Shaped Noise Injection at Inference Time:

Domain Precision, Loop Breaking, and the Terminal Measurement Limit

Jeremy McEntire[*]

Abstract

We investigate whether noise shaped to domain-discriminative directions in activation space can control inference-time output distributions in large language models. Using iterative nullspace projection (INLP) directions as a domain basis, we inject Gaussian noise projected onto target subspaces in the final transformer layers of Qwen 2.5 models at 0.5B, 1.5B, 3B, and 7B parameters. Three applications are tested: domain entropy reduction, repetition loop breaking, and soft guardrails.

We report three findings. First, shaped noise achieves modest domain-specific entropy reductions (up to 6.1% for legal at 7B) and breaks 100% of repetition loops at both 3B and 7B, outperforming temperature scaling and matching repetition penalty on escape rate while achieving near-perfect token uniqueness. Second, cross-domain selectivity is uniformly poor: targeting one domain produces comparable or larger entropy changes in non-target domains. At 7B, injecting along the medical direction reduces legal entropy by 10.7% while reducing medical entropy by only 1.9%. Third, and most significantly, all attempts to correct this cross-domain bleed fail. Scalar interference cancellation worsens selectivity at every lambda value tested. The mathematically optimal correction—computed from the inverse of the empirical response matrix—produces effects uncorrelated with predictions (medical target: predicted $[-1.0, 0, 0, 0]$, actual $[+0.1, -5.2, +1.2, +7.4]$), definitively establishing that the response is nonlinear.

We identify this as the *terminal measurement problem*: the cross-domain response matrix characterizes the system's output but cannot invert the nonlinear mixing that occurs at intermediate transformer layers. In $d = 3584$ dimensions at 7B, the concentration of measure guarantees that INLP direction orthogonality is vacuous—nearly all vectors are nearly orthogonal regardless of functional relationship. The bleed lives in the forward-pass topology, not in the input geometry. Correction must operate at the layer where computational paths diverge, not at the projection layer. This result

[*] Working Paper. Correspondence: jmc@cageandmirror.com

constrains all direction-space intervention methods and motivates layer-resolved noise injection as the natural next instrument.

Keywords: stochastic resonance, shaped noise, inference-time control, activation geometry, cross-domain bleed, terminal measurement, concentration of measure, INLP

1 Introduction

Large language models encode domain knowledge in learned activation subspaces. Prior work has established that these subspaces are identifiable via iterative nullspace projection (Ravfogel et al., 2020), that they carry domain-discriminative signal separable from general linguistic structure (McEntire, 2026c), and that stochastic resonance—the phenomenon where noise improves signal detection in nonlinear systems—operates on these subspaces when specific conditions are met (McEntire, 2026e). This paper asks the natural next question: can noise shaped to these subspaces *control* the model's output distribution at inference time?

The motivation is practical. Larger models are more capable but also more uncertain on domain-specific prompts. Our scale-entropy measurements (Section 3) show that mean output entropy decreases monotonically from 2.38 bits at 0.5B parameters to 1.48 bits at 7B—larger models are more confident, not less. But this aggregate confidence masks domain-specific uncertainty that a specialist model would not exhibit. The gap between generalist and specialist entropy on domain probes is exactly the C1 suboptimality measure identified in McEntire (2026e) as the precondition for stochastic resonance benefit.

If shaped noise can collapse this uncertainty selectively—reducing entropy in a target domain while leaving other domains unaffected—the implications are substantial. Domain precision without fine-tuning. Repetition loop breaking without quality degradation. Continuous guardrails that operate on activation geometry rather than output filtering. One mechanism, three applications.

We test this thesis across ten experimental phases on Qwen 2.5 models from 0.5B to 7B parameters, using INLP-discovered domain directions from McEntire (2026c) as the noise shaping basis.[1] The results are partially positive and partially negative, and the negative result is the more important finding.

The positive results: shaped noise achieves modest entropy reductions (3–6% in target domains at 7B) and breaks 100% of repetition loops at both 3B and 7B, with near-perfect token uniqueness (0.99+). These demonstrate that the mechanism works—shaped activation perturbation does influence output distributions in the predicted direction.

The negative result: cross-domain selectivity is fundamentally limited. When we target one domain, non-target domains respond comparably or more strongly. All attempts to correct this—scalar cancellation (Section 6.2), subspace decomposition (Section 6.1), and optimal

[1] This is Paper VIII in a series. Papers I–VII establish the empirical and theoretical foundations: activation fingerprint convergence (McEntire, 2026a), constellation-indexed composition (McEntire, 2026b), structural transfer via INLP (McEntire, 2026c), capability manifold surveillance (McEntire, 2026d), the communicative variance theorem (McEntire, 2026e), and scale-dependent noise tolerance in social cognition (McEntire, 2026f). Companion papers are available from the author; those on arXiv and SSRN are cited with venue.

linear correction via matrix inversion (Section 6.3)—fail. The failure is not incremental. The empirical response matrix R is invertible (determinant $= -84.5$, condition number $= 21.2$), the linear algebra is clean, and the optimal weight vectors are computable. They simply do not produce the predicted effects when applied. The predicted-vs-actual correlation is approximately zero.

We identify the root cause as the *terminal measurement problem*: the response matrix R is a terminal measurement of a process that occurs at intermediate layers. It characterizes the system's input-output mapping but cannot invert the nonlinear transformations that generate cross-domain bleed during the forward pass. In the high-dimensional activation space ($d = 3584$ at 7B), the concentration of measure (Vershynin, 2018) guarantees that geometric orthogonality of directions is uninformative about functional overlap—nearly all pairs of vectors in $\mathbb{R}^{3584}$ are nearly orthogonal regardless of their computational relationship.

This result constrains the entire class of direction-space intervention methods. Any technique that operates by projecting perturbations onto directions identified at a single layer—whether for steering, editing, or control—faces the same terminal measurement limit. The functional mixing that determines cross-domain bleed occurs at intermediate layers, and no terminal-layer correction can invert it. The constructive implication is a specific next instrument: layer-resolved noise injection, which is the outside-in equivalent of activation patching (Meng et al., 2022) applied at inference time without access to model weights.

2 Methods

2.1 Models and Probes

All experiments use the Qwen 2.5 model family (Qwen Team, 2025) at four scales: 0.5B, 1.5B, 3B, and 7B parameters. Models are loaded in float16 precision on A100 GPUs via HuggingFace Transformers.

We use two probe sets. The *domain probes* comprise 160 prompts (4 domains × 4 syntactic shapes × 10 prompts) from McEntire (2026c), covering medical, legal, code, and science domains. The *cross-domain probes* comprise 150 prompts (25 per domain pair, 6 pairs) from McEntire (2026b), designed to test compositional domain knowledge. An additional 40 general-knowledge probes serve as controls.

2.2 Shaped Noise Injection

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

$$\mathbf{h}' = \mathbf{h} + \sigma \cdot P_S \, \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I_d) \tag{1}$$

where σ is the noise amplitude, $\boldsymbol{\epsilon}$ is isotropic Gaussian noise, and P_S is the projection matrix onto the target subspace S:

$$P_S = D_S^\top D_S \tag{2}$$

where $D_S \in \mathbb{R}^{k \times d}$ is the matrix of k unit-norm domain directions spanning S.

Three injection modes are tested:

- **Positive**: $P_S = D_{\text{target}}^\top D_{\text{target}}$ (boost target domain)
- **Negative**: $P_S = D_{\text{others}}^\top D_{\text{others}}$ (perturb competing domains)
- **Both**: $P_{\text{target}} + P_{\text{others}}$ (combined)

2.3 Domain Basis: INLP Directions

The domain directions D come from iterative nullspace projection (INLP) (Ravfogel et al., 2020) applied to activation differences between domain-specific and general text, as computed in McEntire (2026c). 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 $< 10^{-9}$.

2.4 Metrics

For each probe, we generate 64 tokens with greedy decoding (`output_scores=True`) and compute:

1. **Mean token entropy**: $\bar{H} = \frac{1}{T} \sum_{t=1}^{T} H(p_t)$ where $H(p_t) = -\sum_v p_t(v) \log_2 p_t(v)$
2. **Entropy delta**: $\Delta H = \bar{H}_{\text{injected}} - \bar{H}_{\text{baseline}}$
3. **Entropy reduction ratio**: $\Delta H / \bar{H}_{\text{baseline}}$
4. **Domain selectivity**: $(\Delta H_{\text{target}} - \overline{|\Delta H_{\text{others}}|}) / \max(\overline{|\Delta H_{\text{others}}|}, 0.01)$ when $\Delta H_{\text{target}} < 0$

For loop breaking (Section 5), we additionally measure escape rate, token uniqueness ratio, and output entropy.

3 Scale-Entropy Baselines

Before injection, we establish how output entropy varies with model scale.

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

Mean token entropy decreases monotonically with scale: 2.38 bits at 0.5B to 1.48 bits at 7B, a 37.7% reduction. The decrease is consistent across domains, with code showing the highest entropy at every scale and medical the lowest. Standard deviation also decreases (1.05 to 0.63), indicating that larger models produce more uniformly confident distributions.

This is not counterintuitive once framed correctly. Training is a law-of-large-numbers process: each gradient step is a sample from the loss landscape, and with sufficient samples the parameter estimates converge. A 7B model trained on trillions of tokens has seen orders of magnitude more gradient signal than a 0.5B model, and the central limit theorem guarantees that its posterior over next-token distributions concentrates more tightly. On domain-specific probes—where the training data contains clear signal—the model has effectively performed the trillion coin flips and read the result. Inference reads off a converged estimate. The "probability cloud flattening" posited for large models must occur on genuinely ambiguous prompts where the training signal itself is mixed, not on domain probes where the answer is well-determined by the data.

The practical consequence is that the C1 suboptimality condition from McEntire (2026e)—the gap between current and optimal performance—is small on these probes. The baseline is already good. Any SR benefit must come from residual uncertainty that shaped noise can resolve.

4 Domain Precision via Shaped Noise

4.1 Stochastic Resonance Sweep

We sweep 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 1 reports the best entropy reduction per domain at 7B.

At 3B, effects are smaller: the best result is code/both at -2.7% ($\sigma = 0.01$). Legal/both achieves only -0.8%. The remaining domain-mode combinations show no reduction or slight increases.

Three patterns are notable. First, optimal σ values cluster in the 0.005–0.02 range, consistent with the stochastic resonance inverted-U where small noise helps and larger noise rapidly destroys signal. Second, the best mode varies by domain (positive for medical, both

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 1: Best stochastic resonance results at 7B. Legal/both achieves the strongest entropy reduction. Science shows no improvement at any noise level.

for legal, negative for code), suggesting domain-specific subspace geometry matters. Third, science is completely unresponsive to noise injection at 7B, indicating that the 7B model's science representations are already at or near the entropy floor achievable within the INLP subspace.

4.2 Cross-Domain Response Matrix

The critical question is not whether shaped noise reduces entropy but whether it does so *selectively*. Table 2 reports the full 4×4 response matrix R at both scales, where entry R_{ij} is the percentage entropy change in domain j when noise is injected along domain i's directions.

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 2: Cross-domain response matrices. Asterisks (*) mark diagonal (target) entries. At 3B, noise predominantly increases entropy across all domains. At 7B, medical targeting produces a -10.7% entropy change in the legal domain—5.6× the self-effect. Selectivity is poor at both scales.

The 7B heatmap 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. Code targeting has a self-effect of −1.2% with +3.0% bleed into science.

At 3B, the picture is different but equally poor: noise predominantly *increases* entropy across all domains, with the diagonal entries (self-effects) often smaller than off-diagonal entries. Science targeting at 3B produces +12.3% self-entropy increase, indicating that the model's science representations are being destabilized rather than refined.

The selectivity index is negative for legal and code at both scales. Only medical at 7B achieves a positive selectivity (0.63), and even there the off-diagonal response dominates.

5 Loop Breaking

Independent of domain selectivity, shaped noise has a straightforward application: breaking degenerate attractor states (repetition loops). We test 20 prompts known to induce repetitive generation and compare four methods.

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. Repetition penalty is competitive on escape rate but produces lower uniqueness. Temperature scaling fails to escape half of loops.

Shaped SR breaks every loop at both scales with token uniqueness ≥ 0.99, meaning nearly every generated token is distinct. Baseline generation fails to escape 70–80% of loops. Temperature scaling ($T = 1.5$) improves escape rate to 55–70% but with low uniqueness (0.41–0.44), indicating partial but not complete loop breaking. Repetition penalty ($\alpha = 1.2$) achieves 95–100% escape with good uniqueness (0.86–0.88).

The trade-off is entropy. Shaped SR produces output entropy of 8.4–10.0 bits, 3–5× higher than repetition penalty (2.9–3.1 bits) and 9–10× higher than baseline (0.9–1.1 bits). This reflects over-perturbation: the noise is strong enough to break any attractor but also flattens the distribution substantially. For applications where loop escape is the sole objective, repetition penalty achieves comparable results with less distributional disruption.

The result confirms that shaped noise influences the generation process in the predicted direction—it destabilizes degenerate modes. The mechanism works. The question that follows is whether the effect can be made selective, which the remainder of the paper addresses.

6 The Selectivity Barrier

The cross-domain response matrix (Table 2) shows that shaped noise lacks selectivity. We undertake three progressively stronger attempts to correct this.

6.1 Subspace Decomposition

One hypothesis for the bleed is that INLP directions, despite being orthogonal by construction, share variance attributable to general linguistic structure ("register" effects). If so, projecting out the register subspace should yield purer domain directions with less cross-domain response.

We compute register directions via SVD of the full direction matrix $D \in \mathbb{R}^{k\times d}$ and project each domain's INLP directions onto the nullspace of the top-8 singular vectors. At 7B, the register singular values show a 4.7× spectral gap between the first and second components ($\sigma_1 = 684.5$, $\sigma_2 = 282.0$), confirming a dominant shared component.

However, the residual directions are nearly unchanged. Mean cosine similarity between raw and register-projected directions exceeds 0.97 at 7B across all domains (medical: 0.985, legal: 0.971, code: 0.984, science: 0.987). The register projection removes negligible variance because the INLP directions are already orthogonal to each other and nearly orthogonal to the register subspace.

Meanwhile, contrastive directions (computed from activation differences without INLP's nullspace constraint) show *higher* cross-domain similarity: mean pairwise cosine ≈ 0.25, with maximum cosines reaching 0.82–0.87. This confirms that the cross-domain similarity in the model's representations is genuine, not an artifact of the direction-finding method. The INLP orthogonality is a property of the algorithm's construction, not of the model's geometry. When that construction constraint is relaxed, the shared structure reappears.

Remark 6.1. The cross-domain bleed observed in Table 2 is a property of the model's forward pass, not of the injected directions. Refining the directions does not change the bleed.

6.2 Scalar Interference Cancellation

If the bleed cannot be removed from the directions, perhaps it can be cancelled in the output. The Phase 4 response matrix R provides the empirical mapping from injection direction to output effect. For each target domain, we estimate a scalar cancellation coefficient λ_{ij} for each non-target domain j:

$$\lambda_{ij} = \frac{R_{ij}}{R_{jj}} \tag{3}$$

and construct a cancelled direction:

$$\mathbf{d}_{\text{cancel}} = \bar{\mathbf{d}}_i - \sum_{j \neq i} \lambda_{ij} \cdot \bar{\mathbf{d}}_j \tag{4}$$

where $\bar{\mathbf{d}}_i$ is the mean of domain i's INLP directions.

We sweep a multiplier $m \in \{0, 0.25, 0.5, 1.0, 2.0, 4.0\}$ on the λ coefficients to find the best operating point. Table 4 reports results at 7B.

Domain	**Best** m	**Self** ΔH	**Mean |Bleed|**	**Sel.**
Medical	4.00	-7.2%	15.3%	-0.47
Legal	1.00	-0.3%	6.3%	-0.05
Code	0.00	$+0.5\%$	5.6%	-0.46
Science	0.25	-0.2%	4.8%	-0.04

Table 4: Scalar interference cancellation at 7B. All best selectivities are negative, meaning bleed exceeds self-effect at every operating point. Code's best multiplier is 0.00 (no cancellation at all).

Every domain's best selectivity is negative. Medical targeting at $m = 4.0$ achieves a -7.2% self-effect but generates 15.3% mean bleed—the cure is worse than the disease. Code's best operating point is $m = 0$ (no correction at all), meaning any cancellation makes things worse. The scalar approach fails because the cancellation coefficients λ_{ij} are computed from the same response matrix they attempt to correct, and the system is nonlinear: scaling λ does not scale the response proportionally.

6.3 Optimal Linear Cancellation

The scalar approach constrains each domain pair independently. A stronger test is the globally optimal linear combination: given the full response matrix $R \in \mathbb{R}^{4\times4}$, find the weight vector $\mathbf{w}_k$ such that the predicted response $R^\top \mathbf{w}_k = -\mathbf{e}_k$ (unit reduction in target domain k, zero effect elsewhere).

If R is invertible, the solution is $\mathbf{w}_k = -R^{-\top}\mathbf{e}_k$. We verify that R is indeed invertible:

$$\begin{aligned} \det(R) &= -84.5 \quad (5) \\ \operatorname{rank}(R) &= 4 \quad (6) \\ \kappa(R) &= 21.2 \quad (7) \end{aligned}$$

The condition number is moderate, so numerical stability is not a concern. The optimal weight vectors exist and are computable.

We construct a weighted noise injector that combines all domain directions with the R^{-1}-derived weights:

$$\mathbf{d}_{\text{combined}} = \sum_{j=1}^{4} w_j \cdot \bar{\mathbf{d}}_j, \quad P = \hat{\mathbf{d}}_{\text{combined}}\hat{\mathbf{d}}_{\text{combined}}^{\top} \tag{8}$$

and inject noise through this projection at σ selected by a preliminary sweep.

Phase 10b: 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: Heatmap under R^{-1} optimal weight vectors. Medical self-effect has reversed sign (predicted −1.0%, actual +0.1%). Legal bleed persists at −5.2% despite the prediction of zero. The mathematically optimal linear correction fails completely.

6.3.1 The Linearity Test

The R^{-1} weight vectors provide a direct test of whether the system's response is linear. Under linearity, the predicted response for the medical target is:

$$R^{\top}\mathbf{w}_{\text{med}} = [-1.0,\ 0.0,\ 0.0,\ 0.0] \tag{9}$$

The actual measured response is:

$$\text{Actual} = [+0.1,\ -5.2,\ +1.2,\ +7.4] \tag{10}$$

The self-effect has the wrong sign. The legal bleed (−5.2%) persists at half its original magnitude despite the 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 R is invertible, the optimal weight vectors are computable, and they produce effects that bear no resemblance to the predictions. The system is fundamentally nonlinear: the mapping from injected direction to output entropy cannot be described by any linear operator.

	Phase 4		**Phase 10**		**Phase 10b**	
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. Phase 4: raw INLP directions. Phase 10: scalar cancellation. Phase 10b: R^{-1} optimal. Each correction changes the response pattern entirely rather than improving selectivity, confirming nonlinearity.

Table 6 shows the progression across all three approaches. Each correction attempt does not incrementally improve the heatmap—it produces a qualitatively different pattern. Raw injection (Phase 4) shows modest self-effects with substantial bleed. Scalar cancellation (Phase 10) amplifies both self-effect and bleed. Optimal linear correction (Phase 10b) produces near-zero self-effects with persistent bleed. The system responds to each intervention differently, not as a linear function of the input.

7 The Terminal Measurement Problem

The results of Sections 4.2–6.3 establish an empirical fact: no linear combination of INLP directions injected at the final layers produces selective domain-specific entropy changes. This section provides the theoretical explanation.

7.1 Concentration of Measure in Activation Space

The INLP directions at 7B live in $\mathbb{R}^{3584}$. In high dimensions, the concentration of measure phenomenon (Vershynin, 2018; Ledoux, 2001) guarantees that geometric relationships become degenerate. Specifically, for uniformly random unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$:

$$\mathbb{E}[|\langle \mathbf{u}, \mathbf{v} \rangle|] = O(d^{-1/2}) \tag{11}$$

At $d = 3584$, this is approximately 0.017. Nearly all pairs of vectors are nearly orthogonal, regardless of their functional relationship.

The INLP algorithm produces directions with pairwise cosine similarity $< 10^{-9}$—machine-zero orthogonality. But this is not informative. In $\mathbb{R}^{3584}$, any set of 36 vectors (the full INLP basis at 7B) would be nearly orthogonal even if drawn at random. The orthogonality is a consequence of the dimensionality, not evidence of functional independence.

The volume of the unit ball B_d in $\mathbb{R}^d$ scales as:

$$\mathrm{Vol}(B_d) = \frac{\pi^{d/2}}{\Gamma(d/2 + 1)} \tag{12}$$

which approaches zero super-exponentially for large d. The mass of a high-dimensional Gaussian concentrates in a thin shell at radius $\sqrt{d}$. Directions that are geometrically orthogonal can activate overlapping computational pathways through the transformer's nonlinear layers, because the distance between their projections through nonlinear transformations is not constrained by their angle in the input space.

7.2 Terminal Characterization vs. Intermediate Correction

The response matrix R from Phase 4 is a *terminal measurement*: it records the output-layer consequences of perturbations injected at the final layers. But the cross-domain bleed is *generated* at intermediate layers, where the transformer's attention mechanism and feedforward networks mix information across the residual stream.

Consider the forward pass as a composition of nonlinear maps $f = f_L \circ f_{L-1} \circ \cdots \circ f_1$. Noise injected at layer ℓ propagates through $f_L \circ \cdots \circ f_{\ell+1}$ before reaching the output. Even if the injected perturbation is perfectly aligned with one domain's direction at layer ℓ, the subsequent nonlinear transformations can project it onto other domains' representations.

The response matrix captures:

$$R_{ij} = \left.\frac{\partial \bar{H}_j}{\partial \sigma_i}\right|_{\sigma=0^+} + O(\sigma^2) \tag{13}$$

where $\bar{H}_j$ is the mean entropy on domain j probes and σ_i is the noise amplitude along domain i's direction. Under linearity, R^{-1} would provide the correction. But the forward pass is not linear in the injected perturbation—the $O(\sigma^2)$ terms are not negligible, and more critically, the Jacobian $\partial f / \partial \mathbf{h}_\ell$ at the injection point depends on the input, making the mapping input-dependent even at first order.

Remark 7.1. The response matrix R is the Jacobian of the system evaluated at a single operating point. Under sufficient smoothness, R^{-1} provides a valid local correction in the limit $\sigma \to 0^+$. The experimental evidence shows that this linearization domain is effectively empty: even at $\sigma = 0.005$ (the smallest nonzero value tested), the response deviates substantially from the linear prediction. The nonlinearity is not a higher-order correction—it dominates at every testable amplitude.

7.3 The Legal Direction as Shared Substrate

The 7B legal direction provides the sharpest illustration of why direction-space intervention fails. Three converging measurements identify it as shared substrate rather than domain-specific signal.

Near-zero self-effect. Legal targeting produces a self-effect of only -0.4% (Table 2) with $+3.8\%$ bleed into medical. The legal INLP direction moves other domains more than it moves legal itself.

Zero optimal self-weight. In the R^{-1} optimization, the optimal weight vector for the legal target assigns weight $w_{\text{legal}} = 0.000$ to the legal direction. 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 is to combine medical, code, and science directions only—legal is redundant for its own targeting.

Asymmetric lambda coefficients. The scalar cancellation analysis (Section 6.2) estimates the bleed ratio $\lambda_{\text{legal}\to\text{medical}} = -10.35$. The negative sign means the legal direction's effect on medical entropy runs *opposite* to the legal direction's effect on legal entropy. This is not attenuation—it is a qualitatively different response, consistent with the legal direction activating a shared processing pathway that medical and legal representations both traverse but in different functional modes.

The interpretation is that legal language draws on the same formal register, structured argumentation, and technical vocabulary that other domains use. Legal text is medical reasoning applied to statutes, scientific methodology applied to evidence standards, code-like precision applied to contractual language. The INLP algorithm finds the direction that best *separates* legal from non-legal in a linear classifier, but in $\mathbb{R}^{3584}$ this separating hyperplane is dominated by the shared substrate rather than any domain-specific residual. What INLP calls "the legal direction" is more accurately the direction of maximum shared formality.

This has a concrete implication for the broader activation steering literature: if a linear probe achieves high classification accuracy on a concept, it does not follow that the probe direction is a viable intervention target. 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, which the legal case shows is not guaranteed and may not even be the common case in high dimensions.

8 Soft Guardrails: A Constructive Null

We test whether shaped noise can steer generation away from a "forbidden" subspace. Using 8 prompts designed to elicit responses in a target domain, we compute the top-8 SVD components of their activation fingerprints and inject repulsive noise (negative projection) at $\sigma \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$.

The result is a complete null at both 3B and 7B. Generated text is identical across all σ values. Entropy measurements are constant to three decimal places. The SVD-derived forbidden subspace has no detectable influence on generation.

This null is constructive. It confirms that ad-hoc subspace construction (SVD from a small number of prompts) does not produce directions with functional significance for the model's generation process. The INLP directions used in Sections 4–6 do produce measurable effects precisely because they are *discriminative*—trained to separate domains in activation space. The guardrail null demonstrates that the mechanism requires learned discriminative directions, not arbitrary low-rank approximations.

Combined with the selectivity barrier, this constrains the guardrail application: shaped noise can influence generation when injected along discriminative directions, but the cross-domain bleed prevents targeted steering. Effective soft guardrails would require either (a) discriminative directions with better functional selectivity, or (b) injection at intermediate layers where domain-specific representations are more separated.

9 Discussion

9.1 What Works and What Does Not

Shaped noise injection is an effective mechanism for breaking degenerate attractors and a weak mechanism for domain-specific entropy control. Table 7 summarizes the findings.

Application	**Effectiveness**	**Selective?**
Loop breaking	100% escape, 0.99 unique	N/A
Domain precision	3–6% entropy reduction	No
Interference cancel.	Worsens selectivity	No
Soft guardrails	No measurable effect	N/A

Table 7: Summary of shaped noise injection applications.

Loop breaking succeeds because it does not require selectivity. The goal is to destabilize any attractor, and undirected noise along domain directions suffices. Domain precision partially succeeds in that it reduces target-domain entropy, but the cross-domain effects are comparable or larger, making the reduction non-selective. Interference cancellation fails because the system's response is nonlinear. Soft guardrails fail because ad-hoc subspace construction does not produce functionally meaningful directions.

9.2 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 at that layer faces the same constraint. Activation addition (Turner et al., 2023), representation engineering (Zou et al., 2023), and linear probe-based interventions all rely on the assumption that directions identified via linear methods have approximately linear effects on output behavior.

Our results suggest this assumption breaks down when *selectivity* is required. Three specific findings constrain the activation steering paradigm:

Orthogonality is vacuous. In $\mathbb{R}^{3584}$, the expected absolute cosine between random unit vectors is ≈ 0.017. INLP directions achieve $< 10^{-9}$—but this is not meaningfully different from random in terms of functional independence. The concentration of measure guarantees that any reasonable number of directions will be nearly orthogonal regardless of their computational relationship. Papers that report "orthogonal concept directions" as evidence of clean separation are reporting a geometric tautology, not an empirical finding. The relevant question is not whether directions are orthogonal in activation space but whether perturbations along those directions produce orthogonal effects in the output—and our Phase 10b results show they do not.

Classification accuracy does not imply intervention precision. The INLP directions achieve near-perfect domain classification accuracy (McEntire, 2026c). But as Section 7.3 demonstrates, high classification accuracy can be achieved via shared substrate rather than domain-specific signal. The legal direction classifies legal text correctly while carrying almost no causal signal for legal-specific generation. This dissociation between classification and intervention is predicted by the high-dimensional geometry: separating hyperplanes are easy to find (the blessing of dimensionality for classification) but the directions normal to those hyperplanes need not align with the causal pathways that generate domain-specific behavior (the curse of dimensionality for intervention).

The response is qualitatively nonlinear. The three-phase progression (Table 6) shows that each correction attempt produces a qualitatively different response pattern rather than an incremental improvement. This is not a case where linear methods are approximately correct and need refinement. The predicted and actual response vectors are uncorrelated. Methods that assume linear superposition of steering vectors—adding a "helpfulness" direction and subtracting a "toxicity" direction—face the same nonlinear mixing we document here. The net effect of a composite intervention cannot be predicted from the individual effects.

These findings do not invalidate activation steering as a technique—steering vectors produce measurable effects, as our domain precision results confirm. They constrain the *precision* claims: a steering vector that increases helpfulness will also change formality, domain confidence, and other properties that share computational pathways through intermediate layers. The bleed is not a bug to be fixed; it is a consequence of the architecture.

9.3 Connection to Communicative Variance

The communicative variance framework (McEntire, 2026e) predicts that stochastic resonance benefit is gated by C1 suboptimality: noise helps only when the current operating point is suboptimal. Our Phase 1 baselines show that the 7B model's entropy on domain probes is already low (1.48 bits mean), leaving little room for SR benefit. The modest 3–6% reductions at 7B are consistent with small C1 gaps.

The framework also predicts that SR should show an inverted-U relationship with noise amplitude. We observe this across all effective domain-mode combinations, with optimal σ in the 0.005–0.02 range and rapid degradation above $\sigma = 0.05$. The prediction holds locally (within-domain entropy) even though the cross-domain response prevents selective application.

9.4 Toward Layer-Resolved Injection

The terminal measurement problem points to a specific next instrument: injecting noise at each transformer layer individually and measuring the per-layer, per-domain entropy response. This produces a *layer-resolved response tensor* $R_{ij}^{(\ell)}$ that maps the nonlinear mixing across the forward pass.

This is the outside-in equivalent of activation patching (Meng et al., 2022): instead of ablating intermediate activations to identify causal structure, we inject calibrated noise and measure the downstream effect. The advantage is that it operates at inference time without access to model internals beyond forward hooks. The layer where domain-specific paths diverge—where $R^{(\ell)}$ transitions from non-selective to selective—identifies the computational bottleneck that the terminal measurement cannot resolve.

10 Limitations

Model family. All experiments use the Qwen 2.5 family. The terminal measurement problem is architecturally general (it follows from the nonlinearity of transformer layers), but the specific bleed patterns may differ across architectures.

Domain basis. INLP directions capture linear separability, not domain identity. Other direction-finding methods (DAS, sparse probing, SAE features) may yield directions with different functional properties, though they face the same high-dimensional geometry.

Injection layers. We inject at the final four layers. Injecting deeper may produce different bleed patterns, but would also amplify through more nonlinear transformations. The layer-resolved experiment (Section 9.4) would address this systematically.

Probe coverage. 160 domain probes and 150 cross-domain probes provide reasonable coverage but may not represent the full distribution of domain-specific inputs.

Sigma selection. Phase 10b used sigma values from a preliminary sweep; the selectivity formula favored sigma values that were not always empirically optimal. This does not affect the linearity conclusion, since the nonlinearity is fundamental, not a function of sigma.

11 Conclusion

We tested shaped noise injection as an inference-time mechanism for controlling output distributions in large language models. The mechanism works: noise projected onto domain-discriminative INLP directions reduces target-domain entropy by 3–6% and breaks 100% of repetition loops. But the mechanism is not selective: cross-domain bleed is comparable to or larger than the target effect, and no linear correction—scalar, per-domain, or globally optimal—can improve selectivity.

The root cause is the terminal measurement problem. The response matrix characterizes the system's output-layer behavior but cannot invert the nonlinear mixing at intermediate layers. In $d = 3584$ dimensions, direction-space orthogonality is geometrically guaranteed and functionally meaningless. The bleed lives in the forward-pass topology.

This result establishes a fundamental constraint on direction-space interventions: any method that identifies and perturbs along directions at a single layer must contend with the nonlinear propagation that follows. The constructive implication is that layer-resolved injection—mapping the nonlinear mixing across layers—is the natural next instrument for achieving the selectivity that terminal-layer intervention cannot.

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.

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. *ICLR 2020.*

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

McEntire, J. (2026a). Training once is enough: Activation fingerprint convergence. *SSRN Preprint.* Paper II in this series.

McEntire, J. (2026b). Constellation-indexed model composition. *Companion paper (Paper III in this series).*

McEntire, J. (2026c). Structural transfer via activation space decomposition. *Companion paper (Paper IV in this series).*

McEntire, J. (2026d). Capability manifold surveillance. *Companion paper (Paper V in this series).*

McEntire, J. (2026e). The source of creation is dysfunction: The generative lossy channel and five sufficient conditions for net-beneficial noise. *Companion paper (Paper VI in this series).*

McEntire, J. (2026f). GenAI is socially awkward: Scale-dependent noise tolerance in social cognition. *Companion paper (Paper VII in this series).*

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*.