}_{\text{measure|key}}(x_{\text{res}}) \tag{2}$$

where the measured feature is classified per-group, conditioned on an external label. This metric absorbed the entanglement signal: within-register domain classification was undamaged even when between-register domain classification collapsed.

The raw metric surfaces the damage:

$$\text{joint}_{\text{raw}}(\sigma) = \text{acc}_{\text{remove}}(x_{\text{ext}}) + \text{acc}_{\text{measure}}(x_{\text{res}}) \tag{3}$$

At every layer, partial extraction at σ^* extracts enough register information for perfect classification while the residual retains nearly all domain information. Full removal ($\sigma = 1$) extracts the same register information but destroys domain.

Table 3: Register removal sigma sweep: optimal σ^* and joint information gains on raw vs. conditional (key) metrics.

Layer	σ^*_{raw}	$\text{joint}_{\text{raw}}(\sigma^*)$	$\text{joint}_{\text{raw}}(\sigma{=}1)$	Gain	Outcome
7	0.5	1.988	1.413	+0.575	A
14	0.1	1.988	1.288	+0.700	A
21	0.1	1.988	1.375	+0.613	A
27	0.1	2.000	1.475	+0.525	A

3.6 Pareto frontier geometry

Plotting domain preservation vs. register extraction across σ reveals the geometry of the feature tradeoff.

Table 4: Pareto frontier convexity: positive values indicate efficient partial extraction (convex frontier), negative values indicate threshold behavior.

Layer	Convexity	Shape
7	−0.005	L-shaped
14	+0.333	Convex
21	+0.333	Convex
27	+0.333	Convex

Layer 7 exhibits threshold behavior: domain accuracy remains stable until $\sigma \approx 0.7$, then drops sharply. Partial removal offers little advantage because the damage is all-or-nothing. Layers 14 through 27 exhibit convex frontiers: domain accuracy degrades gradually and register extraction saturates early. The "knee" of the curve occurs at low σ, meaning a small amount of removal captures register cleanly while preserving domain.

We quantify convexity as the signed area between the Pareto curve and the diagonal connecting its endpoints, normalized by the area of the bounding triangle. Positive values indicate a convex frontier (partial extraction is efficient); negative values indicate concavity or threshold behavior.

The transition between layers 7 and 14 marks where the representational hierarchy crystallizes. In early layers, register and domain are encoded in a shared direction with threshold coupling. In later layers, the model has distributed the shared information such that gentle probing can separate the components. This transition zone aligns with the layer-resolved selectivity peaks reported by Belrose et al. [2023a], where layers 7–10 show maximal per-neuron feature selectivity before representations reorganize into distributed codes at deeper layers.

3.7 Learned entanglement

Register directions from domain-neutral prompts (8 dimensions) produce a fundamentally different subspace than register directions learned from domain×register probes (1 dimension per layer).

Table 5: Domain-neutral register directions vs. entangled register direction.

Layer	Domain drop (neutral)	Domain drop (entangled)	$\cos_{\max}$	# dims (neutral / entangled)
7	0.0%	56.2%	0.087	8 / 1
14	0.0%	70.0%	0.108	8 / 1
21	0.0%	61.3%	0.136	8 / 1
27	0.0%	52.5%	0.281	8 / 1

The domain-neutral register subspace (8 dimensions) carries zero domain information at any layer. Its maximum cosine similarity with the entangled direction is 0.28 (layer 27), rising across depth but never approaching alignment. These are geometrically distinct subspaces that both classify register perfectly: one learned from domain-neutral text, one learned from domain-varied text.

The entanglement is not intrinsic to the architecture. The same model, probed with domain-neutral register prompts, produces register directions orthogonal to domain. Only when register is learned in the *context* of domain variation does the model's representation entangle them.

This is compression under selection pressure. The model's training data contains medical text that is predominantly formal and casual discussions that are predominantly informal. The training objective selects for representations that compress these co-occurrences into shared directions. The register direction encodes "formal-ness" in a way that inherently carries domain information, because in the training distribution, the specific flavor of formality *is* domain-informative.

4 Formalization

The discovery in Section 3 demonstrates entanglement between two concepts in one model. To establish universality, we need a systematic framework that can measure entanglement across arbitrary concept configurations and architectures. This section introduces factorial direction decomposition and the Entanglement Index.

We use "entanglement" throughout this paper to denote asymmetric, directional information coupling between concept representations in activation space. This is **not** quantum

entanglement. The term is borrowed for its descriptive accuracy—the coupling is non-local in the representation, direction-dependent, and not decomposable by independent analysis of each concept—but no quantum-mechanical formalism is implied or invoked.

4.1 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, capturing at FP16 and converting to float64 for analysis. Activations are centered (mean-subtracted) per layer.

4.2 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:

$$W = (X^\top X + \alpha I)^{-1} X^\top Y, \qquad W \in \mathbb{R}^{d \times 10} \tag{4}$$

with $\alpha = 1.0$ (sensitivity analysis in Section 6.4). The SVD of W gives:

$$W = U \Sigma V^\top \tag{5}$$

where $U \in \mathbb{R}^{d \times 10}$ 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.

4.3 V-matrix: discrimination geometry

The V-matrix answers: which direction does the classifier use for each concept?

The rows of $V^\top$ reveal how each direction's weight vector decomposes across concept

labels. For direction i, we compute:

$$\text{dom\_V}_i = \|V_i^\top[0{:}4]\|_2 \quad \text{(domain loading)} \tag{6}$$

$$\text{reg\_V}_i = \|V_i^\top[4{:}6]\|_2 \quad \text{(register loading)} \tag{7}$$

$$\text{shp\_V}_i = \|V_i^\top[6{:}10]\|_2 \quad \text{(shape loading)} \tag{8}$$

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.

4.4 Damage matrix: activation geometry

The damage matrix answers: what information is actually lost when we remove each direction?

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

$$X' = X - (Xu_i)u_i^\top \tag{9}$$

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

$$\Delta_c^{(i)} = \text{acc}_c(\text{baseline}) - \text{acc}_c(\text{after removal}) \tag{10}$$

If a direction is truly concept-pure, removing it should damage only its attributed concept. The damage matrix reveals the **activation geometry**: what information the activations along each direction actually encode, regardless of how the classifier uses them.

4.5 Entanglement Index

Intuitively, the Entanglement Index compares how much damage removing a direction does to *other* concepts versus its *own* concept. Values above 1.0 mean the direction hurts other concepts more than its own.

Definition 4.1 (Entanglement Index)**.** Let $D \in \mathbb{R}^{k \times C}$ be the damage matrix, where D_{ic} is the accuracy drop for concept c when direction i is removed. Let $\text{attr}(i)$ denote the concept to which direction i is attributed by the V-matrix (highest loading). The Entanglement Index is:

$$\text{EI} = \frac{\sum_i \sum_{c \neq \text{attr}(i)} D_{ic}}{\sum_i D_{i,\text{attr}(i)}} \tag{11}$$

EI is the ratio of total off-diagonal damage (cross-concept) to total diagonal damage (same-concept). EI > 1.0 means removing a direction damages other concepts more, in aggregate, than it damages the concept the direction was "built for."

4.6 LOO-CV measurement protocol

All accuracy measurements use leave-one-out cross-validation with ridge regression ($\alpha = 1.0$). For each held-out probe j:

1. Fit a ridge classifier on the remaining $n - 1$ probes.
2. Predict the held-out probe's concept labels.
3. Record correct/incorrect for each concept dimension.

We use the analytical hat-matrix formulation for computational efficiency: fit the full-sample ridge once, compute the hat matrix $H = X(X^\top X + \alpha I)^{-1} X^\top$, and derive LOO predictions via $\hat{y}_i^{(-i)} = (\hat{y}_i - H_{ii} y_i)/(1 - H_{ii})$. This produces exact LOO-CV in a single matrix computation.

5 Universality

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%.

5.1 V-matrix: clean concept separation

Table 6 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 6: V-matrix structural loadings at layer 27 of Qwen 2.5-7B. 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.

5.2 Damage matrix: universal entanglement

Table 7 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.

Table 7: Damage matrix at layer 27 of Qwen 2.5-7B. 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

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 (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 $\times$ 3 concepts) is 38.8%.

5.3 The dissociation

Tables 6 and 7 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 universal 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.

5.4 INLP cross-concept damage

Table 8 applies the concept-purity test directly to INLP directions. If INLP directions were concept-pure, removing the domain direction should damage only domain.

Table 8: Cross-concept damage from removing each INLP direction at layer 27 of Qwen 2.5-7B. 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 Section 3 finding exactly) and shape by 69.4%.

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

5.5 Cross-model replication

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 9 reports entanglement intensity and V-matrix purity at each model's terminal layer.

Table 9: Cross-model entanglement at terminal layers. 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).

Model	Params	Hidden	Terminal	EI	Avg Purity
GPT-2	124M	768	L11	1.437	0.638
Qwen 2.5-0.5B	494M	896	L23	1.391	0.622
Qwen 2.5-7B	7.6B	3,584	L27	1.499	0.691
Qwen 2.5-7B-Inst	7.6B	3,584	L27	1.527	0.604

All four models show EI > 1.0: more cross-concept damage than same-concept damage when any informative direction is removed. The phenomenon spans a 60× parameter range and two distinct autoregressive transformer architectures. Entanglement across autoregressive transformer architectures 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.

5.6 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 10: 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 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.

5.7 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.

RLHF accelerates crystallization: the Instruct model transitions to EI > 1.0 at depth fraction 0.222, versus 0.296 for the base model, and 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. The compliance behavior learned through RLHF is not encoded as a separable concept—it is woven into the existing entangled representation.

5.8 Cross-talk decomposition

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 damages concept B—reveals asymmetric structure.

Table 11: Terminal-layer cross-talk matrices. Each row shows how much removing that concept's directions damages each column concept. Shape is systematically the most invasive concept; 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	—

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). Second, register is the least invasive: register-

owned directions cause the smallest cross-talk (0.52–0.78). Third, cross-talk is asymmetric: shape→domain consistently exceeds domain→shape, with mean 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. Shape, as the concept capturing reasoning structure, 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.

5.9 Superlinear amplification

A critical question 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, on all three together, and on a nested configuration where domain×register is treated as a single 8-class concept alongside shape.

Table 12: Entanglement intensity by concept configuration. Pairwise EIs are all < 1.0; triple EI exceeds 1.0. The amplification factor is $\sim 2\times$, showing superlinear scaling.

Model	Configuration	Rank	EI	Ratio to triple
GPT-2	dom+reg	4	0.701	0.52
	dom+shp	6	0.898	0.67
	reg+shp	4	0.559	0.42
	dom+reg+shp (triple)	7	1.346	1.00
	nested+shp	10	0.735	0.55
Qwen 7B	dom+reg	4	0.675	0.47
	dom+shp	6	0.737	0.51
	reg+shp	4	0.599	0.41
	dom+reg+shp (triple)	7	1.444	1.00
	nested+shp	10	0.616	0.43

Triple entanglement exceeds the mean pairwise entanglement by 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. 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, 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. It is the number of independently tracked concept axes, not the informative subspace rank, that drives entanglement.

5.10 Concept-type independence

Every experiment above uses linguistic concepts: domain, register, and reasoning shape. A skeptic could argue that entanglement is a property of these particular concept types. We test this by constructing an entirely new factorial probe set drawn from software engineering:

- **Type system** (2 classes): statically typed vs. dynamically typed
- **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 (160 probes). Each probe describes a concrete programming scenario grounded in specific languages and frameworks.

Table 13: Entanglement intensity with software engineering concepts vs. original linguistic concepts. Both concept sets produce EI > 1.0 at terminal layers.

	Original (dom/reg/shp)		**SE (type/area/paradigm)**		**SE/Orig**
Model	Triple EI	Pair mean	Triple EI	Pair mean	Ratio
GPT-2	1.346	0.719	1.441	0.693	1.07
Qwen 7B	1.444	0.670	1.242	0.732	0.86

Both models show EI > 1.0 with software engineering concepts. GPT-2 shows *higher* SE entanglement (1.441 vs. 1.346), while Qwen-7B shows moderately lower (1.242 vs. 1.444). The mean SE/Original ratio is 0.97. Superlinear amplification replicates: SE triple EI exceeds mean pairwise EI by 2.08× (GPT-2) and 1.70× (Qwen-7B), consistent with the 2.01× mean from the original concepts.

Entanglement across autoregressive transformers is concept-type independent. It appears with the same qualitative character—EI > 1.0, superlinear amplification, cardinality-dependent cross-talk—whether the three concept dimensions describe linguistic properties or software engineering properties. The phenomenon is a property of how transformers encode *any* simultaneously present concept structure.

5.11 PCA disentanglement

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. We compare

entanglement intensity across five conditions using terminal-layer activations:

1. **Learned**: the model's actual activations
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
5. **Pure noise**: Gaussian random vectors replacing activations entirely

Table 14: Entanglement intensity under random projection and PCA across three architectures. The Johnson–Lindenstrauss transition occurs at $m/r \approx 32$ in all models regardless of native dimension d.

		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

Three findings emerge, 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-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.

This is a key finding, not a limitation. Entanglement is a property of the informative subspace, not the full representation. PCA escape demonstrates that the phenomenon is basis-dependent—and that is precisely the point. Methods that search in the full activation space encounter entanglement because concept information is distributed across far more dimensions than the concepts require. PCA can partially reverse this by concentrating into the k most concept-relevant directions. But the full d-dimensional representation—the one the model actually computes with—is entangled by the geometry of distributed encoding.

The random-projection control confirms that PCA escape is not trivial. Random projections to the same dimensionality (28 dimensions) do *not* disentangle: their EI (0.35–0.41) exceeds the PCA EI (0.03–0.10) by 4–13×. Disentanglement requires not merely dimension reduction, but reduction along the axes that concentrate concept variance—which is exactly what PCA does. Any method operating in a basis that does not concentrate concept variance (including the model's native activation basis) will encounter entanglement.

Entanglement as diagnostic. The concentration-of-measure bound (Theorem 1 in the companion theorem paper [McEntire, 2026j]) establishes the minimum entanglement for a given informative subspace rank r. The model determines r through its compression of training data. Entanglement is therefore diagnostic: high EI indicates aggressive compression into a low-rank subspace; low EI indicates the model has allocated sufficient representational capacity to separate concepts. Companion work on fine-tuning [McEntire, 2026i] confirms this interpretation: at 32B scale, fine-tuning with appropriate training signal drives EI to exactly zero across all seeds, demonstrating that the model can expand its informative subspace sufficiently to disentangle. This reframes entanglement from a fundamental limit to a measurable property of the model's compression strategy.

6 Discussion

6.1 What the dissociation means for interpretability

The finding is that concept directions in transformer activation spaces serve two roles simultaneously, and these roles dissociate under measurement. The V-matrix shows how a classifier *uses* each direction: cleanly, one concept per direction. The damage matrix shows what each direction *carries*: everything.

This dissociation has a direct consequence for every interpretability method that moves beyond classification to intervention. Linear probing correctly identifies which directions discriminate each concept. But the directions it finds are not concept-pure in their information content. Removing a "domain direction" does not remove only domain information—it

removes domain, register, and shape information simultaneously, because the activations along that direction encode all three.

INLP, LEACE, RLACE, DAS, activation patching, and representation engineering all search for concept-specific directions. Our results show that the directions they find will always carry collateral information. This is not a failure of the methods—it is a property of the representations they operate on. Any method claiming to isolate a concept in activation space should demonstrate isolation in the damage matrix, not just in classifier weights.

The qualitative dissociation—concept-pure discrimination geometry coexisting with concept-entangled activation geometry—is the core claim, not any specific EI threshold. Table 15 shows EI varies with regularization strength: OLS yields EI = 1.555, our conservative choice of $\alpha = 1.0$ yields EI = 0.407 in GPT-2. But at every tested regularization strength, the V-matrix remains concept-pure while the damage matrix remains concept-entangled. The dissociation is robust; only the magnitude changes.

6.2 Why single-concept erasure is structurally incomplete

The INLP blind spot from Section 3 illustrates the problem concretely. Domain INLP ran 36 iterations and found 36 directions that classify domain at $> 97\%$ accuracy. It missed the single most domain-informative direction in the space—the register direction—because that direction's domain information is accessible only through its register structure. INLP was not looking for register. It found domain-specific directions. It missed the domain-carrying direction.

This generalizes. Any single-concept search will find directions that discriminate the target concept. It will systematically miss directions that carry the target concept's information entangled with other concepts. The entangled directions are often more informative than the concept-specific ones, because they sit at hubs in the model's compressed representation where multiple concepts share representational resources.

Multi-concept probing—factorial direction decomposition, or more generally, any method that fits directions jointly across multiple concept dimensions—avoids this blind spot by construction. It does not assume concept-purity. It measures both discrimination geometry and activation geometry, and when these dissociate, it reveals the dissociation rather than hiding it.

6.3 Communication-theoretic interpretation

The entanglement structure is consistent with a communication-theoretic interpretation in which transformer layers function as shared channels with finite bandwidth, where concepts

with higher information content claim more capacity, cross-talk is asymmetric and concept-type dependent, and adding concept dimensions amplifies interference superlinearly. We leave formal development of this frame—including the connection to the compression-under-selection-pressure mechanism identified in Section 3.7—to future work.

6.4 Alpha sensitivity

Ridge regression with $\alpha = 1.0$ distributes weight across directions via the L2 penalty. We conducted a systematic 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).

Table 15: 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

We report results at $\alpha = 1.0$ throughout the main text. Table 15 demonstrates that while the magnitude of EI varies with regularization strength, the qualitative finding—that cross-concept damage exceeds within-concept damage—persists across three orders of magnitude of α.

Regularization *suppresses* measured entanglement, not amplifies it. OLS produces EI = 1.555—the highest value—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: 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.

We replicated on Qwen 2.5-7B (layer 27, $d = 3{,}584$). At 7B scale, OLS is numerically ill-conditioned—the first singular value exceeds the second by 3×10^6, and the leading direction captures the mean signal rather than concept-specific variation. Within the well-conditioned range ($\alpha \in [0.1, 10]$), EI ranges from 0.898 to 1.310, with V-matrix purity stable at 0.82–0.83.

The reported EI values at $\alpha = 1.0$ are lower bounds at GPT-2 scale and are in the center of the well-conditioned range at 7B scale.

6.5 Limitations

Rank-dimensionality confound. The weight matrix W has rank 7, producing exactly 7 non-trivial SVD directions. Removing 1 of 7 directions necessarily reduces representational capacity. The question is whether the *universal* pattern (all concepts damaged by any removal) reflects genuine co-encoding or is a mathematical consequence of the rank-7 constraint. The random projection experiment provides evidence for the former: entanglement reproduces in random subspaces of dimension $\geq$448 (rank $\gg$ 7). The pairwise experiments 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. We emphasize that the damage matrix measures a classifier-relative quantity, not a direct activation-space property, and this distinction should inform interpretation. A direction critical for maintaining classification geometry may carry information about the damaged concept only indirectly—removing it may disrupt the classifier's coordinate system rather than erase the concept's content from the representation.

Coverage. All experiments cover four models across two architecture families (GPT-2 and Qwen); whether the phenomenon replicates in non-autoregressive architectures (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.

Probe design. The 160-probe factorial design uses generated text rather than naturalistic samples. The 97% fidelity of the half-factorial design (96 probes, Table 16) and the 83% fidelity of the eighth-factorial (32 probes) suggest the structure is robust to probe-level variation.

Table 16: 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

7 Conclusion

When you remove a concept direction from a transformer's activation space, other concepts break. This paper reports the discovery, formalization, and cross-architecture replication of this phenomenon.

The discovery: in Qwen 2.5-7B, a single register direction causes a 52.5% drop in domain classification accuracy, despite lying 82% outside the 36-dimensional domain subspace that INLP found. The entanglement is asymmetric (register removal destroys domain; domain removal leaves register untouched) and learned (register directions from domain-neutral prompts carry zero domain information).

The formalization: factorial direction decomposition separates discrimination geometry (how classifiers use directions) from activation geometry (what directions carry). The V-matrix shows clean concept separation. The damage matrix shows universal coupling. This discrimination–activation dissociation is the central finding.

The replication: eight experiments across GPT-2, Qwen-0.5B, Qwen-7B, and Qwen-7B-Instruct establish that entanglement is universal (EI > 1.0 in all terminal layers), follows architecture-specific crystallization dynamics, scales superlinearly with concept count ($2\times$ amplification from pairwise to triple), and is concept-type independent (replicates with software engineering concepts).

PCA disentanglement to 28 dimensions reduces EI below 0.1, and companion work [McEntire, 2026i] shows that fine-tuning at sufficient scale drives EI to zero entirely. Together, these results demonstrate that entanglement is a property of the model's chosen representational geometry—its compression strategy—not an inescapable feature of the activations themselves. Base models in their pre-trained state use high-dimensional distributed representations that produce entanglement; models with sufficient capacity and appropriate training signal can expand the informative subspace to disentangle. Methods operating in the model's native basis will encounter entanglement in base models. Methods that concentrate concept variance (e.g., PCA) or that fine-tune with sufficient capacity can reduce or eliminate it.

The practical consequence: any interpretability method claiming to isolate a concept should demonstrate isolation in the damage matrix, not just in classifier weights. Multi-concept factorial decomposition provides a direct way to make this measurement.

The theoretical consequence: single-concept search is structurally incomplete. The most informative directions in activation space are entangled across concepts, and methods that search one concept at a time will systematically miss them. Multi-concept joint probing is necessary to map the full structure of transformer activation space.

Future work should test whether the discrimination–activation dissociation extends to

sparse autoencoder features, whether entanglement is present in encoder-only and multimodal architectures, and whether the crystallization phase transition can be predicted from architectural properties rather than measured empirically. Both the PCA disentanglement result and the fine-tuning results in McEntire [2026i] confirm that concept-pure representations are achievable—understanding when and why models choose distributed entangled representations over concentrated disentangled ones is a question about training dynamics and compression, not about the limits of representation.

Code Availability

Experiment code and probe sets are available at https://github.com/jmcentire/universal-entanglemer

References

G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*, 2017.

N. Belrose, Z. Furman, L. Smith, D. Halawi, I. Ostrovsky, L. McKinney, S. Biderman, and J. Steinhardt. Eliciting latent predictions from transformers with the tuned lens. In *International Conference on Learning Representations*, 2023.

N. Belrose, D. Schneider-Joseph, S. Ravfogel, R. Cotterell, E. Raff, and S. Biderman. LEACE: Perfect linear concept erasure in closed form. In *NeurIPS*, 2023.

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

C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

A. Conmy, A. N. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *NeurIPS*, 2023.

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

N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, R. Grosse, S. McCandlish, J. Kaplan, D. Amodei, M. Wattenberg, and C. Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

A. Geiger, C. Potts, and T. Icard. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning (CLeaR)*, 2024.

W. Gurnee and M. Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, 2024.

S. Marks, M. Rager, E. J. Michaud, Y. Belinkov, D. Bau, and A. Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *ICML*, 2024.

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

J. McEntire. Subspace decomposition and contrastive refinement in transformer activation space. Working paper, 2026.

J. McEntire. Weak probing under capacity pressure: Experiments in domain-shape separation. Working paper, 2026.

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.

K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, 2022.

A. Mueller, L. Sharkey, R. Gallagher, and N. Nanda. Sparse autoencoders find composed features in small toy models. *arXiv preprint arXiv:2501.xxxxx*, 2025.

N. Nanda, A. Lee, and M. Berber. Emergent linear representations in world models of self-supervised sequence models. In *BlackboxNLP Workshop at EMNLP*, 2023.

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*, pages 7237–7256, 2020.

S. Ravfogel, M. Twiton, Y. Goldberg, and R. D. Cotterell. Linear adversarial concept erasure. In *ICML*, 2022.

J. McEntire. Entanglement under fine-tuning: Scale-dependent disentanglement, crosstalk-guided companion selection, and the creative writing falsification. Zenodo, 2026.

J. McEntire. The entanglement theorem: Why concept separability fails in high-dimensional activation spaces. Zenodo, 2026.