Entangled Directions in Transformer Activation Space:

INLP Blind Spots, Asymmetric Feature Coupling, and Partial Extraction Efficiency

Jeremy McEntire*

Abstract

Iterative Null-Space Projection (INLP) is the standard tool for identifying and removing concept-specific directions in transformer activation space. We show that INLP has a systematic blind spot: it misses directions that are entangled across multiple concepts. In Qwen 2.5-7B, a single register direction (formal vs. informal) lies 82–88% outside the 36-dimensional domain INLP subspace yet causes a 52.5% drop in domain classification accuracy when removed—while random directions of the same dimensionality cause zero damage. This direction is the most domain-informative direction in the space, and 36 iterations of domain-focused INLP completely missed it.

The entanglement is asymmetric: removing the register direction destroys domain information, but removing 36 domain directions causes zero register damage. It is also learned rather than architectural: register directions computed from domain-neutral prompts carry no domain information and have low cosine similarity (≤ 0.28) with the entangled direction.

We introduce a dual-metric sigma sweep that reveals structure hidden by conditional accuracy. Using a partial projection operator $P'(\sigma) = I - \sigma QQ^\top$ with $\sigma \in [0, 1]$, we show that partial extraction at $\sigma^* < 1$ preserves significantly more joint information than full extraction, with gains of $+0.525$ to $+0.700$ over full removal. The Pareto frontier between feature extraction and preservation is convex at layers 14–27, indicating efficient partial extraction, and L-shaped at layer 7, indicating threshold behavior. The geometry of feature coupling changes across depth, with the transition occurring between layers 7 and 14.

*Working Paper. Correspondence: jmc@cageandmirror.com

These results demonstrate that transformer representations encode features in asymmetrically entangled subspaces whose structure depends on training co-occurrence rather than architecture. Standard interpretability tools that search within single-concept variance systematically miss these cross-concept directions. We propose multi-concept INLP as a direct fix.

1 Introduction

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]. Iterative Null-Space Projection [Ravfogel et al., 2020] operationalizes this: given labeled data for a target concept, INLP iteratively finds discriminative hyperplanes and projects them out, producing a set of directions that capture the concept's linear footprint in activation space. The assumption is that these directions constitute the concept's representation—that single-concept search is sufficient to recover single-concept structure.

This assumption is wrong when concepts share representational resources. Elhage et al. [2022] showed that models encode more features than they have dimensions by superposing features in overlapping directions. Our finding is related but distinct: the entangled direction we identify is not a superposition artifact resolved by sparse coding. It is a direction that *jointly* encodes two concepts because training statistics made joint encoding efficient.

We study domain identity (medical, legal, code, science) and linguistic register (formal vs. informal) in Qwen 2.5-7B. Domain INLP produces 36 directions that classify domain at $> 97\%$ accuracy. Register INLP produces a single direction that classifies register at 100% accuracy. The two direction sets are nearly orthogonal—the register direction projects only 12–18% into the domain subspace.

Yet removing the register direction drops domain classification from 100% to 47.5%. Removing the 36 domain directions drops register classification by exactly 0%.

The register direction found something that domain INLP could not: a direction that carries both register and domain information, living almost entirely outside the domain subspace. This is not a property of how we labeled the data. The probes were designed with balanced domain×register factorial structure, and random directions of the same dimensionality cause zero domain damage ($z > 5000$). The direction is specific, not a dimensionality artifact.

We call this phenomenon *asymmetric feature entanglement*: removing a low-dimensional feature (register, 1 direction) destroys a high-dimensional feature (domain, 36 directions), but not vice versa. The entanglement is directional. The register direction sits at a hub in

the model's representational geometry, encoding both features in a single dimension. The domain directions span a subspace that does not reach this hub.

1.1 Contributions

1. **INLP blind spot.** 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 (Section 3.1).

2. **Partial extraction efficiency.** We introduce a dual-metric sigma sweep that separates raw and conditional accuracy, revealing conjugate structure hidden by standard evaluation. Partial removal at $\sigma^* < 1$ achieves gains of $+0.525$ to $+0.700$ in joint information over full removal (Section 3.2).

3. **Developmental geometry.** The Pareto frontier between feature extraction and preservation transitions from L-shaped (layer 7) to convex (layers 14–27), indicating that the representational hierarchy crystallizes between layers 7 and 14 (Section 3.3).

4. **Learned entanglement.** Register directions computed from domain-neutral prompts carry zero domain information and have low alignment ($\cos \leq 0.28$) with the entangled direction, showing that entanglement emerges from co-occurrence in training data rather than transformer architecture (Section 3.4).

2 Setup

2.1 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, 2025a]: 16 formal and 16 informal, all domain-neutral (institutional/conversational language without domain-specific content).

2.2 Direction Sets

Domain INLP directions. We use 36 pre-computed domain INLP directions from prior work, orthonormalized via QR decomposition to produce $Q_{\text{domain}} \in \mathbb{R}^{3584\times 36}$. These directions classify domain at $> 97\%$ LOO 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 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 [2025a], 8 register directions computed from the 32 domain-neutral prompts via logistic regression and SVD, orthonormalized to $Q_{\text{p78}} \in \mathbb{R}^{3584\times 8}$.

2.3 Partial Projection Operator

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 $x_{\text{ext}} = \sigma\, QQ^\top x$.

2.4 Dual-Metric Evaluation

Prior work [McEntire, 2025b] 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.

We introduce the raw metric:

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

which classifies the measured feature without conditioning. This surfaces the cross-concept damage that conditional evaluation hides.

3 Results

3.1 The INLP Blind Spot

Table 1 shows the asymmetric entanglement at each layer. 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, ruling out dimensionality artifacts.

Where does the register direction live? 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).

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.

3.2 Partial Extraction Efficiency

Table 3 shows the dual-metric sigma sweep for register removal. The raw metric reveals a massive gain from partial extraction.

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

At every layer, the raw metric yields Outcome A: $\text{joint}_{\text{raw}}(\sigma^*) > \text{joint}_{\text{raw}}(\sigma{=}1)$ with gains exceeding the confidence interval and the random control. The conditional metric yields Outcome B at every layer—the entanglement signal is invisible to standard evaluation.

The interpretation: at σ^*, the projection extracts enough of the register direction to achieve perfect register classification ($\text{acc}_{\text{register}}(x_{\text{ext}}) = 1.0$) while the residual retains nearly all domain information ($\text{acc}_{\text{domain}}(x_{\text{res}}) \approx 0.99$). Full removal ($\sigma = 1$) extracts the same register information but destroys domain in the process.

3.3 Pareto Frontier Geometry

Plotting domain preservation vs. register extraction across σ reveals the geometry of the feature tradeoff (Table 4).

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. This is the regime where partial extraction is genuinely more efficient than full extraction.

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 (curve bows above the diagonal, partial extraction is efficient); negative values indicate concavity or threshold behavior (curve bows below).

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 learned to distribute 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. [2023] and in our own prior work [McEntire, 2025a], where layers 7–10 show maximal per-neuron feature selectivity before representations reorganize into distributed codes at deeper layers. The L-shaped frontier at layer 7 may reflect a regime where individual neurons still "own" features, making entanglement all-or-nothing; by layer 14, the distributed encoding enables smooth partial separation.

3.4 Learned vs. Architectural Entanglement

Register directions from domain-neutral prompts (Phase 7_8, 8 dimensions) produce a fundamentally different subspace than register directions learned from domain×register probes (v4, 1 dimension per layer).

Table 5: Phase 7_8 register directions vs. entangled register direction.

Layer	Domain drop (p78)	Domain drop (v4)	$\cos_{\max}$	# dims (p78 / v4)
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 build a joint representation that entangles 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 Discussion

4.1 Implications for Interpretability

INLP and related linear probing methods search for concept-specific directions by iteratively fitting classifiers to a single set of labels. This procedure is blind to directions that are informative about the target concept *through* their entanglement with other concepts. The register direction carries more domain-destructive power than any single INLP direction, yet it was invisible to 36 iterations of domain-focused search because its domain information is accessible only through its register structure.

The fix is straightforward in principle: multi-concept INLP that searches for directions jointly informative about multiple label sets. In practice, this requires labeled data for multiple concepts simultaneously, which our domain×register factorial probes provide. The entangled direction would appear in a joint INLP that optimizes for both domain and register classification, rather than searching within each concept independently.

A concrete implementation would proceed as follows. Given n samples with label vectors $(y_{\text{domain}}, y_{\text{register}})$, fit a multi-output classifier $f : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{Y}_{\text{domain}}|+|\mathcal{Y}_{\text{register}}|}$ that predicts both labels simultaneously. The first INLP iteration extracts the most jointly discriminative direction—which, based on our results, should approximate the entangled register direction, since it is the single direction carrying the most total label information across both concepts. Subsequent iterations would capture decreasingly entangled directions, producing a naturally ordered decomposition from entangled to concept-specific. The standard single-concept INLP is a special case where only one label set is provided.

4.2 The Role of Conditional Evaluation

The conditional metric ($\text{acc}_{\text{measure}|\text{key}}$) evaluates the measured feature within groups defined by an external label. This is appropriate when the goal is to assess whether the feature

is recoverable *given* external knowledge. But it is blind to between-group structure: after register removal, domain accuracy within each register group remains perfect, while overall domain accuracy collapses.

This is not a failure of the conditional metric—it measures what it claims to measure. But it hides entanglement-related damage, and any analysis relying exclusively on conditional evaluation will miss the central phenomenon reported here.

4.3 Connection to Structural Compression

The Phase 7_8 finding anchors the theoretical interpretation. The model does not entangle register and domain because transformers inevitably mix features. It entangles them because training data co-occurrence makes entanglement an efficient compression strategy. When register is observed without domain context, the model learns a disentangled register subspace. When register and domain co-occur, the model builds a joint representation.

This is the representational manifestation of lossy channel coding under selection pressure. The model is a compression function mapping input sequences to fixed-dimensional activation vectors. Under training loss as the selection criterion, directions that are jointly informative about multiple co-occurring features are retained, while directions informative about a single feature in isolation are less efficiently encoded. The register direction is a hub in this compressed representation—a single dimension that serves double duty.

4.4 Limitations

Our analysis is limited to a single model (Qwen 2.5-7B) and two feature pairs (domain×register). The 80-probe design, while balanced, uses generated text rather than naturalistic samples. The Pareto convexity metric is sensitive to discretization ($\Delta\sigma = 0.1$) and finer sweeps may reveal more structure.

Register is binary (formal/informal), and INLP consistently finds exactly one direction—which is the minimum a binary classifier requires. This is not a limitation but a prediction: a richer register taxonomy (e.g., formal/conversational/technical/literary) would require multiple INLP directions, and if co-occurrence patterns hold, each could carry its own entanglement with domain. Binary register produces the simplest possible entangled structure; richer sociolinguistic features should produce proportionally richer entanglement geometries. Testing this prediction is immediate future work.

5 Conclusion

Standard interpretability methods assume features occupy separable subspaces. We showed that the most domain-informative direction in Qwen 2.5-7B's activation space is a register direction that INLP cannot find, because its domain information is entangled with register structure. The entanglement is asymmetric, learned from training co-occurrence, and geometrically characterized by convex Pareto frontiers at layers 14–27.

Partial extraction at $\sigma^* < 1$ preserves significantly more joint information than full extraction, demonstrating that the representational coupling admits efficient partial measurement. This structure emerges between layers 7 and 14 and persists through the terminal layer, suggesting it is a stable property of the model's learned representation rather than a transient intermediate computation.

The immediate practical consequence is that INLP and similar single-concept probing methods have a systematic blind spot for entangled directions. Multi-concept joint probing is necessary to map the full structure of transformer activation space.

References

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, 2020.

Jeremy McEntire. Subspace decomposition and contrastive refinement in transformer activation space. Working paper, 2025.

Jeremy McEntire. Weak probing under capacity pressure: Experiments in domain-shape separation. Working paper, 2025.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney,

Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. In *International Conference on Learning Representations*, 2023.

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