The Geometry of Failure:

Why Direction-Space Interventions Cannot Achieve Domain Selectivity in Neural Networks

Jeremy McEntire*

April 2026

Abstract

In plain terms: we tried to use the internal structure of neural networks to control what domains they generate text about, and we proved why that approach has fundamental geometric limits.

When you ask a language model about medicine, it draws on the same formal reasoning it uses for law. This shared computational substrate makes it fundamentally difficult to perturb one domain without disturbing others. We report a systematic attempt to achieve domain-selective control of neural network outputs via direction-space intervention using linear discriminative directions (INLP, DAS, LEACE), and the four independent measurements that diagnose why it fails.

Using iterative nullspace projection (INLP) directions as a domain basis, we inject shaped noise into the activation space of Qwen 2.5 models (0.5B–7B parameters) targeting four domains: medical, legal, code, and science. The intervention produces measurable effects—3–6% entropy reduction in target domains and 100% repetition loop escape—but cross-domain selectivity is uniformly poor. At 7B, targeting medical reduces legal entropy 5.6× more than medical entropy. Three correction attempts (subspace decomposition, scalar cancellation, R^{-1}-optimal weighting) all fail; the predicted and actual response vectors are uncorrelated.

We diagnose this failure through four converging measurements. (1) *Layer-resolved selectivity*: mean selectivity peaks at layer 10 ($\bar{s} = 0.57$) in a 28-layer model; no layer achieves $\bar{s} > 1.0$. (2) *Spectral isotropy*: the INLP/random Jacobian-vector-product amplification ratio is 0.991; the forward pass treats domain directions identically to random directions. (3) *Concentration barrier*: we prove that any k directions in a space

*Correspondence: jmc@cageandmirror.com

with effective dimensionality d_{eff} capture at most a variance fraction k/d_{eff}, and verify this bound at all 28 layers ($d_{\text{eff}}^{\text{LT}}$ mean = 19.2). (4) *Information-theoretic quantification*: domain-specific output perturbation is 1–2 bits against ∼15 bits of domain-agnostic perturbation, consistent with an approximate bound of 2.24 bits from the concentration barrier.

The mechanism succeeds when selectivity is not required: shaped noise breaks 100% of repetition loops with near-perfect token uniqueness, demonstrating that the intervention works—it simply cannot be made selective. The four measurements converge: domain-selective intervention via direction-space methods using linear discriminative directions is constrained by a geometric barrier, not an engineering limitation. The forward pass is an isotropic amplifier that provides no directional leverage. Classification accuracy does not imply intervention precision—the concentration barrier is why entanglement between domains exists.

Keywords: activation geometry, domain selectivity, concentration of measure, shaped noise, INLP, stochastic resonance, effective dimensionality, terminal measurement limit

1 Introduction

When you ask a language model about medicine, the internal representations it activates overlap substantially with those it uses for law. Both domains demand structured argumentation, technical vocabulary, evidential reasoning, and formal register. This overlap is not a defect of the model's training—it is a consequence of the shared computational substrate that makes general-purpose language models general-purpose. But it creates a fundamental problem for anyone who wants to intervene on one domain without disturbing others.

This paper reports what happens when you try. We shaped noise to domain-discriminative directions in activation space and injected it into transformer models at inference time, attempting to control the output distribution for one domain while leaving others untouched. The intervention produced measurable effects: modest entropy reductions (3–6%) in target domains and perfect repetition loop breaking. But it failed at the task that matters: domain selectivity. Targeting medical reduced legal entropy 5.6× more than medical entropy. Three progressively stronger correction attempts—subspace decomposition, scalar cancellation, and the mathematically optimal linear correction via response-matrix inversion—all failed. The predicted and actual response vectors were uncorrelated.

This paper reports both a positive result (loop breaking) and a systematic negative result (domain selectivity). The positive result confirms the mechanism functions; the negative results explain why it cannot be targeted.

We then asked why. Four independent measurements, conducted across all 28 layers of Qwen 2.5 7B (Qwen Team, 2025), converge on the same answer: the failure is geometric, not engineering.

First, we mapped selectivity at every layer and found it peaks weakly at intermediate layers ($\bar{s} = 0.57$ at layer 10) and collapses toward both input and output layers. No layer achieves mean selectivity above 1.0. Second, we measured how the forward pass amplifies perturbations along INLP directions versus random directions and found no difference: the INLP/random amplification ratio is 0.991. The forward pass does not know which directions are domain-discriminative. Third, we proved and verified a concentration barrier: in a space with effective dimensionality $d_{\text{eff}} \approx 20$, any $k = 36$ directions capture at most a fraction $k/d_{\text{eff}} \approx 1.8$ of the total variance, bounding the geometric footprint of any domain-specific intervention. Fourth, we measured the information content of the perturbation and found that domain-specific output perturbation is 1–2 bits against ∼15 bits of domain-agnostic perturbation, consistent with the concentration barrier bound.

These results have implications beyond our specific intervention method. The concentration barrier constrains direction-space interventions using linear discriminative directions

(INLP, DAS, LEACE)—and likely extends to activation addition (Turner et al., 2023), representation engineering (Zou et al., 2023), ROME (Meng et al., 2022), and linear probe-based steering—because they all rely on the assumption that directions identified via linear methods have approximately linear effects on output behavior. Our results show this assumption breaks down precisely when selectivity is required. The barrier is a property of high-dimensional computation, not a limitation of any particular method. We emphasize that these results apply to interventions along linearly discriminative directions. Methods that operate nonlinearly, across multiple layers simultaneously, or in non-discriminative subspaces may achieve selectivity that linear single-layer injection cannot.

The connection to the broader structure of neural network representations is direct. McEntire (2026d) established that domain-specific and structural information are entangled in activation space: INLP decomposition separates them for classification purposes, but the forward pass mixes them during inference. The concentration barrier is the geometric mechanism that produces this entanglement. In high dimensions, the geometric footprint of domain-specific directions is necessarily small relative to the full activation space, and the forward pass provides no spectral mechanism to amplify it. The entanglement is not incidental—it is guaranteed by the geometry.

The paper is organized as follows. Section 2 reviews related work on direction-space intervention and concentration of measure. Section 3 describes the experimental setup. Section 4 reports the intervention and its failure. Section 5 presents the four diagnostic measurements. Section 6 synthesizes the results and draws implications. Sections 7 and 8 address limitations and conclude.

2 Background and Related Work

Direction-space intervention in neural networks rests on a simple premise: if a linear probe can identify a direction in activation space that correlates with a concept, then perturbing activations along that direction should influence the model's behavior with respect to that concept. The premise has been productive. Activation addition (Turner et al., 2023) shows that adding a fixed vector to intermediate representations shifts model behavior in interpretable ways. Representation engineering (Zou et al., 2023) identifies "reading vectors" for concepts like honesty and uses them for both monitoring and control. ROME (Meng et al., 2022) edits factual associations by rank-one updates to specific MLP layers. These methods achieve measurable steering effects, and they share a common structure: identify a direction, apply a perturbation along it, observe the downstream consequence.

The implicit assumption in all of these methods is that the relationship between direction

and effect is approximately linear. If adding a helpfulness vector increases helpfulness, then adding twice the vector should increase it roughly twice as much, and subtracting it should decrease it. Our results test this assumption to destruction.

2.1 INLP and Linear Probing

Iterative nullspace projection (INLP) (Ravfogel et al., 2020) finds directions in activation space that maximally separate categories (in our case, domains) by training a sequence of linear classifiers and projecting into the nullspace of each. The resulting directions are exactly orthogonal by construction and achieve near-perfect classification accuracy. INLP was designed for information removal—projecting out protected attributes—but the directions it discovers are natural candidates for intervention, since they capture the linear structure that distinguishes domains.

The critical distinction is between *classification* and *intervention*. Classification requires only that a direction correlates with the target concept—shared substrate that happens to predict the label will serve. Intervention requires that perturbation along the direction causally affects only the target concept. In high dimensions, the gap between these two requirements is large.

2.2 Concentration of Measure

The concentration of measure phenomenon (Ledoux, 2001; Vershynin, 2018) describes how geometric intuitions from low dimensions fail catastrophically in high dimensions. For uniformly random unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$:

$$\mathbb{E}[|\langle \mathbf{u}, \mathbf{v} \rangle|] = O(d^{-1/2}) \tag{1}$$

At $d = 3584$ (the hidden dimension of Qwen 2.5 7B), this is approximately 0.017. Nearly all pairs of vectors are nearly orthogonal, regardless of their functional relationship. The volume of the unit ball approaches zero super-exponentially. Mass concentrates in thin shells. Directions that are geometrically orthogonal can activate overlapping computational pathways, because the distance between their projections through nonlinear transformations is not constrained by their angle in the input space.

This has a direct consequence for direction-space intervention: orthogonality of concept directions—frequently reported as evidence of clean separation—is a geometric tautology in high dimensions, not an empirical finding. Any 36 vectors in $\mathbb{R}^{3584}$ will be nearly orthogonal even if drawn at random.

2.3 Stochastic Resonance

Stochastic resonance (SR) (Benzi et al., 1981; Gammaitoni et al., 1998) is the phenomenon where noise improves signal detection in nonlinear systems. The communicative variance framework (McEntire, 2026b) identifies five sufficient conditions (C1–C5) under which SR produces net-beneficial effects in neural networks, with C1 (suboptimality) as the gate. Our noise injection protocol is a controlled test of domain-specific SR: shaped noise should benefit domain-specific outputs when the model's baseline is suboptimal on those domains. The inverted-U relationship between noise amplitude and effect size is a hallmark of SR and appears throughout our measurements.

3 Methods

3.1 Models

All experiments use the Qwen 2.5 model family (Qwen Team, 2025). Scale experiments (Section 4) use four scales: 0.5B, 1.5B, 3B, and 7B parameters. The diagnostic measurements (Section 5) focus on the 7B model (28 transformer layers, hidden dimension $d = 3584$). Models are loaded in float16 precision on A100 GPUs via HuggingFace Transformers.

3.2 Probes and Domains

We use 160 domain probes (40 per domain, 4 syntactic shapes) covering medical, legal, code, and science domains, drawn from McEntire (2026a). An additional 10 general-knowledge probes serve as controls. For loop-breaking experiments, 20 prompts known to induce repetitive generation are used. For cross-domain experiments, 150 prompts (25 per domain pair, 6 pairs) from McEntire (2026c) test compositional domain knowledge.

3.3 INLP Directions

Domain directions come from INLP (Ravfogel et al., 2020) applied to activation differences between domain-specific and general text, as computed in McEntire (2026a). At 3B, this yields 13 directions in $\mathbb{R}^{2048}$ (4 medical, 3 each for legal, code, science). At 7B, 36 directions in $\mathbb{R}^{3584}$ (9 per domain). By construction, INLP directions across domains are exactly orthogonal: mean pairwise cosine similarity $< 10^{-9}$.

3.4 Shaped Noise Injection Protocol

The core mechanism registers forward hooks on transformer layers. At each hooked layer ℓ, the hidden state $\mathbf{h} \in \mathbb{R}^{b \times s \times d}$ is modified:

$$\mathbf{h}' = \mathbf{h} + \sigma \cdot \|h^{(\ell)}_{\text{last}}\| \cdot P_S \, \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I_d) \tag{2}$$

where σ is the noise amplitude relative to hidden state norm, $\boldsymbol{\epsilon}$ is isotropic Gaussian noise, and $P_S = D_S^\top D_S$ is the projection matrix onto the target subspace S spanned by k unit-norm domain directions $D_S \in \mathbb{R}^{k \times d}$.

For the initial intervention experiments (Section 4), noise is injected at the final four transformer layers. For the layer-resolved diagnostic (Section 5.1), noise is injected at a single layer at a time.

3.5 Metrics

Entropy metrics. For each probe, we generate tokens with greedy decoding and record the full logit distribution. Mean per-token entropy: $\bar{H} = \frac{1}{T}\sum_{t=1}^{T} H(p_t)$ where $H(p_t) = -\sum_v p_t(v) \log_2 p_t(v)$. The cross-domain response matrix $\mathbf{R}$ records percentage entropy change:

$$\mathbf{R}_{d \to j} = \frac{H_{d \to j} - H_{\text{baseline},j}}{H_{\text{baseline},j}} \times 100\% \tag{3}$$

where $H_{d \to j}$ is mean entropy on domain j probes when injecting domain d noise.

Selectivity. For target domain d:

$$s_d = \frac{\mathbf{R}_{d \to d} - \overline{\mathbf{R}_{d \to \neg d}}}{\max(\sigma_{\mathbf{R}},\, 0.01)} \tag{4}$$

where $\overline{\mathbf{R}_{d \to \neg d}}$ is the mean cross-domain effect and $\sigma_{\mathbf{R}}$ is the standard deviation of all four response values. Positive selectivity means the self-domain effect exceeds bleed.

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

$$A(\ell, \ell', v) = \frac{\|h^{(\ell')}_{\text{perturbed}} - h^{(\ell')}_{\text{clean}}\|}{\epsilon \cdot \|v\|} \tag{5}$$

computed via finite differences with $\epsilon = 0.01$.

Effective dimensionality. The participation ratio of the activation covariance:

$$d_{\text{eff}} = \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2} \tag{6}$$

where λ_i are the eigenvalues of the centered activation covariance at a given layer.

KL divergence. We approximate $D_{\mathrm{KL}}(P_\sigma \| P_0)$ using the union of top-100 tokens from both distributions at each generation step:

$$D_{\mathrm{KL}}(P_\sigma \| P_0) \approx \sum_{i \in S_P \cup S_Q} p_\sigma(i) \log_2 \frac{p_\sigma(i)}{p_0(i)} \tag{7}$$

4 The Intervention and Its Failure

4.1 Scale-Entropy Baselines

Before injection, we establish how output entropy varies with model scale. Table 1 reports mean per-token entropy across 64 generated tokens for each domain at four model scales.

Model	Mean $\bar{H}$	Std	Medical	Legal	Code	Science
Qwen 0.5B	2.382	1.046	2.034	2.429	2.607	2.457
Qwen 1.5B	1.960	0.920	1.760	1.990	2.119	1.969
Qwen 3B	1.769	1.006	1.407	1.584	2.283	1.803
Qwen 7B	1.484	0.628	1.337	1.513	1.678	1.407

Table 1: Scale-entropy baselines. Mean per-token entropy decreases monotonically from 2.38 bits at 0.5B to 1.48 bits at 7B, a 37.7% reduction. Code shows the highest entropy at every scale; medical the lowest.

Mean entropy decreases monotonically with scale: larger models are more confident, not less. This is not counterintuitive. Training is a law-of-large-numbers process: a 7B model trained on trillions of tokens has seen orders of magnitude more gradient signal, and the central limit theorem guarantees its next-token posterior concentrates more tightly. The practical consequence is that the C1 suboptimality gap (McEntire, 2026b)—the precondition for stochastic resonance benefit—is small on domain probes at 7B.

4.2 Domain Precision via Shaped Noise

We swept noise amplitude $\sigma \in \{0, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0\}$ across all domain-mode combinations. Table 2 reports the best entropy reduction per domain at 7B.

Three patterns emerge. Optimal σ values cluster in the 0.005–0.02 range, consistent with the stochastic resonance inverted-U. The best mode varies by domain (positive for medical, both for legal, negative for code), suggesting domain-specific subspace geometry matters. Science is unresponsive at 7B, indicating its representations are already near the achievable entropy floor within the INLP subspace.

Domain	Best Mode	σ^*	$\Delta\bar{H}$	Reduction
Medical	positive	0.020	-0.045	-3.3%
Legal	both	0.010	-0.092	-6.1%
Code	negative	0.005	-0.054	-3.2%
Science	—	—	—	0.0%

Table 2: Best stochastic resonance results at 7B. Optimal σ values cluster in the 0.005–0.02 range, consistent with the SR inverted-U. Science is completely unresponsive.

4.3 Loop Breaking

Independent of selectivity, shaped noise breaks degenerate attractor states. We compared four methods on 20 loop-inducing prompts at both 3B and 7B (Table 3).

Scale	Method	Escape	Escapes	Unique	$\bar{H}$
3B	Baseline	20%	4/20	0.17	0.90
	Temperature	70%	14/20	0.44	—
	Rep. penalty	100%	20/20	0.86	2.94
	Shaped SR	100%	20/20	**1.00**	8.39
7B	Baseline	30%	6/20	0.28	1.05
	Temperature	55%	11/20	0.41	—
	Rep. penalty	95%	19/20	0.88	3.12
	Shaped SR	100%	20/20	**0.99**	10.03

Table 3: Loop breaking comparison. Shaped SR achieves 100% escape rate with near-perfect token uniqueness at both scales. The trade-off is high output entropy (8.4–10.0 bits), reflecting distributional flattening from the noise.

Shaped SR breaks every loop at both scales with token uniqueness ≥ 0.99. The mechanism works—it destabilizes degenerate modes. Loop breaking succeeds because it does not require selectivity. The question is whether the effect can be made selective.

4.4 The Cross-Domain Response Matrix

Table 4 reports the full 4×4 response matrix at both 3B and 7B, where entry $\mathbf{R}_{ij}$ is the percentage entropy change in domain j when noise is injected along domain i's directions.

The 7B matrix reveals the core problem. When targeting medical, the largest effect is on *legal* (-10.7%), not medical (-1.9%). The medical INLP direction reduces legal entropy $5.6\times$ more than it reduces medical entropy. Legal targeting produces a self-effect of only -0.4% with $+3.8\%$ bleed into medical. At 3B, noise predominantly *increases* entropy across all domains, with off-diagonal entries often exceeding diagonal entries.

3B Response Matrix ($\sigma = 0.01$)					
Target ↓	**Med.**	**Legal**	**Code**	**Sci.**	**Sel.**
Medical	+2.9*	+1.3	−1.5	+2.7	1.19
Legal	+8.6	+3.6*	−0.2	+4.3	−0.19
Code	+6.3	+8.2	−3.3*	+3.7	−2.15
Science	+6.8	+1.0	+1.6	+12.3*	2.01
7B Response Matrix (domain-optimal σ)					
Target ↓	**Med.**	**Legal**	**Code**	**Sci.**	**Sel.**
Medical	−1.9*	−10.7	−1.5	−1.0	0.63
Legal	+3.8	−0.4*	−1.7	−0.1	−0.52
Code	+0.7	−2.0	−1.2*	+3.0	−0.91
Science	+0.6	−1.3	−0.1	+1.2*	1.55

Table 4: Cross-domain response matrices. Asterisks (*) mark diagonal (target) entries. At 7B, medical targeting produces a −10.7% change in legal entropy—5.6× the medical self-effect of −1.9%. Selectivity is poor at both scales.

4.5 Three Failed Corrections

We made three progressively stronger attempts to correct the cross-domain bleed.

4.5.1 Subspace Decomposition

One hypothesis is that INLP directions, despite orthogonality, share variance attributable to general linguistic structure. We computed register directions via SVD of the full direction matrix $D \in \mathbb{R}^{k \times d}$ and projected each domain's directions into the nullspace of the top-8 singular vectors. The spectral gap confirmed a dominant shared component ($\sigma_1 = 684.5$, $\sigma_2 = 282.0$, ratio 4.7×). However, the residual directions were nearly unchanged: mean cosine similarity between raw and register-projected directions exceeded 0.97 across all domains. Meanwhile, contrastive directions (without INLP's nullspace constraint) showed high cross-domain similarity (mean pairwise cosine ≈ 0.25, max 0.82–0.87), confirming the cross-domain bleed is a property of the model's forward pass, not of the direction-finding algorithm.

4.5.2 Scalar Interference Cancellation

If the bleed cannot be removed from the directions, perhaps it can be cancelled in the output. For each target domain i, we estimated scalar cancellation coefficients $\lambda_{ij} = R_{ij}/R_{jj}$ for each

non-target domain j and constructed cancelled directions:

$$\mathbf{d}_{\text{cancel}} = \bar{\mathbf{d}}_i - \sum_{j \neq i} \lambda_{ij} \cdot \bar{\mathbf{d}}_j \tag{8}$$

We swept a multiplier $m \in \{0, 0.25, 0.5, 1.0, 2.0, 4.0\}$. Every domain's best selectivity was negative. Medical targeting at $m = 4.0$ achieved a -7.2% self-effect but generated 15.3% mean bleed. Code's best operating point was $m = 0$ (no cancellation at all).

4.5.3 Optimal Linear Correction via R^{-1}

The strongest test: given the full response matrix $\mathbf{R} \in \mathbb{R}^{4\times4}$, find $\mathbf{w}_k$ such that $\mathbf{R}^\top \mathbf{w}_k = -\mathbf{e}_k$. If $\mathbf{R}$ is invertible, the solution is $\mathbf{w}_k = -\mathbf{R}^{-\top}\mathbf{e}_k$. We verified invertibility:

$$\det(\mathbf{R}) = -84.5 \tag{9}$$

$$\text{rank}(\mathbf{R}) = 4 \tag{10}$$

$$\kappa(\mathbf{R}) = 21.2 \tag{11}$$

The optimal weight vectors exist and are computable. Table 5 shows what happens when they are applied.

R^{-1} **Optimal Weight Vectors (7B)**					
Target ↓	**Med.**	**Legal**	**Code**	**Sci.**	**Sel.**
Medical	+0.1*	−5.2	+1.2	+7.4	−0.23
Legal	−3.6	−0.1*	+0.7	+1.4	0.20
Code	−1.1	−0.5	+0.3*	+1.6	0.30
Science	−2.4	+1.1	+1.2	−2.1*	−1.21

Table 5: Response under R^{-1}-optimal weight vectors. For the medical target, the prediction was $[-1.0, 0, 0, 0]$; the actual response was $[+0.1, -5.2, +1.2, +7.4]$. The self-effect has the wrong sign. The mathematically optimal linear correction fails completely.

For the medical target, the predicted response was $[-1.0, 0.0, 0.0, 0.0]$. The actual response was $[+0.1, -5.2, +1.2, +7.4]$. The self-effect has the wrong sign. Legal bleed persists at -5.2% despite a prediction of zero. Science shows $+7.4\%$ where zero was predicted. The predicted and actual vectors are uncorrelated.

This is the strongest possible evidence against linearity. The response matrix is invertible, the linear algebra is clean, and the optimal weight vectors produce effects that bear no resemblance to predictions. The system is fundamentally nonlinear.

4.6 Progression Across Corrections

Table 6 shows the full progression. Each correction attempt does not incrementally improve the response—it produces a qualitatively different pattern. Raw injection shows modest self-effects with substantial bleed. Scalar cancellation amplifies both. Optimal linear correction produces near-zero self-effects with persistent bleed. The system responds to each intervention differently, confirming the mapping from direction to output entropy is not any linear function of the input.

	Raw INLP		**Scalar Cancel.**		**R^{-1} Optimal**	
Target	**Self**	**Sel.**	**Self**	**Sel.**	**Self**	**Sel.**
Medical	−1.9	0.63	+6.3	−1.68	+0.1	−0.23
Legal	−0.4	−0.52	+3.4	0.74	−0.1	0.20
Code	−1.2	−0.91	−2.7	−1.05	+0.3	0.30
Science	+1.2	1.55	+7.8	2.05	−2.1	−1.21

Table 6: Progressive correction attempts at 7B. Each correction changes the response pattern entirely rather than improving selectivity, confirming nonlinearity.

4.7 The Legal Direction as Shared Substrate

The legal INLP direction provides the sharpest illustration of why direction-space intervention fails. Three measurements converge:

Near-zero self-effect. Legal targeting produces −0.4% self-effect with +3.8% bleed into medical. The legal direction moves other domains more than it moves legal.

Zero optimal self-weight. The R^{-1} optimization assigns weight $w_{\text{legal}} = 0.000$ to the legal direction for the legal target. The algorithm has discovered that the legal INLP direction contains no uniquely legal signal that survives the forward pass. To target legal, the optimal strategy combines medical, code, and science directions only.

Asymmetric cancellation. The cancellation coefficient $\lambda_{\text{legal}\to\text{medical}} = -10.35$. The negative sign means the legal direction's effect on medical runs *opposite* to its effect on legal. This is not attenuation—it is a qualitatively different response, consistent with the legal direction activating a shared processing pathway.

The interpretation is that legal language draws on the same formal register, structured argumentation, and technical vocabulary as other domains. What INLP calls "the legal direction" is more accurately the direction of maximum shared formality. Classification exploits this shared substrate because it correlates with the label; intervention fails because perturbing shared substrate affects all domains that share it.

5 Diagnosing the Failure

The intervention experiments established the *terminal measurement limit*: the cross-domain response matrix characterizes the system's output but cannot invert the nonlinear mixing at intermediate layers. The following four measurements diagnose *why* this limit exists, each approaching the problem from a different angle. Together they show the failure is geometric.

5.1 Layer-Resolved Selectivity

If the terminal layers are the wrong place to intervene, perhaps intermediate layers are better. To find out, we injected shaped noise at individual layers across the full depth of the 28-layer model.

We sampled nine layers evenly across the stack: $\ell \in \{0, 3, 7, 10, 14, 17, 20, 24, 27\}$. At each layer, for each of four target domains, we injected noise shaped to that domain's INLP subspace (9 directions per domain) and measured entropy changes on all 160 domain probes plus 10 general probes. Two noise scales were tested: $\sigma = 0.05$ (matching the terminal injection range) and $\sigma = 0.2$ (providing $4\times$ stronger perturbation).

5.1.1 Baselines

Table 7 shows baseline entropy for the layer-resolved experiment (32-token generation, no injection).

Table 7: 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

5.1.2 Selectivity Profile

Table 8 presents the full layer-resolved selectivity data at $\sigma = 0.05$.

The selectivity curve forms an inverted-U with a clear peak:

Selectivity peaks at layer 10 (36% depth) with $\bar{s} = 0.57$ and declines toward both input and output layers. The positive-selectivity window spans layers 3–14. No layer achieves $\bar{s} > 1.0$, meaning no layer produces reliable diagonal dominance in the response matrix.

Table 8: 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.

Table 9: 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)

5.1.3 Stronger Perturbation ($\sigma = 0.2$)

At $\sigma = 0.2$, effect magnitudes are 4–10× larger but selectivity is generally worse. Table 10 summarizes. Layer 10 peaks at both sigmas ($\bar{s} = 0.57$ at $\sigma = 0.05$, $\bar{s} = 0.75$ at $\sigma = 0.2$). Terminal layer 27 at $\sigma = 0.2$ shows +115% medical self-effect with +83% bleed—the representation is obliterated.

5.1.4 Response Matrix at the Selectivity Peak

Table 11 shows the full 4×4 response matrix at the peak-selectivity layer 10 ($\sigma = 0.2$).

Even at the best layer, the matrix is far from diagonal. Targeting code produces the largest effect on *science* (+21.1%), not code (+6.6%). The modest peak at layer 10 is real but insufficient for practical domain-selective intervention. The terminal measurement limit

Table 10: Summary metrics at $\sigma = 0.2$.

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

Table 11: Response matrix $\mathbf{R}^{(10)}$ at $\sigma = 0.2$. Diagonal entries 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 %. Selectivity: med=1.46, leg=0.74, code=−0.71, sci=1.50.

generalizes to an *all-layer measurement limit*.

5.1.5 Domain Asymmetry

A striking feature persists across all layers. At $\sigma = 0.05$:

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

Medical and legal INLP directions produce larger effects on *other* domains than on their targets. The INLP subspace for these domains captures shared structure that bleeds across boundaries. This asymmetry is not an artifact of direction quality—all four domain direction sets achieve comparable classification accuracy ($> 90\%$) under linear probing (McEntire, 2026a). The asymmetry reflects forward-pass topology: code and science are functionally more distinct than medical and legal.

5.2 Spectral Isotropy of the Forward Pass

The layer-resolved measurements show selectivity peaks at intermediate layers but do not explain *why*. Two hypotheses suggest themselves.

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 affect the bulk of the representation.

We test both using Jacobian-vector products (JVPs) and PCA-INLP alignment measurements.

5.2.1 JVP Amplification Spectrum

For each of seven injection layers ($\ell \in \{0, 4, 9, 14, 18, 22, 27\}$), we computed the amplification factor to the terminal layer for four INLP mean directions (one per domain) and eight random unit directions, averaged over eight domain probes (Table 12).

Table 12: Amplification to terminal layer 27. $\bar{A}_{\text{INLP}}$: mean over 4 INLP directions. $\bar{A}_{\text{rand}}$: mean over 4 random directions. Layer 27 omitted (zero downstream propagation).

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

Values near 1.0 indicate isotropic amplification (no preferential treatment of INLP vs. random directions); values significantly above or below 1.0 would indicate spectral bias.

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. The forward pass does not know which directions are INLP directions.

Per-domain amplification (Table 13) shows that the code INLP direction has notably higher amplification from layer 0 (2938 vs. ∼1700–1900 for other domains), but individual

Table 13: Per-domain INLP amplification to terminal 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

random directions show comparable spread. This reflects direction geometry, not a systematic domain effect.

Three structural features of the amplification spectrum are notable regardless of direction type. First, amplification from injection to terminal grows monotonically with distance, roughly 1.06× per layer. Second, the final five layers (22→27) contribute disproportionately—a terminal spike. Third, individual random directions span a 10× range, reflecting direction-dependent coupling to the Jacobian's singular structure, with INLP directions showing comparable variance.

5.2.2 PCA-INLP Alignment

At each sampled layer, we captured last-token activations across 160 domain probes, computed the top-20 PCA directions, and measured mean absolute cosine similarity between PCA and INLP directions (720 cosine values per layer). For $d = 3584$ and $k = 20$, the expected random baseline is mean $|\cos\theta| \approx 0.013$ (Table 14).

Table 14: PCA-INLP alignment at each layer. 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 selectivity peaks), alignment is near the random baseline.

H2 is rejected. Alignment and selectivity are *dissociated*: alignment peaks at layer 27, where mean selectivity is $\bar{s} = -0.10$. 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 the most anti-selective domain. Alignment with top PCA directions does not produce selectivity—it produces bleed, because the top PCA components carry domain-agnostic variance.

5.2.3 The Forward Pass as Isotropic Amplifier

Both hypotheses are rejected. The forward pass amplifies all perturbations roughly equally, regardless of alignment with INLP or PCA directions. Each layer multiplies perturbation magnitude by $\sim$1.06$\times$ on average. The nonlinearity does not selectively filter INLP directions—it treats them as generic. Without preferential amplification, the only leverage for selectivity comes from the geometric overlap between INLP directions and the activation subspace—exactly what the concentration barrier measures.

The terminal amplification spike (layers 22–27) 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.

5.3 The Concentration Barrier

The spectral measurements show the forward pass provides no directional leverage. The concentration barrier theorem formalizes the consequence: geometry alone bounds achievable selectivity.

Intuitively, this theorem says that in a high-dimensional space, no small set of directions can capture a large fraction of the total variance. Domain-specific information lives in a subspace, but that subspace is too small relative to the full space to dominate the output.

Theorem 5.1 (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:*

$$d_{\mathit{eff}} = \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2} \tag{12}$$

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

$$\frac{\sum_{j=1}^{k} v_j^\top \Sigma v_j}{\mathrm{tr}(\Sigma)} \leq \frac{k}{d_{\mathit{eff}}} \tag{13}$$

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

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

$$v^\top \Sigma v = \sum_i \lambda_i (u_i^\top v)^2 \tag{14}$$

By Cauchy–Schwarz: $v^\top \Sigma v \leq \lambda_{\max}$. For k directions:

$$\sum_{j=1}^{k} v_j^\top \Sigma v_j \leq k \cdot \lambda_{\max} \tag{15}$$

Meanwhile, since $\sum_i \lambda_i^2 \geq \lambda_{\max} \sum_i \lambda_i$ (as $\lambda_{\max} \geq \lambda_i$ for all i):

$$d_{\text{eff}} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2} \leq \frac{\sum_i \lambda_i}{\lambda_{\max}} \tag{16}$$

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

$$\frac{\sum_{j=1}^{k} v_j^\top \Sigma v_j}{\mathrm{tr}(\Sigma)} \leq \frac{k \cdot \lambda_{\max}}{\mathrm{tr}(\Sigma)} \leq \frac{k}{d_{\text{eff}}} \tag{17}$$

□

Remark 5.2. The bound assumes only that activations are centered and that the participation ratio is well-defined. It does not assume uniform or Gaussian variance distribution—the participation ratio d_{eff} captures the actual spectrum shape. The bound is tightest when variance is concentrated (low d_{eff}) and loosest when variance is diffuse (high d_{eff}).

The bound says: if the activation space has high effective dimensionality, then any k directions capture a vanishingly small fraction of the variance. With $k = 36$ INLP directions in a space with $d_{\text{eff}} \approx 20$, the bound is $36/20 = 1.8$—a small fraction of total variance. The connection to behavioral selectivity is indirect but empirically consistent: where the variance fraction is bounded to be small, perturbation along those directions has limited domain-specific leverage.

5.3.1 Effective Dimensionality Across All 28 Layers

We measured d_{eff} at every transformer layer under two activation pooling strategies. Last-token pooling ($a_{\text{LT}}^{(\ell)} = h_T^{(\ell)}$) matches the INLP direction computation from McEntire (2026a). Mean pooling ($a_{\text{MP}}^{(\ell)} = \frac{1}{T}\sum_t h_t^{(\ell)}$) captures the global representation structure.

Table 15 presents the complete profile. Both INLP variance fraction and Fisher separation between domains are shown alongside the concentration barrier bound.

The two pooling methods reveal strikingly different dimensionality profiles.

Table 15: Effective dimensionality, INLP variance fraction, Fisher separation, and barrier bound at all 28 layers 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

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: sharp dip at layer 1, recovery through layers 2–7, broad plateau around 20–22 at layers 8–14, slight decline through layers 15–20, and 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, recovering to 18.9 and 22.5 only at layers 26–27. A participation ratio of 1.0 means the eigenspectrum is dominated by a single eigenvalue—the centered activation vectors all lie along essentially one direction. This is a previously undocumented form of representation anisotropy that is invisible to position-aware measurements. Domain classification from mean-pooled activations achieves 90–100%

accuracy 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 in preparation for the vocabulary projection.

5.3.2 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 15 shows $k/d_{\text{eff}}^{\text{LT}}$, ranging 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.

The bound is loose by a factor of 11–557× at layers 0–18. At the terminal layer, the gap tightens: bound 1.61 vs. actual 12.46%, a factor of ∼13×. The looseness reflects the theorem's worst-case nature: it bounds variance captured by *any* k directions, while INLP directions are biased toward domain-relevant variance, which is a small fraction of total variance.

5.3.3 The INLP Gradient

The INLP variance fraction increases monotonically toward terminal layers under both pooling:

- **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 indicates the forward pass progressively constructs domain-discriminative structure. INLP directions (learned from terminal-layer activations) are progressively more aligned with the activation space as we approach their layer of origin. It also explains why early-layer injection may fail: perturbation along INLP directions at layer 3 affects only 1.6% of activation variance, and subsequent layers may attenuate this further.

5.3.4 Fisher Separation

The Fisher discriminant ratio (mean across six domain pairs) follows a non-monotonic profile under last-token extraction, peaking at layers 8–10 (Fisher ≈ 0.24), dipping through layers 15–21, and rising 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 ∼0.06 at layers 3–18, consistent with $d_{\text{eff}} = 1.0$.

5.4 Information-Theoretic Quantification

The concentration barrier bounds how much variance domain directions can capture. The final measurement asks: how much domain-specific information does the perturbation actually carry to the output?

We injected domain-shaped noise at the optimal layer (layer 10) across a sigma sweep $\sigma \in \{0.01, 0.05, 0.1, 0.2, 0.5, 1.0\}$ and measured the KL divergence between noisy and clean output distributions for 20 probes per domain (32 tokens each).

5.4.1 KL Divergence Profiles

Table 16 presents the KL divergence for each domain across the sigma sweep. All four domains exhibit inverted-U profiles, confirming stochastic resonance in the information domain.

Table 16: Mean $D_{\mathrm{KL}}(P_\sigma \| P_0)$ in bits, by domain and σ. Bold: peak value per domain.

Domain	$\sigma = 0.01$	0.05	0.1	0.2	0.5	1.0
Medical	2.79	8.34	9.76	13.55	**14.55**	12.43
Legal	2.49	7.04	9.93	**15.12**	14.62	11.18
Code	2.65	7.84	9.48	**15.04**	14.60	11.36
Science	0.52	7.00	8.90	13.05	**14.03**	11.31

20 probes × 32 tokens per condition. σ relative to hidden state norm.

KL rises steeply from $\sigma = 0.01$ to $\sigma = 0.2$, peaks at $\sigma = 0.2$–0.5, then declines at $\sigma = 1.0$. Peak KL is ∼14–15 bits for all domains. Medical and science peak at $\sigma = 0.5$; legal and code peak at $\sigma = 0.2$. The decline at $\sigma = 1.0$ (15–26% drop from peak) confirms that excessive noise destroys information rather than adding it—the SR inverted-U in information space.

5.4.2 Cross-Domain KL Matrix

Table 17 presents the KL divergence measured on each domain's probes when injecting noise shaped for each target domain at the target's optimal sigma.

Table 17: Cross-domain KL matrix at optimal σ^* for each target. Rows: noise target. Columns: measurement domain. Self-KL in bold.

Target ↓ / Meas. →	Medical	Legal	Code	Science
Medical ($\sigma^* = 0.5$)	**14.76**	14.53	14.05	11.89
Legal ($\sigma^* = 0.2$)	16.05	**16.63**	15.13	13.13
Code ($\sigma^* = 0.2$)	14.56	16.93	**13.85**	13.16
Science ($\sigma^* = 0.5$)	15.38	15.26	15.76	**14.00**

All values in bits. Computed at the sigma that maximizes target-domain KL.

The matrix is far from diagonal. Code-targeted noise produces larger KL on legal (16.93) than on code (13.85)—code noise perturbs legal outputs more than code outputs. Science-targeted noise produces larger KL on code (15.76) and medical (15.38) than on science (14.00).

5.4.3 Domain-Specific Differential

The central distinction is between *total* output perturbation and *domain-specific* output perturbation. Table 18 separates the two.

Table 18: Domain-specific analysis. Δ_d: self-KL minus mean cross-KL (domain-specific differential). IS: information selectivity (self-KL / mean cross-KL). Bound: approximate upper bound from d_{eff}.

Domain	Peak KL (bits)	IS	Δ_d (bits)	Bound (bits)	Consistent?
Medical	14.55	1.095	+1.28	2.24	Yes
Legal	15.12	1.126	+1.86	2.24	Yes
Code	15.04	0.931	−1.03	2.24	Yes
Science	14.03	0.905	−1.47	2.24	Yes

$\Delta_d = D_{\text{KL}}^{(d \to d)} - \text{mean}(D_{\text{KL}}^{(d \to \neg d)})$. Bound: $\log_2(1 + k^2/d_{\text{eff}})$ with $k = 9$, $d_{\text{eff}} = 21.7$. IS near 1.0 indicates isotropic perturbation (no domain preference); values significantly above 1.0 would indicate selective amplification.

The total output perturbation is ∼15 bits—enough to distinguish $2^{15} \approx 32{,}000$ output states. But this perturbation is overwhelmingly domain-agnostic: shaped noise changes the output distribution substantially but does so across all domains roughly equally.

The domain-specific differential Δ_d is small: +1.28 bits (medical), +1.86 bits (legal), −1.03 bits (code), −1.47 bits (science). Code and science are *information-anti-selective*: noise shaped for these domains is more informative about other domains' outputs.

An approximate bound motivated by the concentration barrier treats the domain-shaped perturbation as a signal in a Gaussian channel, with aggregate signal-to-noise scaling as k^2/d_{eff}:

$$C_{\text{domain}} \lesssim \log_2 \left(1 + \frac{k^2}{d_{\text{eff}}}\right) \tag{18}$$

With $k = 9$ and $d_{\text{eff}} = 21.7$: $\log_2(1 + 81/21.7) = \log_2(4.73) = 2.24$ bits. This is a heuristic bound, not a rigorous derivation—the Gaussian channel assumption and the k^2/d_{eff} scaling are approximate. The primary finding is empirical: all four domain-specific differentials fall within this bound, and the bound is consistent with the concentration barrier as the binding constraint.

Medical and legal show weak information selectivity (IS = 1.09–1.13); code and science are information-anti-selective (IS = 0.91–0.93). This reproduces in information space the same domain asymmetry observed in entropy space (Section 5.1).

6 Discussion

Four independent measurements—layer-resolved selectivity, spectral isotropy, the concentration barrier, and information-theoretic quantification—converge on a single conclusion: domain-selective intervention via direction-space methods is constrained by the geometry of the activation space, not by engineering choices about where or how to inject.

6.1 Convergence of the Four Measurements

Each measurement answers a different question but points to the same barrier.

The **layer-resolved selectivity** measurement (Section 5.1) asks: *where* in the forward pass does selectivity peak? Answer: layer 10, with $\bar{s} = 0.57$—a factor of six better than terminal but still below 1.0. The terminal measurement limit extends across layers.

The **spectral isotropy** measurement (Section 5.2) asks: *why* does selectivity peak there? Answer: not because of any spectral privilege. The INLP/random amplification ratio is 0.991. PCA alignment peaks at terminal layers, dissociated from selectivity. The forward pass operates as an isotropic amplifier, providing no directional leverage.

The **concentration barrier** (Section 5.3) asks: *what* bounds selectivity? Answer: geometry. Any k directions in a space with $d_{\text{eff}} \approx 20$ capture at most k/d_{eff} of the variance. The bound holds empirically at all 28 layers. Without spectral selectivity in the Jacobian, only the geometric overlap k/d_{eff} determines achievable selectivity.

The **information-theoretic quantification** (Section 5.4) asks: *how much* domain-specific information can the perturbation carry? Answer: 1–2 bits, against ∼15 bits of domain-agnostic perturbation. Consistent with the heuristic bound of 2.24 bits from the concentration barrier. The domain-specific channel is barely sufficient for a 4-way distinction and far too narrow for fine-grained control.

6.2 Implications for Activation Steering

The terminal measurement problem is not specific to shaped noise injection. Any method that identifies directions at one layer and applies interventions faces the same geometric constraint. Three findings constrain the paradigm broadly.

Orthogonality is insufficient. We distinguish between algorithmic orthogonality (INLP constructs orthogonal directions by design) and statistical near-orthogonality (random vectors

in $\mathbb{R}^{3584}$ are nearly orthogonal by concentration of measure). The former is meaningful; the latter is not. Our point is that INLP's algorithmic orthogonality does not confer special status in the forward pass—the Jacobian treats INLP directions identically to random directions (amplification ratio 0.991). In $\mathbb{R}^{3584}$, the expected absolute cosine between random unit vectors is ≈ 0.017. INLP achieves $< 10^{-9}$—but this additional precision beyond statistical near-orthogonality does not translate to functional independence. Papers that report "orthogonal concept directions" as evidence of clean separation must distinguish which kind of orthogonality they have demonstrated.

Classification accuracy does not imply intervention precision. INLP achieves near-perfect domain classification accuracy (McEntire, 2026a). But classification exploits *any* separable signal, including shared substrate that happens to correlate with the label. Intervention requires the direction to carry causal signal for the target concept specifically. The legal direction classifies legal text correctly while carrying almost no causal signal for legal-specific generation. This dissociation is predicted by the geometry: separating hyperplanes are easy to find (the blessing of dimensionality for classification) but their normal directions need not align with causal pathways (the curse of dimensionality for intervention).

The response is qualitatively nonlinear. The three-phase progression (Table 6) shows each correction produces a qualitatively different response pattern, not an incremental improvement. Methods that assume linear superposition of steering vectors—adding a "helpfulness" direction and subtracting a "toxicity" direction—face the same nonlinear mixing.

These findings do not invalidate activation steering as a technique—steering vectors produce measurable effects. They constrain the *selectivity* claims: a steering vector that increases helpfulness will also change formality, domain confidence, and other properties that share computational pathways. The bleed is a consequence of the architecture.

Negative results in science are not failures—they are boundary conditions. This paper establishes where direction-space intervention stops working and why. For practitioners: if your interpretability method relies on INLP-derived or similarly-obtained linear directions, expect collateral damage proportional to the concentration barrier. For researchers: the barrier suggests that selective intervention requires either nonlinear methods, multi-layer coordination, or fundamentally different subspace identification strategies.

6.3 The Concentration Barrier and Entanglement

The concentration barrier provides the geometric mechanism underlying the domain entanglement documented in McEntire (2026d). That paper established that INLP decomposition separates domain and structural information for classification but not for intervention. The concentration barrier explains why: in high dimensions, the geometric footprint of domain-

specific directions is necessarily small (k/d_{eff}) relative to the full activation space, and the forward pass (being spectrally isotropic) provides no mechanism to amplify it. The entanglement is not incidental—it is guaranteed by the geometry.

The pooling duality discovered in the concentration barrier analysis adds a methodological warning. Mean-pooled d_{eff} collapses to 1.0 at layers 3–25, making the barrier vacuous ($k/d_{\text{eff}} = 36$). The theorem is only informative when d_{eff} is measured under the same pooling used to compute the directions. This sensitivity is itself a finding: the effective geometry of the activation space depends on the measurement frame.

6.4 Connection to Stochastic Resonance

The inverted-U profiles in both entropy space and KL divergence confirm stochastic resonance in the neural network's output distribution. The communicative variance framework (McEntire, 2026b) predicted this: noise benefits a suboptimal system (C1), and the optimal noise level ($\sigma^* = 0.005$–0.5 depending on domain and metric) lies at the peak of the inverted-U. But the SR framework does not predict selectivity—it predicts improvement at the system level. The concentration barrier limits how much of that improvement can be domain-specific.

6.5 Open Problems

Three directions remain unexplored.

Non-linear intervention. The concentration barrier constrains *linear* direction-space methods. Non-linear interventions—modifying the computation at specific attention heads, using learned non-linear transforms of activations—could in principle circumvent the barrier by operating in the non-linear manifold where domain representations actually separate. The SAE (sparse autoencoder) approach identifies non-linear features that may have better causal specificity than linear directions.

Multi-layer coordinated injection. Our measurements inject at single layers. Coordinated injection across multiple layers, where the perturbation at each layer is designed to counteract the nonlinear mixing introduced by subsequent layers, could potentially achieve higher selectivity. This would require a differentiable model of the inter-layer Jacobian, which is computationally expensive but not infeasible.

Architecture-dependent barriers. The concentration barrier depends on how the forward pass distributes activation variance. Architectures with different mixing patterns (state-space models with linear recurrence, mixture-of-experts with sparse routing, linear attention) should show different d_{eff} profiles and different selectivity ceilings. The domain asymmetry may also be architecture-dependent.

7 Limitations

Single model family. All experiments use the Qwen 2.5 family. The concentration barrier theorem is architecturally general (it follows from the covariance structure, not the architecture), but the specific d_{eff} values, selectivity profiles, and domain asymmetry patterns may differ across architectures (e.g., Llama, Mistral, Mamba).

Four domains. Medical, legal, code, and science provide a limited test bed. The concentration barrier predicts that adding more domains (increasing k) should increase the variance fraction but not the selectivity per domain. Testing with finer-grained domain distinctions would stress the barrier more directly.

Synthetic probes. The 160 domain probes are template-generated. Natural-distribution prompts may produce different activation geometries, particularly for domains (like legal) that overlap substantially with others in natural text.

Terminal and single-layer injection only. The initial intervention used four terminal layers; the diagnostics used single-layer injection. Multi-layer coordinated injection remains untested.

Variance fraction vs. behavioral selectivity. The concentration barrier theorem (Theorem 5.1) bounds the variance fraction captured by k directions. The connection to behavioral selectivity (entropy change ratios, KL divergence) is empirically consistent but not formally derived. The numerical proximity of peak selectivity values (2.0–2.1) to the barrier bound ($\sim$1.8) is suggestive but may be coincidental—the quantities are defined differently. Closing this gap formally remains an open derivation.

Heuristic information bound. The $\log_2(1 + k^2/d_{\text{eff}})$ bound on domain-specific output perturbation is motivated by analogy to a Gaussian channel but is not rigorously derived. The empirical consistency is encouraging but does not constitute a proof.

8 Conclusion

We attempted domain-selective control of neural network outputs via direction-space intervention and failed. The failure is informative.

Shaped noise projected onto INLP domain directions produces measurable effects: 3–6% entropy reductions in target domains, 100% repetition loop escape. But cross-domain selectivity is uniformly poor. At 7B, targeting medical reduces legal entropy 5.6$\times$ more than medical entropy. Three correction attempts—subspace decomposition, scalar cancellation, R^{-1}-optimal weighting—all fail. The predicted and actual responses under the mathematically optimal linear correction are uncorrelated.

Four independent measurements diagnose the failure:

1. Layer-resolved selectivity peaks weakly at intermediate layers ($\bar{s} = 0.57$ at layer 10) and collapses toward both ends. No layer achieves mean selectivity above 1.0.
2. The forward pass is spectrally isotropic: INLP/random JVP amplification ratio = 0.991. PCA alignment is dissociated from selectivity.
3. The concentration barrier bounds the variance fraction capturable by any k directions to k/d_{eff}. With $d_{\text{eff}} \approx 20$, the bound is $\sim$1.8. Verified at all 28 layers.
4. Domain-specific output perturbation is 1–2 bits against $\sim$15 bits of domain-agnostic perturbation, consistent with a heuristic bound of 2.24 bits from the concentration barrier.

The measurements converge. The barrier is geometric—a property of high-dimensional computation—not an engineering limitation of any particular intervention method. Classification accuracy does not imply intervention precision. The concentration barrier constrains direction-space interventions using linear discriminative directions (INLP, DAS, LEACE), and likely extends to activation addition, representation engineering, ROME, and linear probe-based steering—all face the same isotropic forward pass and the same bound on domain-specific leverage. Methods that operate nonlinearly, across multiple layers simultaneously, or in non-discriminative subspaces may achieve selectivity that linear single-layer injection cannot.

The constructive implication points beyond linear methods: non-linear intervention (operating in the manifold where representations actually separate), multi-layer coordinated injection (counteracting nonlinear mixing across layers), and architecture-dependent barrier analysis (testing whether different mixing patterns produce different selectivity ceilings) are the natural next instruments.

References

Benzi, R., Sutera, A., & Vulpiani, A. (1981). The mechanism of stochastic resonance. *Journal of Physics A*, 14(11), L453.

Gammaitoni, L., Hänggi, P., Jung, P., & Marchesoni, F. (1998). Stochastic resonance. *Reviews of Modern Physics*, 70(1), 223.

Ledoux, M. (2001). *The Concentration of Measure Phenomenon*. American Mathematical Society.

McEntire, J. (2026a). Structural transfer via activation space decomposition. *Working paper*.

McEntire, J. (2026b). The source of creation is dysfunction: The generative lossy channel and five sufficient conditions for net-beneficial noise. *Working paper*.

McEntire, J. (2026c). Constellation-indexed model composition. *Working paper*.

McEntire, J. (2026d). Universal structural entanglement in transformer representations. *Zenodo*. DOI: 10.5281/zenodo.19409951.

Meng, K., Bau, D., Mitchell, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. *NeurIPS 2022*.

Qwen Team. (2025). Qwen 2.5 technical report. *arXiv preprint*.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. *ACL 2020*.

Turner, A., Thiergart, L., Udell, D., Leech, G., Mini, U., & MacDiarmid, M. (2023). Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A., Goel, S., Li, N., Lin, Z., Forsyth, M., Bumpus, R., Huang, J., & Steinhardt, J. (2023). Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.