Spectral Geometry of the Forward Pass: How INLP Directions Interact with Layer Jacobians

Jeremy McEntire[1]

March 2026

Abstract

Paper IX showed that domain selectivity from shaped noise injection peaks at intermediate layers (7–10) and declines at both input and terminal layers. Paper XI showed that effective dimensionality ($d_{\text{eff}} \approx 20$) bounds achievable selectivity. This paper asks *why* selectivity peaks where it does by measuring the amplification spectrum of INLP domain directions through the layer Jacobian. Using finite-difference Jacobian-vector products (JVPs), we compute how perturbations along INLP directions propagate through consecutive layer pairs, and compare these amplification factors to random directions and to the PCA structure of activations at each layer.

Both hypotheses are rejected. INLP directions are **not** preferentially amplified relative to random directions: the INLP/random amplification ratio is 0.99 ± 0.05 across all injection layers. PCA-INLP alignment is near-random at intermediate layers (mean $|\cos\theta| = 0.015$–0.020, expected random ≈ 0.013) and increases toward terminal layers (mean 0.038 at layer 27), where selectivity is *lowest*. The forward pass treats INLP directions as generic directions in activation space, consistent with the concentration barrier (Paper XI) as the sole constraint on domain selectivity.

1 Introduction

Papers VIII–IX established that shaped noise injection at any single layer fails to achieve strong domain selectivity, with a modest peak at intermediate layers and a decline toward terminals. Paper XI proved the concentration barrier bounds this from above. What remains unexplained is the *mechanism*: why does selectivity peak at layers 7–10 rather than elsewhere?

Two hypotheses:

H1: Amplification. INLP directions are amplified relative to random directions at intermediate layers, giving shaped perturbations more leverage.

H2: Alignment. INLP directions align with the top principal components of activations at intermediate layers, meaning perturbations along these directions affect the bulk of the representation.

[1]Correspondence: jmc@cageandmirror.com

We test both via:

1. **JVP amplification spectrum**: perturb $h^{(\ell)}$ along INLP direction v, measure $\|h^{(\ell')}_{\text{perturbed}} - h^{(\ell')}_{\text{clean}}\|$ at downstream layers.
2. **PCA-INLP alignment**: compute cosine similarity between top PCA directions at each layer and the INLP direction set.

2 Method

2.1 Finite-Difference JVP

For injection layer ℓ and target layer ℓ', the amplification factor for direction v is:

$$A(\ell, \ell', v) = \frac{\|h^{(\ell')}_{\text{perturbed}} - h^{(\ell')}_{\text{clean}}\|}{\epsilon \cdot \|v\|} \tag{1}$$

where $h^{(\ell)}_{\text{perturbed}} = h^{(\ell)}_{\text{clean}} + \epsilon \cdot v$ with $\epsilon = 0.01$. We compute this for:

- 4 INLP mean directions (one per domain, averaged over 9 directions each)
- 8 random unit directions (baseline)
- 7 injection layers: $\ell \in \{0, 4, 9, 14, 18, 22, 27\}$
- All downstream sampled layers plus the terminal layer

Each amplification is averaged over 8 domain probes as input contexts.

2.2 PCA-INLP Alignment

At each sampled layer, we capture last-token activations across 160 domain probes, compute the top-20 PCA directions, and measure the mean absolute cosine similarity between these PCA directions and the 36 INLP directions (720 cosine values per layer).

2.3 Random Baseline

For $d = 3584$ dimensions and $k = 20$ PCA directions, the expected mean $|\cos\theta|$ between a random unit vector and the PCA subspace is $\sqrt{2/(\pi d)} \approx 0.013$. The expected maximum over 720 cosine pairs is approximately 0.075 by extreme value theory.

3 Results

3.1 JVP Amplification Spectrum

Table 1 presents the mean amplification factor from each injection layer to the terminal layer (27), for INLP directions vs. random directions.

Table 1: Amplification to terminal layer (27). $\bar{A}_{\text{INLP}}$: mean over 4 domain INLP directions. $\bar{A}_{\text{rand}}$: mean over 4 random directions. Ratio: INLP/random.

Inject ℓ	$\bar{A}_{\text{INLP}}$	$\bar{A}_{\text{rand}}$	Ratio
0	2094.9	2003.4	1.046
4	407.6	391.3	1.042
9	373.3	370.2	1.008
14	346.9	353.3	0.982
18	311.1	319.6	0.973
22	255.7	284.9	0.897
Grand mean ratio			0.991

Layer 27 omitted (injection at terminal → zero downstream propagation).

The INLP/random amplification ratio ranges from 0.897 to 1.046, with a grand mean of 0.991. INLP directions are amplified identically to random directions at every injection layer. H1 is rejected: there is no preferential amplification.

3.2 Amplification Structure

Several structural features of the amplification spectrum are notable regardless of direction type:

1. **Monotonic growth**: Amplification from injection to terminal increases monotonically with distance. From layer 0, amplification grows from ∼400 (at layer 4) to ∼2000 (at layer 27), roughly 1.06× per layer.

2. **Terminal spike**: The final five layers (22→27) contribute disproportionately. From layer 22 to 27, the amplification is ∼260, compared to ∼55 from 18 to 22 (4 layers).

3. **High variance in random directions**: Individual random directions span a 10× range (66–698 from layer 0 to layer 4), reflecting direction-dependent coupling to the Jacobian's singular structure. INLP directions show comparable variance across domains (313–586 from layer 0 to layer 4).

The forward pass amplifies all perturbations roughly equally. There is no "privileged" direction basis for the Jacobian.

3.3 Per-Domain Amplification

Table 2 breaks down INLP amplification to terminal by domain.

Table 2: Per-domain INLP amplification to terminal (layer 27) by injection layer.

ℓ	Medical	Legal	Code	Science	$\bar{A}_{\text{rand}}$
0	1854	1893	2938	1694	2003
4	374	432	436	388	391
9	361	411	371	350	370
14	358	338	356	335	353
18	310	303	315	315	320
22	265	257	242	259	285

At layer 0, the code INLP direction has notably higher amplification (2938) than other domains (∼1700–1900). This reflects the specific direction geometry rather than a systematic domain effect — individual random directions show comparable spread.

3.4 PCA-INLP Alignment

Table 3 presents the mean and maximum alignment between INLP directions and top-20 PCA directions at each sampled layer.

Table 3: PCA-INLP alignment at each layer. Mean: average $|\cos\theta|$ over 720 pairs. Max: maximum $|\cos\theta|$. Random baselines: mean ≈ 0.013, max ≈ 0.075.

Layer	Mean $\lvert\cos\theta\rvert$	Max $\lvert\cos\theta\rvert$	Interpretation
0	0.015	0.083	Near random
4	0.017	0.133	Slightly above random
9	0.018	0.098	Near random
14	0.020	0.089	Near random
18	0.020	0.103	Near random
22	0.023	0.264	Elevated max
27	0.038	0.469	Substantially above random

Mean alignment increases monotonically from 0.015 (layer 0) to 0.038 (layer 27), but remains low everywhere. At intermediate layers 9–14 (where Paper IX found peak selectivity), alignment is near the random baseline of 0.013.

The maximum alignment at layers 22 and 27 (0.264 and 0.469) is far above the random expectation of 0.075. Inspection of per-direction maxima reveals these outliers are the first three medical INLP directions (indices 0–2), which align with the top PCA components at terminal layers. The legal directions (indices 9–13) show moderate alignment (max 0.17–0.20). Code and science directions remain near random (max < 0.10).

3.5 Correlation with Paper IX Selectivity

Direct correlation between JVP amplification and Paper IX's selectivity is limited by a design mismatch: our injection layers $\{0, 4, 9, 14, 18, 22, 27\}$ overlap with Paper IX's $\{0, 3, 7, 10, 14, 17, 20, 24, 27\}$ at only three points (0, 14, 27), with layer 27 degenerate. Spearman correlation requires $n \geq 4$. The qualitative comparison below substitutes for the missing statistical test.

3.6 Alignment–Selectivity Dissociation

A critical finding: alignment and selectivity are *dissociated*. Alignment peaks at layer 27, where Paper IX reports mean selectivity $\bar{s} = -0.10$. Alignment is near-random at layers 9–10, where selectivity peaks at $\bar{s} = 0.57$. The medical INLP direction with the highest terminal alignment (0.469) corresponds to a domain that is *anti-selective* at most layers.

This dissociation rules out H2 as an explanation for the selectivity peak. Alignment with top PCA directions does not produce selectivity — it produces bleed, because the top PCA components carry domain-agnostic variance.

4 Discussion

Both hypotheses are rejected:

H1 (Amplification): INLP/random ratio = 0.991. INLP directions receive no preferential treatment from the layer Jacobians.

H2 (Alignment): PCA-INLP alignment peaks at terminal layers, not intermediate layers. Alignment correlates with *anti-selectivity*, not selectivity.

4.1 The Forward Pass as Isotropic Amplifier

The JVP results reveal a simple picture: the forward pass amplifies all perturbations roughly equally, regardless of their alignment with INLP or PCA directions. Each layer multiplies

perturbation magnitude by $\sim 1.06\times$ on average, with a terminal spike. The nonlinearity does not selectively filter INLP directions — it treats them as generic.

This is consistent with the concentration barrier. If the forward pass amplified INLP directions preferentially, one could design interventions that exploit the amplification differential. The near-unit ratio closes this avenue: there is no spectral shortcut around the barrier.

4.2 Connection to the Terminal Measurement Limit

Paper VIII's terminal measurement limit is often described as "nonlinear mixing scrambles INLP directions." The JVP data refines this: the scrambling is not direction-specific. The forward pass does not know which directions are INLP directions and which are random. It amplifies and mixes all equally.

The terminal spike (layers 22–27 contribute disproportionate amplification) explains why terminal injection produces large but unselective effects: perturbation energy is massively amplified at the final layers, overwhelming any directional specificity that might have existed earlier.

4.3 Why Alignment Increases at Terminal Layers

The increasing PCA-INLP alignment toward terminal layers has a natural explanation. Late-layer activations carry more domain-relevant information (the model is "deciding" its output). The INLP directions, computed to separate domains, increasingly align with the principal components that encode this domain information. But this alignment does not produce selectivity because the top PCA components at terminal layers carry *shared* variance that affects all domains, not domain-specific variance.

The medical INLP directions show the strongest terminal alignment (max 0.469) precisely because medical content overlaps most with the dominant activation modes — and precisely because medical is the most anti-selective domain in Paper IX. The directions that align with the bulk of the representation are the ones whose perturbations bleed most broadly.

4.4 Connection to the Concentration Barrier

Paper XI showed that selectivity is bounded by k/d_{eff}. The JVP results provide a complementary perspective: the bound holds not because of some active filtering mechanism but because of the *absence* of one. The forward pass is direction-agnostic. Without preferential amplification, the only leverage for selectivity comes from the geometric overlap between INLP directions and the activation subspace — exactly what k/d_{eff} measures.

If the forward pass preferentially amplified INLP directions (ratio $\gg 1$), the effective k in the bound would increase, potentially allowing higher selectivity. The near-unit ratio means the geometric bound from Paper XI is the complete story.

5 Conclusion

We measured the Jacobian-vector product amplification spectrum for INLP domain directions across seven injection layers in Qwen-2.5 7B. The central findings:

1. INLP directions are not preferentially amplified (INLP/random ratio = 0.991 ± 0.05).
2. PCA-INLP alignment is near-random at intermediate layers and elevated at terminal layers, dissociated from selectivity.
3. The forward pass operates as an isotropic amplifier, treating INLP directions identically to random directions.
4. The terminal amplification spike (layers 22–27) explains Paper VIII's large but unselective effects.

The selectivity peak at intermediate layers (Paper IX) cannot be attributed to preferential amplification or PCA alignment. The concentration barrier (Paper XI) is the sole constraint, and the JVP data explains why: without directional selectivity in the Jacobian, only the geometric overlap k/d_{eff} determines achievable selectivity.

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.

[10] McEntire, J. (2026). Paper XI: The Concentration Barrier.