Causal Basis Discovery for Domain-Selective Noise Injection

Jeremy McEntire[1]

March 2026

Abstract

Papers IV and IX–XI established INLP (Iterative Nullspace Projection) as the default basis for domain-selective noise injection, but INLP is a *classification-optimized* basis: it finds directions that maximally separate domain distributions in activation space, not directions that causally mediate domain-specific computation. This paper tests whether causally-derived bases produce superior selectivity. We extract three alternative bases at the selectivity-peak layer (layer 10, Qwen 2.5-7B): activation patching directions, contrastive activation directions, and gradient-aligned directions. Each is evaluated against INLP using the 4×4 entropy response matrix from Paper IX.

INLP achieves mean selectivity +0.618 with diagonal dominance +1.818. All three causal bases are *anti-selective*: patching (−0.630, dominance −4.073), contrastive (−0.777, dominance −2.159), and gradient (−0.116, dominance −1.157). Pairwise basis overlaps are near-zero (0.011–0.016 mean cosine), confirming that all four bases occupy essentially orthogonal subspaces in $\mathbb{R}^{3584}$. The result establishes a classification-intervention dissociation: directions that best *classify* domains also produce the best *interventions*, while causally-derived directions are actively worse than random in this geometry.

1 Introduction

INLP finds the directions along which domain distributions are maximally linearly separable. These directions achieve >97.5% classification accuracy (Paper IV) and serve as the injection basis for all noise experiments in Papers VII–XII. But "separates domains" is not the same as "mediates domain computation." A razor-thin decision boundary can be a perfect classifier while lying orthogonal to the actual causal pathway through which the model processes domain-specific information.

This paper tests whether the causal pathway produces better noise injection targets. The motivation is a change-of-basis argument: INLP provides one basis for the domain-specific subspace, but other bases spanning different directions might yield higher selectivity if those directions better align with the model's internal computation.

We extract three causally-motivated bases:

[1]Correspondence: jmc@cageandmirror.com

1. **Activation patching**: Directions along which swapping activations between domain and non-domain contexts produces maximal output change. These capture the causal influence of representation on output.

2. **Contrastive activations**: Directions of maximal mean activation difference between domain and non-domain inputs. These capture the statistical signature of domain processing.

3. **Gradient-aligned**: Directions of steepest gradient of output entropy with respect to hidden activations on domain-specific inputs. These capture the sensitivity of the model's output to perturbation along each direction.

Each basis is compared to INLP using the entropy response matrix methodology from Paper IX, applied at the selectivity-peak layer 10 with $\sigma = 0.2$.

2 Methods

2.1 Model and data

All experiments use Qwen 2.5-7B (28 transformer layers, hidden dimension 3584) at layer 10, identified as the selectivity peak in Paper IX. Domain probes: 160 prompts across four domains (medical, legal, code, science; 40 per domain). INLP directions: 36 unit vectors (9 per domain) from Paper IV, computed via iterative nullspace projection on terminal-layer activations.

2.2 Basis extraction

Activation patching. For each domain, we select the 10 highest-response probes (strong domain signal) and 10 lowest (weak domain signal). At layer 10, we capture the last-token hidden state $\mathbf{h}_{\text{strong}}$ for each strong-signal probe and the mean hidden state $\bar{\mathbf{h}}_{\text{weak}}$ across the weak-signal set. The patching direction is the normalized difference:

$$\mathbf{v}_{\text{patch}} = \frac{\mathbf{h}_{\text{strong}} - \bar{\mathbf{h}}_{\text{weak}}}{\|\mathbf{h}_{\text{strong}} - \bar{\mathbf{h}}_{\text{weak}}\|}$$

We extract 9 such directions per domain (using the top-9 strong probes), yielding 36 patching directions total.

Contrastive activations. For each domain, we compute the mean activation vector across all 40 domain probes and the mean across all 120 non-domain probes at layer 10. The

contrastive direction is the normalized difference of these means. To obtain 9 directions per domain, we apply PCA to the centered within-domain activations and take the top-9 principal components of the domain-minus-global residual.

Gradient-aligned. For each domain probe, we compute $\partial H / \partial \mathbf{h}^{(10)}$, where H is the output token entropy and $\mathbf{h}^{(10)}$ is the layer-10 hidden state. This requires a backward pass through layers 10–27 plus the language model head. The per-probe gradient is normalized; we then apply SVD across all domain probes to extract the top-9 gradient-aligned directions per domain.

2.3 Evaluation protocol

Each basis is evaluated using the same protocol as Paper IX: inject shaped noise along the basis directions at layer 10 ($\sigma = 0.2$, 9 directions per target domain), measure entropy change across all 160 probes, and compute the 4×4 response matrix where entry R_{ij} is the mean entropy change on domain-j probes when noise is shaped for domain i.

Metrics:

- **Diagonal dominance**: $\bar{R}_{ii} - \bar{R}_{i \neq j}$ (positive = selective)
- **Mean selectivity**: Average per-domain selectivity (diagonal minus mean off-diagonal, per row)
- **Per-domain selectivity**: Row-level selectivity for each target domain

2.4 Basis overlap analysis

For each pair of bases, we compute the mean absolute cosine similarity between all direction pairs. In $\mathbb{R}^{3584}$, two random unit vectors have expected cosine $\sim 1/\sqrt{3584} \approx 0.017$. Overlaps near this floor indicate the bases span essentially orthogonal subspaces.

3 Results

3.1 Head-to-head selectivity comparison

Table 1 summarizes the four bases.

INLP is the only basis with positive selectivity. All three causal bases have negative diagonal dominance — noise injected along their directions produces *more* entropy change in non-target domains than in the target domain.

Table 1: Basis comparison at layer 10, $\sigma = 0.2$, Qwen 2.5-7B.

Basis	Mean Selectivity	Diag. Dominance	$\bar{R}_{ii}$	$\bar{R}_{i \neq j}$
INLP	+0.618	+1.818	9.383	7.565
Gradient	−0.116	−1.157	3.896	5.054
Patching	−0.630	−4.073	11.554	15.627
Contrastive	−0.777	−2.159	11.066	13.225

3.2 Response matrices

Table 2 shows the full 4×4 response matrices. Rows are injection targets (whose directions are used); columns are measurement domains.

Table 2: Entropy response matrices (% change). Diagonal entries are self-domain effects.

Basis	Target	Medical	Legal	Code	Science
4*INLP	Medical	**16.93**	4.04	3.28	7.64
	Legal	7.74	**1.88**	14.49	19.48
	Code	2.74	−0.99	**6.08**	19.75
	Science	5.93	2.95	3.74	**12.65**
4*Patching	Medical	**5.89**	19.10	8.06	19.15
	Legal	8.70	**20.31**	10.67	24.94
	Code	30.15	15.30	**9.31**	18.42
	Science	12.34	12.11	8.58	**10.70**
4*Contrastive	Medical	**15.68**	9.18	13.26	20.19
	Legal	7.56	**7.56**	5.50	18.52
	Code	16.82	20.81	**11.24**	10.56
	Science	10.72	12.53	13.04	**9.79**
4*Gradient	Medical	**17.78**	18.72	23.65	16.60
	Legal	17.64	**9.06**	−5.23	1.12
	Code	5.33	−8.09	**2.43**	−0.55
	Science	13.09	−10.68	−10.94	**−13.68**

The INLP matrix shows clear diagonal enhancement for medical (+16.93%) and science (+12.65%). Legal shows a paradox: legal-targeted noise produces more entropy change in code (+14.49%) and science (+19.48%) than in legal itself (+1.88%), consistent with the "shared substrate" finding from Paper IX.

The patching matrix is dominated by off-diagonal effects. Medical-targeted patching directions produce their largest effect on legal (+19.10%) and science (+19.15%), not medical (+5.89%). Code-targeted patching produces a massive medical effect (+30.15%).

The gradient matrix shows a striking pattern: science-targeted gradient directions produce

negative entropy changes in science (−13.68%), legal (−10.68%), and code (−10.94%). Noise along gradient-sensitive directions can *reduce* entropy — a form of constructive interference — but not selectively.

3.3 Per-domain selectivity

Table 3: Per-domain selectivity by basis.

Basis	Medical	Legal	Code	Science
INLP	+2.20	−1.80	−0.14	+2.21
Patching	−1.56	+0.83	−1.58	−0.21
Contrastive	+0.37	−0.58	−1.15	−1.75
Gradient	−0.70	+0.53	+0.71	−1.00

The gradient basis achieves positive selectivity for code (+0.71), outperforming INLP (−0.14) in that domain. This is the single case where a causal basis beats INLP, and it occurs for the domain where INLP is weakest. The gradient directions appear to find the causal pathway for code-domain output better than the classification boundary does. However, the gradient basis is anti-selective overall.

3.4 Basis overlaps

Table 4: Mean absolute cosine similarity between basis pairs. Random baseline ≈ 0.017.

	Patching	Contrastive	Gradient
INLP	0.015	0.016	0.011
Patching	—	0.158	0.022
Contrastive	—	—	0.018

All INLP-vs-causal overlaps are at or below the random baseline, confirming that INLP directions and causal directions lie in nearly orthogonal subspaces. The patching-contrastive overlap (0.158) is the exception: these two bases share some directions, which makes sense because both are derived from activation differences between domain and non-domain inputs (one via swapping, the other via averaging).

4 Discussion

4.1 The classification-intervention dissociation

The central result is counterintuitive: the basis that best *classifies* domain activations also produces the best *interventions*, while bases derived from causal analysis of the model's computation are actively worse. This is not a minor quantitative difference — the causal bases have negative selectivity, meaning they are worse than injecting noise along random directions.

Why? The answer lies in the geometry of the forward pass established in Papers X–XI. The Jacobian between layers is an isotropic amplifier (Paper X, INLP/random amplification ratio = 0.991). This means the model treats all directions approximately equally during forward propagation. There is no preferred "causal pathway" for domain information that differs from the classification boundary.

INLP's advantage is that it finds the classification boundary — the hyperplane that maximally separates domain distributions. When noise is injected along this boundary, it pushes activations toward or away from the domain cluster, producing domain-correlated entropy changes. Causal directions (what changes output when perturbed) are not the same thing as discriminative directions (what separates domains in representation space).

4.2 Why causal bases fail

The near-zero overlaps (0.011–0.016) confirm the concentration barrier from Paper XI. In $\mathbb{R}^{3584}$, any two 9-dimensional subspaces chosen by different methods will be nearly orthogonal. The INLP, patching, contrastive, and gradient methods each find a different 36-dimensional subspace of the 3584-dimensional activation space, and these subspaces have essentially no intersection.

This is not a failure of the causal methods — they correctly identify directions that are causally relevant to the model's computation. But causal relevance is not domain-selective relevance. The gradient directions capture where the model is most sensitive to perturbation. The patching directions capture where activation swapping changes output most. Neither of these is the same as where domain information is *concentrated*, because domain information is concentrated (by INLP's construction) along the classification boundary.

4.3 The code/gradient exception

The gradient basis outperforms INLP for code domain selectivity (+0.71 vs. −0.14). Code is the domain where INLP shows weakest selectivity across all papers. One interpretation:

the code domain's computational pathway is less isotropic than the other domains — the gradient finds a genuine causal channel that INLP's classification boundary misses. This is consistent with code probes producing qualitatively different activation patterns (structured syntax vs. natural language).

4.4 Implications for activation steering

These results constrain the design space for activation-level interventions in multi-agent systems. If the goal is domain-selective modulation of an LLM's behavior, INLP-type classification boundaries are the correct injection basis, not causally-derived directions. This is because:

1. The forward pass is isotropic (Paper X), so there is no causal "highway" to exploit.
2. Domain information occupies a geometrically tiny subspace (Paper XI), so interventions must target exactly that subspace.
3. Classification-optimized directions are the only ones that reliably point at the domain subspace.

5 Conclusion

INLP directions produce the only basis with positive domain selectivity. Three causally-derived alternatives are all anti-selective, with near-zero overlap to INLP. The classification-intervention dissociation is explained by the isotropic forward pass: in a geometry where all directions are amplified equally, only the classification boundary provides domain-specific targeting. Causal directions identify perturbation-sensitive subspaces, but these subspaces are orthogonal to the domain-discriminative subspace in 3584 dimensions.

The single exception — gradient basis outperforming INLP for code — suggests that code domain processing may involve a less isotropic computational pathway, a finding worth investigating in future work.

Data Availability

All results, including response matrices, basis overlap matrices, and extracted causal directions, are archived at `huggingface.co/datasets/jmcentire/paper8-data` under `paper14/`.

Series: Activation Geometry of Domain-Selective Noise Injection, Paper XIV.