The Concentration Barrier:
Effective Dimensionality Bounds Domain Selectivity in Neural Network Activations

Jeremy McEntire[1]

March 2026

Abstract

Paper VIII in this series showed that shaped noise injection at terminal transformer layers cannot achieve domain-selective entropy effects. Paper IX mapped the response tensor at each layer. This paper establishes the mathematical explanation: the *concentration barrier*. We prove that in an activation space with effective dimensionality d_{eff} (measured by participation ratio), the maximum achievable selectivity from k domain directions is bounded by k/d_{eff}. We measure d_{eff} at every transformer layer of Qwen-2.5 7B under both last-token and mean-pooled extraction, finding $d_{\text{eff}}^{\text{LT}} \in [4.7, 26.0]$ (mean 19.2) while $d_{\text{eff}}^{\text{MP}}$ collapses to 1.0 at layers 3–25 before recovering at layer 27. The INLP variance fraction increases from 1.3% at early layers to 12.5% at the terminal layer, yet the concentration barrier bound $\text{INLP}_{\text{var}} \leq k/d_{\text{eff}}$ holds empirically at all 28 layers. The pooling-dependent collapse reveals a previously undocumented form of representation anisotropy that is invisible to position-aware measurements.

1 Introduction

Paper VIII's terminal measurement limit and Paper IX's layer-resolved selectivity curve both demand a *theoretical* explanation. Why can't linear interventions in high-dimensional activation space achieve domain selectivity? The answer lies in the concentration of measure phenomenon: in high dimensions, almost all directions are nearly orthogonal to any fixed subspace.

We formalize this as the *concentration barrier*:

Theorem 1 (Concentration Barrier). *Let $h \in \mathbb{R}^d$ be the hidden state at a transformer layer with covariance Σ having eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. Define the effective dimensionality as the participation ratio:*

$$d_{\mathit{eff}} = \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2} \tag{1}$$

[1] Correspondence: `jmc@cageandmirror.com`

For any set of k unit directions $\{v_1, \ldots, v_k\}$, the fraction of total variance captured by these directions satisfies:

$$\frac{\sum_{j=1}^{k} v_j^\top \Sigma v_j}{\mathrm{tr}(\Sigma)} \leq \frac{k}{d_{\mathit{eff}}} \tag{2}$$

with equality when $d_{\mathit{eff}} = k$ (i.e., all variance lies in the k directions).

The bound says: if the activation space has high effective dimensionality, then any k directions capture a vanishingly small fraction of the variance. Since INLP produces $k = 9$ directions per domain (36 total) in a space with $d = 3584$ dimensions, the fraction is small unless $d_{\text{eff}} \approx k$.

1.1 Implications for Selectivity

The bound constrains how much of the activation variance any fixed set of directions can capture. The implication for intervention is indirect but important: if perturbation along k domain directions affects only a fraction k/d_{eff} of the total activation variance, then the perturbation's domain-specific leverage is correspondingly limited. The formal bound is on variance fraction; the connection to behavioral selectivity (the entropy change ratios measured in Papers VIII and IX) is not derived here but is empirically consistent — Papers VIII and IX observe low selectivity at all layers, as the variance fraction bound would predict.

2 Method

2.1 Models and Data

- **Primary**: Qwen-2.5 7B (28 layers, $d = 3584$)
- **Probes**: 160 domain probes (40 per domain, 4 structural shapes)
- **INLP directions**: 36 (9 per domain) from Paper IV, unit-normalized

2.2 Activation Capture

We register forward hooks on all 28 transformer layers. For each of 160 probes, we perform one forward pass and capture the hidden state at every layer under *two* pooling strategies:

Last-token pooling:

$$a_{\text{LT}}^{(\ell)} = h_T^{(\ell)} \tag{3}$$

where T is the final token position. This matches Paper VIII's noise injection mechanism and Paper IV's INLP direction computation.

Mean pooling:

$$a_{\text{MP}}^{(\ell)} = \frac{1}{T} \sum_{t=1}^{T} h_t^{(\ell)} \tag{4}$$

This captures the global representation structure, averaging over all positions.

Each pooling yields a $(160, 3584)$ activation matrix at each layer.

2.3 Per-Layer Metrics

For each layer ℓ and each pooling method:

1. **Participation ratio**: $d_{\text{eff}}^{(\ell)}$ from the SVD of centered activations.
2. **INLP variance fraction**: $\|AP_{\text{INLP}}\|_F^2 / \|A\|_F^2$, measuring what fraction of activation variance lies in the INLP subspace.
3. **Fisher separation**: $F = \|\mu_A - \mu_B\|^2 / (\text{tr}(C_A) + \text{tr}(C_B))$ averaged over all domain pairs.
4. **Bound test**: Whether $\text{INLP}_{\text{var}} \leq k/d_{\text{eff}}$ holds at each layer.

3 Results

3.1 Effective Dimensionality Profile

Table 1 shows the effective dimensionality at each of 28 transformer layers under both pooling methods.

The two pooling methods reveal strikingly different dimensionality profiles.

Last-token d_{eff} ranges from 4.7 (layer 1) to 26.0 (layer 26), with a mean of 19.2. It follows a characteristic arc: a sharp dip at layer 1, recovery through layers 2–7, a broad plateau around 20–22 at layers 8–14, slight decline through layers 15–20, and a secondary rise at layers 22–26 before settling at 22.3 at the terminal layer.

Mean-pooled d_{eff} shows a radically different pattern: high values ($\sim$40) at layers 0–2, then complete collapse to 1.0 at layers 3–25, before recovering to 18.9 and 22.5 at layers 26–27. The collapse indicates that position-averaged activations are dominated by a single principal component through 23 of 28 layers.

3.2 The Anisotropy Collapse

The mean-pooled $d_{\text{eff}} = 1.0$ at layers 3–25 deserves explanation. A participation ratio of 1.0 means the eigenspectrum is dominated by a single eigenvalue — the centered activation vectors

Table 1: Effective dimensionality (d_{eff}), INLP variance fraction, Fisher separation, and theoretical bound at each layer of Qwen-2.5 7B. LT = last-token, MP = mean-pooled.

Layer	$d_{\text{eff}}^{\text{LT}}$	$d_{\text{eff}}^{\text{MP}}$	INLP$^{\text{LT}}$%	INLP$^{\text{MP}}$%	Fish.$^{\text{LT}}$	Fish.$^{\text{MP}}$	Bound
0	16.5	43.9	1.28	1.61	0.164	0.137	2.18
1	4.7	39.4	1.37	1.54	0.104	0.142	7.63
2	12.3	40.5	1.94	1.87	0.215	0.171	2.93
3	17.7	1.0	1.58	2.75	0.149	0.060	2.03
4	17.1	1.0	1.44	2.73	0.132	0.060	2.11
5	15.8	1.0	1.37	2.73	0.139	0.061	2.28
6	15.2	1.0	1.53	2.73	0.159	0.061	2.36
7	18.8	1.0	1.54	2.73	0.159	0.060	1.92
8	22.0	1.0	1.74	2.73	0.210	0.060	1.64
9	22.5	1.0	1.84	2.73	0.234	0.062	1.60
10	21.7	1.0	2.00	2.74	0.240	0.063	1.66
11	21.7	1.0	2.04	2.73	0.228	0.062	1.66
12	20.8	1.0	1.98	2.72	0.220	0.062	1.73
13	21.5	1.0	1.92	2.71	0.199	0.063	1.67
14	21.0	1.0	2.00	2.72	0.198	0.063	1.71
15	20.5	1.0	2.06	2.72	0.181	0.063	1.75
16	19.8	1.0	2.02	2.74	0.174	0.063	1.82
17	19.5	1.0	2.08	2.74	0.168	0.063	1.85
18	18.6	1.1	2.08	2.75	0.151	0.062	1.94
19	17.9	1.1	2.84	2.83	0.188	0.067	2.01
20	17.7	1.1	2.79	2.91	0.186	0.069	2.04
21	17.9	1.2	2.92	3.07	0.180	0.076	2.01
22	19.3	1.3	3.61	3.46	0.213	0.088	1.86
23	21.4	1.4	4.17	4.02	0.220	0.101	1.68
24	22.9	1.6	4.94	4.89	0.244	0.117	1.57
25	23.6	1.9	5.19	5.86	0.230	0.126	1.53
26	26.0	18.9	6.12	14.18	0.242	0.250	1.38
27	22.3	22.5	12.46	28.38	0.312	0.406	1.61

all lie along essentially one direction. This is a manifestation of the *representation anisotropy* phenomenon: after the embedding layers, the mean-pooled representation concentrates onto a narrow cone, with all 160 probes producing nearly identical mean-pooled activations up to a rank-1 perturbation.

Critically, this collapse is *invisible to position-aware measurements*. The last-token d_{eff} at the same layers ranges from 15 to 22, indicating rich multi-dimensional structure in the position-specific activations that mean-pooling destroys. Domain classification from mean-pooled activations achieves 90–100% accuracy (from v1 measurements) even at $d_{\text{eff}} = 1.0$, because the classifier exploits structure in the small residual variance that the participation ratio does not weight.

At layers 26–27, the two pooling methods converge ($d_{\text{eff}}^{\text{LT}} = 22.3$, $d_{\text{eff}}^{\text{MP}} = 22.5$ at layer 27), suggesting the representation "unfolds" near the output layer, distributing variance across dimensions in preparation for the vocabulary projection.

3.3 INLP Variance Fraction

The INLP variance fraction (proportion of activation variance captured by the 36 INLP directions) increases monotonically toward terminal layers under both pooling methods:

- **Last-token**: 1.3% at layer 0 → 12.5% at layer 27 (9.7× increase)
- **Mean-pooled**: 1.6% at layer 0 → 28.4% at layer 27 (17.8× increase)

This gradient has a clear interpretation: INLP directions become progressively more aligned with the activation space as representations are transformed through the forward pass. The directions were learned from terminal-layer activations (Paper IV), so their alignment with earlier layers is incidental.

3.4 Fisher Separation

The Fisher discriminant ratio (mean across all six domain pairs) follows a non-monotonic profile under last-token extraction. It peaks at layers 8–10 (Fisher ≈ 0.24), dips through layers 15–21, and rises to its maximum at layer 27 (Fisher $= 0.31$). This profile suggests two regimes of domain separability: a mid-layer regime where abstract features differentiate domains, and a terminal regime where task-specific representations diverge.

Under mean-pooling, Fisher separation is suppressed to $\sim$0.06 at layers 3–18 (consistent with $d_{\text{eff}} = 1.0$) before rising at layers 26–27.

4 The Bound

4.1 Proof of Theorem 1

Proof. Let $\Sigma = U \Lambda U^\top$ be the eigendecomposition with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$. For any unit vector v:

$$v^\top \Sigma v = \sum_i \lambda_i (u_i^\top v)^2 \tag{5}$$

By the Cauchy–Schwarz inequality applied to the weighted inner product:

$$v^\top \Sigma v \leq \lambda_{\max} \sum_i (u_i^\top v)^2 = \lambda_{\max} \tag{6}$$

For k directions:

$$\sum_{j=1}^{k} v_j^\top \Sigma v_j \leq k \cdot \lambda_{\max} \tag{7}$$

Meanwhile:

$$\operatorname{tr}(\Sigma) = \sum_i \lambda_i, \quad d_{\text{eff}} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2} \leq \frac{(\sum_i \lambda_i)^2}{\lambda_{\max} \sum_i \lambda_i} = \frac{\sum_i \lambda_i}{\lambda_{\max}} \tag{8}$$

Therefore $\lambda_{\max} \leq \operatorname{tr}(\Sigma)/d_{\text{eff}}$, giving:

$$\frac{\sum_{j=1}^{k} v_j^\top \Sigma v_j}{\operatorname{tr}(\Sigma)} \leq \frac{k \cdot \lambda_{\max}}{\operatorname{tr}(\Sigma)} \leq \frac{k}{d_{\text{eff}}} \tag{9}$$

□

4.2 Empirical Test of the Bound

The concentration barrier bound $\text{INLP}_{\text{var}} \leq k/d_{\text{eff}}$ holds at all 28 layers under last-token extraction. The bound column in Table 1 shows $k/d_{\text{eff}}^{\text{LT}}$, which ranges from 1.38 (layer 26) to 7.63 (layer 1). The actual INLP variance fraction ranges from 1.28% to 12.46%, always well below the bound.

Tightness. The bound is loose by a factor of 11–557× at layers 0–18, where $d_{\text{eff}}^{\text{LT}} \approx 20$ gives a bound of ∼1.7 but the actual INLP variance is ∼1.5–2%. At the terminal layer, the bound tightens: $k/d_{\text{eff}} = 1.61$ vs. actual 12.46%, a gap of ∼13×. The bound is a necessary condition, not a sufficient characterization.

Why the bound is loose. The theorem bounds the variance captured by *any* k directions. INLP directions are not arbitrary — they are computed to be maximally discriminative between domains, which biases them toward the subspace of *domain-relevant* variance. In an activation space where domain-relevant variance is a small fraction of total variance, the INLP directions will capture far less than k/d_{eff} of the total variance even if they capture most of the domain-relevant component.

5 Discussion

5.1 Why the Terminal Measurement Limit Exists

Paper VIII established that shaped noise injection at terminal layers cannot achieve domain-selective entropy effects, even when the response matrix $\mathbf{R}$ is invertible. The concentration barrier provides one component of the explanation: at the terminal layer ($d_{\text{eff}} = 22.3$), any 36 directions capture at most $36/22.3 = 1.61$ of the total variance. The INLP directions actually capture 12.46% — more than at any other layer — but this is still domain-entangled variance (Paper VIII's cross-domain bleed), not domain-selective variance.

The bound alone does not explain the terminal measurement limit; the limit arises from a combination of (a) the concentration barrier limiting how much variance any fixed directions can capture, and (b) the entanglement of domain-specific and cross-domain variance within the INLP subspace at terminal layers. The concentration barrier is necessary but not sufficient.

5.2 The Pooling Duality

The most unexpected finding is the dramatic divergence between last-token and mean-pooled d_{eff}. At layers 3–25, the last-token representation maintains $d_{\text{eff}} \approx 20$ while the mean-pooled representation collapses to $d_{\text{eff}} = 1.0$. This means:

1. The position-specific representations (what each token "sees" at each layer) are multi-dimensional.
2. The position-averaged representation is dominated by a single direction — the "global activation mode."
3. Mean-pooling destroys the multi-dimensional structure by averaging out position-dependent features.

This has implications for representation probing methodologies. Mean-pooled representations are commonly used in probing classifiers, yet they discard the very dimensional structure that determines intervention effectiveness. The concentration barrier theorem gives different (and vacuous, at $d_{\text{eff}} = 1$) bounds depending on which pooling is used.

5.3 The INLP Gradient

The monotonic increase of INLP variance fraction from 1.3% to 12.5% across layers is consistent with a model in which domain-discriminative structure is *constructed* through the forward pass. Early layers encode token-level features with little domain specificity; later layers compose these into domain-relevant representations. The INLP directions, learned at the terminal layer, are progressively more aligned with the activation subspace as we approach their layer of origin.

This gradient also explains why early-layer injection might fail (Paper IX Outcome C): perturbation along INLP directions at layer 3 affects only 1.6% of the activation variance, and subsequent layers may attenuate this small perturbation further.

5.4 Implications for Papers X and XII

The per-layer d_{eff} and INLP variance profiles provide the denominators for Paper XII's channel capacity bound: $C \leq \log_2(1 + \Delta^2/d_{\text{eff}})$, where Δ is the signal-to-noise ratio of the domain perturbation. Paper X's spectral geometry analysis can examine whether INLP directions align with the amplified or attenuated modes of the layer Jacobian, connecting the static variance structure measured here to the dynamic propagation structure of the forward pass.

6 Conclusion

The concentration barrier theorem bounds the fraction of activation variance capturable by any k fixed directions: INLP variance fraction $\leq k/d_{\text{eff}}$. We verify this bound empirically at all 28 layers of Qwen-2.5 7B. The connection to behavioral selectivity (entropy change ratios) is indirect: the variance fraction limits the domain-specific leverage of shaped noise, consistent with the low selectivity observed in Papers VIII–IX, though the formal mapping remains an open derivation.

The key findings are:

1. Last-token d_{eff} ranges from 4.7 to 26.0 (mean 19.2), giving a bound of $\sim$1.4–7.6 on the variance fraction capturable by 36 INLP directions. The actual fraction is 1.3–12.5%.

2. Mean-pooled d_{eff} collapses to 1.0 at layers 3–25, a form of representation anisotropy invisible to position-aware measurements.

3. INLP variance fraction increases monotonically toward terminal layers, indicating the forward pass progressively constructs domain-discriminative structure.

4. The bound holds at all layers but is loose (11–557×), indicating that the variance structure is more favorable to the INLP directions than a worst-case analysis predicts, yet the actual variance captured remains far below the theoretical maximum.

The concentration barrier is a necessary condition for Paper VIII's terminal measurement limit but not a sufficient explanation. The additional ingredient — cross-domain entanglement within the INLP subspace — is measured by Paper IX's layer-resolved response tensor and will be formalized in Papers X and XII.

References

[1] McEntire, J. (2026). Paper I: Leap+Verify. *arXiv:2602.19580.*

[2] McEntire, J. (2026). Paper II: Ensemble Collapse. *SSRN*.

[3] McEntire, J. (2026). Paper III: Constellation Composition.

[4] McEntire, J. (2026). Paper IV: Structural Transfer.

[5] McEntire, J. (2026). Paper V: Capability Manifold Surveillance.

[6] McEntire, J. (2026). Paper VI: Communicative Variance.

[7] McEntire, J. (2026). Paper VII: GenAI Is Socially Awkward.

[8] McEntire, J. (2026). Paper VIII: Shaped Noise Injection.

[9] McEntire, J. (2026). Paper IX: Layer-Resolved Response Tensor. [This series]