Layer-Resolved Response Tensor: Where Domain Selectivity Lives in the Forward Pass

Jeremy McEntire[1]

March 2026

Abstract

Paper VIII established the *terminal measurement limit*: shaped noise injected at the final transformer layers cannot achieve domain-selective entropy effects because nonlinear mixing at intermediate layers scrambles the signal. This paper maps the full selectivity profile $\mathbf{R}^{(\ell)}$ at nine sampled transformer layers in Qwen-2.5 7B (28 layers total), measuring where in the forward pass domain selectivity peaks.

We find a **weak Outcome A**: mean selectivity peaks at intermediate layers 7–10 ($\bar{s} \approx 0.5$) and declines toward both input and output layers. However, absolute selectivity is modest at all layers — no layer achieves diagonal dominance in the 4×4 response matrix. We observe strong **domain asymmetry**: code and science domains consistently exhibit positive selectivity (self-effect exceeds bleed) while medical and legal domains exhibit negative selectivity (bleed exceeds self-effect) across all layers. These findings are quantitatively consistent with Paper XI's concentration barrier: the effective dimensionality $d_{\text{eff}} \approx 20$ bounds achievable selectivity to $k/d_{\text{eff}} \approx 1.8$, and no layer exceeds this bound. The terminal measurement limit generalizes to an *all-layer measurement limit*: domain-selective intervention via single-layer shaped noise injection is fundamentally constrained by activation geometry, not by layer choice.

1 Introduction

Paper VIII (Shaped Noise Injection) demonstrated that domain-shaped noise at terminal layers fails to achieve selective entropy control. The response matrix $\mathbf{R}$ at the output layer is invertible ($\det = -84.5$, $\kappa = 21.2$) but the $\mathbf{R}^{-1}$-optimal weight vectors produce predicted-vs-actual correlations ≈ 0. The forward pass transforms INLP directions nonlinearly between intermediate and terminal layers.

This result leaves open the question: *where* in the forward pass does domain selectivity peak? Three outcomes are possible:

Outcome A Selectivity peaks at intermediate layers ($\ell \approx 10$–20). Mid-layer injection achieves what terminal injection cannot.

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

Outcome B Selectivity is uniformly low at all layers. INLP directions never achieve clean domain targeting.

Outcome C Selectivity is highest at early layers but with negligible effect magnitude, as perturbations are diluted by the remaining forward pass.

We resolve this by injecting shaped noise at individual layers and measuring the full response profile at each.

2 Method

2.1 Model and Directions

We use Qwen-2.5 7B with 28 transformer layers ($d = 3584$). INLP directions ($k = 36$, 9 per domain) are from Paper IV's structural transfer decomposition, computed on last-token activations of the same model. Domains: medical, legal, code, science.

2.2 Single-Layer Noise Injection

We extend Paper VIII's noise injection framework to hook exactly one transformer layer. At layer ℓ, for target domain d:

$$h^{(\ell)}_{\text{last}} \leftarrow h^{(\ell)}_{\text{last}} + \sigma \cdot \|h^{(\ell)}_{\text{last}}\| \cdot P_d \epsilon \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$, $P_d = D_d^\top D_d$ is the projection onto domain d's INLP subspace (9 directions per domain), and σ is the noise scale relative to hidden state norm.

2.3 Experimental Protocol

Layer sampling. Nine layers are sampled evenly across the 28-layer stack: $\ell \in \{0, 3, 7, 10, 14, 17, 20, 24, 27\}$.

Sigma selection. Papers VIII v1–v3 used adaptive sigma sweeps that proved unreliable (criteria favored either no effect or destructive over-perturbation). We instead use two fixed sigma values:

- $\sigma = 0.05$: matches Paper VIII's terminal injection range
- $\sigma = 0.2$: provides 4× stronger perturbation, compensating for single-layer vs. four-layer injection

Baselines. Generate 32 tokens for each of 160 domain probes (40 per domain) and 10 general probes with no injection, recording mean per-token entropy. (Paper VIII used

64-token generation; the shorter window here produces higher baseline entropy due to less conditioning context, so absolute values are not directly comparable across papers.)

Heatmap. At each of 9 layers × 2 sigmas, inject noise shaped for each of 4 target domains. For each target, generate 32 tokens for all 160 domain probes and 10 general probes. Compute:

$$\mathbf{R}^{(\ell)}_{d\to j} = \frac{H^{(\ell)}_{d\to j} - H_{\text{baseline},j}}{H_{\text{baseline},j}} \times 100\% \tag{2}$$

where $H^{(\ell)}_{d\to j}$ is mean entropy on domain j probes when injecting domain d noise at layer ℓ.

2.4 Selectivity Metric

For each target domain d at layer ℓ:

$$s^{(\ell)}_d = \frac{\mathbf{R}^{(\ell)}_{d\to d} - \overline{\mathbf{R}^{(\ell)}_{d\to\neg d}}}{\max(\sigma_{\mathbf{R}},\, 0.01)} \tag{3}$$

where $\overline{\mathbf{R}^{(\ell)}_{d\to\neg d}}$ is the mean cross-domain effect and $\sigma_{\mathbf{R}}$ is the standard deviation of all four delta values. Positive selectivity means the self-domain effect exceeds bleed; negative means the opposite.

Layer-level mean selectivity: $\bar{s}^{(\ell)} = \frac{1}{4}\sum_d s^{(\ell)}_d$.

3 Results

3.1 Baselines

Table 1 shows baseline entropy (no injection) for each domain.

Table 1: Baseline mean per-token entropy (32 tokens, no injection).

Domain	Mean H (bits)
Medical	1.565
Legal	1.764
Code	1.909
Science	1.761
General	1.029

3.2 Layer-Resolved Selectivity ($\sigma = 0.05$)

Table 2 presents the self-effect (entropy change on same-domain probes), mean absolute bleed (entropy change on other-domain probes), and selectivity for each target domain at each layer.

Table 2: Per-domain self-effect, bleed, and selectivity at $\sigma = 0.05$. Bold entries indicate positive selectivity > 1.0.

	Medical			Legal			Code			Science			
ℓ	self	bleed	sel	self	bleed	sel	self	bleed	sel	self	bleed	sel	$\bar{s}$
0	$+0.5$	2.2	-0.1	$+0.6$	1.0	-0.1	-0.1	0.8	-1.5	-2.1	1.7	-1.3	-0.73
3	$+0.1$	2.9	0.5	-1.1	1.6	-0.8	$+1.2$	1.2	0.6	$+0.2$	2.9	**1.1**	0.37
7	$+1.9$	1.7	0.2	-3.2	2.0	-1.0	$+2.1$	2.3	0.8	$+1.3$	1.2	**2.0**	0.50
10	-2.8	1.9	-1.4	$+0.9$	1.2	1.0	$+2.4$	0.8	**1.7**	$+0.8$	0.9	**1.1**	0.57
14	-0.8	2.0	-1.2	$+1.5$	2.6	-1.1	$+2.3$	2.0	**1.2**	$+4.8$	2.5	**1.5**	0.12
17	-2.7	3.0	-1.3	$+0.9$	1.2	-1.8	$+2.8$	2.7	0.7	$+2.4$	3.0	**1.2**	-0.31
20	$+1.1$	2.4	-1.2	-1.1	2.6	-1.1	$+1.5$	2.4	**2.0**	-3.8	1.4	-1.4	-0.42
24	$+4.5$	1.1	**2.1**	-3.5	2.0	-1.8	-2.8	0.6	-2.1	-1.9	1.6	-0.8	-0.66
27	$+4.3$	5.9	-0.1	-0.8	4.5	-1.9	$+6.2$	5.7	0.8	$+5.7$	3.8	0.8	-0.10

Self-effect and bleed in %; selectivity is dimensionless (z-score). $\bar{s}$: layer mean selectivity.

3.3 Selectivity Curve

Figure 3 (described numerically) shows the selectivity profile across layers:

Table 3: Mean selectivity $\bar{s}^{(\ell)}$ vs. layer depth at $\sigma = 0.05$.

Layer ℓ	$\bar{s}^{(\ell)}$	Profile
0	-0.73	Anti-selective (input layer)
3	$+0.37$	Weak positive
7	$+0.50$	Moderate positive
10	$+0.57$	**Peak selectivity**
14	$+0.12$	Near zero
17	-0.31	Negative
20	-0.42	Negative
24	-0.66	Anti-selective
27	-0.10	Near zero (terminal)

The selectivity curve forms an inverted-U with peak at layer 10 (36% depth). Layers 0–3 and 17–27 show negative or near-zero mean selectivity. The positive-selectivity window spans layers 3–14.

3.4 Layer-Resolved Selectivity ($\sigma = 0.2$)

Table 4 summarizes the sigma=0.2 results. Effect magnitudes are 4–10× larger than sigma=0.05, but selectivity is generally worse.

Table 4: Summary metrics at $\sigma = 0.2$. DiagMean: mean diagonal of $\mathbf{R}^{(\ell)}$. OffDiag: mean off-diagonal. $\bar{s}$: mean selectivity.

ℓ	DiagMean (%)	OffDiag (%)	DiagDom	$\bar{s}$
0	+5.0	+2.2	+2.8	+0.83
3	+4.3	+7.2	−2.9	−0.33
7	+8.3	+10.0	−1.7	−0.66
10	+10.6	+8.7	+1.9	+0.75
14	+16.2	+20.1	−3.9	−0.13
17	+19.3	+21.2	−1.9	−0.34
20	+14.5	+16.8	−2.3	−0.56
24	+16.2	+16.2	−0.0	−0.25
27	+69.2	+66.2	+3.0	−0.50

Three findings emerge from the sigma comparison:

1. **Layer 10 peaks at both sigmas**: mean $\bar{s} = 0.57$ ($\sigma = 0.05$) and $\bar{s} = 0.75$ ($\sigma = 0.2$). This is the most robust intermediate peak.
2. **Layer 0 is sigma-sensitive**: $\bar{s} = -0.73$ at $\sigma = 0.05$ but $+0.83$ at $\sigma = 0.2$. Larger perturbation at the very first layer allows signal to propagate through the full forward pass.
3. **Terminal layer is destructive**: at $\sigma = 0.2$, layer 27 shows +115% medical self-effect with +83% bleed — the representation is obliterated.

3.5 Response Matrices at Key Layers

Table 5 shows the full 4 × 4 response matrix at the selectivity-peak layer 10 ($\sigma = 0.2$).

Table 5: Response matrix $\mathbf{R}^{(10)}$ at $\sigma = 0.2$. Diagonal entries (self-domain effects) in bold.

Target ↓ / Meas. →	Medical	Legal	Code	Science
Medical	+**12.0**	+4.9	+2.6	+11.1
Legal	+4.3	+**10.3**	+5.1	+13.2
Code	+5.8	+6.4	+**6.6**	+21.1
Science	+12.5	+10.0	+7.2	+**13.6**

All values in %. sel: med=1.46, leg=0.74, code=−0.71, sci=1.50.

At layer 10, medical, legal, and science all have positive selectivity — the self-domain effect exceeds the cross-domain mean. Code is the exception: targeting code produces the largest effect on *science* (+21.1%), not code (+6.6%).

3.6 Domain Asymmetry

A striking feature is the asymmetry between domains. Across all 9 layers:

- **Code** achieves positive selectivity at 7 of 9 layers (exceptions: 0, 24). Best: sel = 2.0 at layer 20.
- **Science** achieves positive selectivity at 6 of 9 layers (exceptions: 0, 20, 24). Best: sel = 2.0 at layer 7.
- **Medical** achieves positive selectivity at only 1 of 9 layers (layer 24, sel = 2.1). Anti-selective at 7 layers.
- **Legal** achieves positive selectivity at only 1 of 9 layers (layer 10, sel = 1.0). Anti-selective at 7 layers.

Medical and legal INLP directions produce larger effects on *other* domains than on their target domain. The INLP subspace for these domains captures shared structure that bleeds across domain boundaries.

3.7 Terminal Layer Behavior

At layer 27 (terminal), all four domains show elevated entropy (+4–6% self-effect) with high bleed (+4–6%). This matches Paper VIII's finding: terminal injection produces large effects but zero selectivity. The mean selectivity at layer 27 ($\bar{s} = -0.10$) is higher than Paper VIII's terminal result, likely because our single-layer protocol applies perturbation at one layer rather than four.

3.8 Effect Magnitude vs. Selectivity

There is an inverse relationship between effect magnitude and selectivity across layers:

- Layers 7–10 (peak selectivity): self-effects are modest (±1–3%)
- Layer 27 (terminal): self-effects are large (+4–6%) but bleed is comparably large
- Layer 24: anomalous — medical has the largest selectivity (2.1) but other domains are anti-selective

The perturbation at intermediate layers produces smaller but more targeted effects. At terminal layers, effects are larger but undifferentiated. This is consistent with the concentration barrier: intermediate layers have lower effective dimensionality (Paper XI finds $d_{\text{eff}} \approx 15$–$22$ at layers 7–14) allowing marginally better selectivity.

4 Discussion

4.1 Outcome Classification

The results represent a **weak Outcome A** embedded within Outcome B:

1. There is a clear intermediate peak (layer 10, $\bar{s} = 0.57$), confirming selectivity is not uniform across layers.

2. The peak is modest — $\bar{s} = 0.57$ means the self-domain effect is on average 0.57 standard deviations above the cross-domain mean.

3. No layer achieves $\bar{s} > 1.0$, meaning no layer produces reliable diagonal dominance in the response matrix.

The intermediate peak exists but is too weak for practical domain-selective intervention. The terminal measurement limit from Paper VIII generalizes to an *all-layer measurement limit*.

4.2 The Domain Asymmetry Problem

The 4:1 ratio between code/science (mostly positive selectivity) and medical/legal (mostly negative) suggests that INLP directions for different domains have qualitatively different relationships to the forward pass:

- **Code and science** INLP directions align with domain-specific computational pathways. Noise along these directions preferentially perturbs the model's processing of same-domain content.

- **Medical and legal** INLP directions align with *shared* structure. Noise along these directions disrupts multiple domains simultaneously, with bleed often exceeding the self-effect.

This asymmetry is not an artifact of direction quality — Paper IV showed all four domain direction sets have comparable classification accuracy ($> 90\%$) under linear probing. The asymmetry reflects forward-pass topology: code and science are functionally more distinct from each other and from the remaining domains than medical and legal are.

4.3 Connection to the Concentration Barrier (Paper XI)

Paper XI establishes that effective dimensionality d_{eff} bounds achievable selectivity via:

$$\text{INLP variance fraction} \leq \frac{k}{d_{\text{eff}}} \tag{4}$$

At the peak-selectivity layers 7–14, Paper XI measured:

- Last-token $d_{\text{eff}} \approx 15$–$22$
- $k = 36$ directions
- Bound: ≤ 1.6–2.4

Our maximum individual selectivity (code at layer 20: sel = 2.0; science at layer 7: sel = 2.0; medical at layer 24: sel = 2.1) clusters near the concentration barrier bound. The mean selectivity ($\bar{s} \leq 0.57$) remains well below, consistent with the bound applying per-direction while selectivity averages over domains.

The concentration barrier theorem predicts that selectivity is bounded by geometry, not by layer choice. Our results confirm this: varying the injection layer shifts the selectivity profile but cannot exceed the geometric bound.

4.4 Implications for Paper VIII

Paper VIII's terminal measurement limit is a special case of a deeper phenomenon. The response matrix $\mathbf{R}$ is layer-dependent but selectivity-bounded at every layer. Specifically:

- Terminal $\mathbf{R}$ has large off-diagonal entries relative to diagonal (low selectivity)
- Intermediate $\mathbf{R}$ has smaller entries overall but modestly better diagonal-to-off-diagonal ratio (higher selectivity)
- The $\mathbf{R}^{-1}$ correction strategy fails at all layers because the nonlinearity is distributed throughout the forward pass, not localized at any layer

4.5 Implications for Papers X–XII

These results constrain the remaining papers in the series:

- **Paper X (Spectral Geometry):** The Jacobian's singular structure should show INLP directions being attenuated at layers 17+ (where selectivity collapses) and partially preserved at layers 7–14 (where selectivity peaks).

- **Paper XI (Concentration Barrier):** Confirmed — the bound holds empirically at all layers.
- **Paper XII (Channel Capacity):** The maximum information gain from shaped noise is bounded by the concentration barrier. The optimal injection layer (10) provides the ceiling, and that ceiling is low.

5 Conclusion

We mapped domain selectivity across all layers of Qwen-2.5 7B by injecting shaped noise at nine sampled transformer layers. The key findings are:

1. Selectivity peaks at intermediate layers (7–10, $\bar{s} \approx 0.5$) and decays toward both input and terminal layers.
2. The peak is too modest for practical domain-selective intervention — no layer achieves mean selectivity > 1.0.
3. Strong domain asymmetry: code and science respond selectively; medical and legal are anti-selective at most layers.
4. Terminal layer behavior (large effects, zero selectivity) extends Paper VIII's finding.
5. The concentration barrier theorem (Paper XI) correctly predicts the selectivity ceiling at all layers.

The terminal measurement limit is not about terminals — it is about geometry. Single-layer shaped noise injection, at any depth, cannot overcome the concentration barrier imposed by high-dimensional activation spaces with $d_{\text{eff}} \approx 20$. Domain-selective intervention through noise injection appears to be fundamentally bounded.

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 XI: The Concentration Barrier.