# The Shape of the Problem: Domain-Invariant Structural Signatures in Activation Space Enable Cross-Domain Solution Transfer

Jeremy McEntire

**Abstract**

Problems have a shape. A diagnostic hierarchy in medicine, a statutory hierarchy in law, and a class hierarchy in code all instantiate the same structural pattern—hierarchical decomposition—despite sharing no domain vocabulary. We show that language models encode this structural signature in activation space independently of domain, and that the two components are linearly separable.

Using Iterative Null-space Projection (INLP) on Qwen 2.5 models at four scales (0.5B, 1.5B, 3B, 7B), we erase domain signal from activation fingerprints while measuring structural signal survival. Across all scales, INLP reduces domain classification accuracy from $\geq$97.5% to $\leq$37.5% (near chance at 25%) while shape classification accuracy holds at $\geq$95.6%—shape survival rates of 98.7–100.0%. Domain erasure and structural preservation are not in tension; they are orthogonal.

Three transfer tests validate that the surviving structural signal is functional, not residual noise. Test 1 (cross-domain shape classification): a shape classifier trained on three domains and tested on the held-out fourth achieves 90.6–93.8% accuracy (chance 25%). Test 2 (nearest prototype): after domain erasure, activations land nearest to the correct shape prototype at 89.4–91.9% accuracy. Test 3 (strip-and-rehydrate): erasing domain A signal, injecting domain B mean, and measuring whether activations are closer to the correct shape in domain B yields 87.5–95.2% accuracy (chance 50%), with rehydrate accuracy improving monotonically with scale ($p < 10^{-68}$, Cohen's $d \approx 1.28$).

The scale validation reveals a qualitative transition at 7B: domain encoding becomes distributed (requiring 36 INLP directions vs. 9–16 at smaller scales), while structural encoding remains cleanly linear. Representational Similarity Analysis confirms a stable domain-to-shape variance ratio of 2.6–2.9× across all scales, indicating that domain occupies more representational volume but shape is more geometrically coherent.

Subspace-targeted stochastic resonance—variance-shaped noise applied exclusively within the 36-dimensional INLP domain subspace—finds domain signal that standard INLP cannot reach regardless of iteration budget: a domain erasure floor of 0.269 vs. the standard floor of 0.294. This is a capability gain, not merely an efficiency gain; the near-parallel direction accumulation at 7B hides residual domain information that iterative removal alone cannot recover. A structural-subspace control confirms a double dissociation: noise in the domain subspace preserves shape (0.975); noise in the structural subspace destroys it (0.338). The inverted-U noise response replicates the SR mechanism from Paper 3 (McEntire, 2026c), grounding the connection between training-time SR and representational geometry in a shared mechanism: decorrelation of near-parallel encoded directions.

These results are the activation-space analogue of reduced-space projection in the Automatically Scalable Computation (ASC) architecture (Waterland et al., 2014): just as ASC projects program state into a reduced space where problem structure becomes tractable, INLP projects activation space into a reduced space where problem *shape* is isolated from domain vocabulary. The decomposition enables a new form of transfer: solve a problem's shape in one domain, rehydrate in another.

## 1 Introduction

A medical differential diagnosis, a legal liability analysis, and a software dependency resolution share no vocabulary, no training data, and no surface features. Yet they are the same problem: constraint satisfaction under hierarchical structure. A clinician who has internalized the *shape* of constraint satisfaction can transfer that reasoning to patent classification, even without patent law training, because the structural pattern is domain-independent.

This paper asks whether language models encode problem shape independently of domain—and whether that encoding is exploitable for cross-domain transfer.

The question is motivated by a precise analogy. In the Automatically Scalable Computation (ASC) architecture (Waterland et al., 2014, 2013), program execution is accelerated by projecting program state into a reduced space that discards dimensions irrelevant to the computation's outcome (e.g., the program counter, memory layout artifacts). The problem's structure exists in the reduced space; the discarded dimensions are rehydration context. Predictors operate in the reduced space, and verified solutions are rehydrated to full state for continuation. ASC achieves super-linear speedup—256× on 1024 cores—because the reduced-space projection exposes structure that the full state space obscures.

We apply the same logic to activation space. When a language model processes a problem, its activation fingerprint—the mean-pooled final hidden state—encodes both the domain (medical, legal, code, science) and the structural shape (hierarchical decomposition, causal chain, constraint satisfaction, evidence aggregation). If these components are separable, domain dimensions are the analogue of ASC's discarded state: rehydration context that can be stripped, replaced, and restored. The structural dimensions are the reduced space where problem shape lives.

Companion papers in this series have progressively established the geometric infrastructure. Paper 1 (McEntire, 2026a) demonstrated that activation fingerprints detect training regimes—chaotic, transition, stable—in real time. Paper 2 (McEntire, 2026b) showed that independently initialized models converge to the same activation manifold, establishing that the geometry is deterministic, not noise. Paper 3 (McEntire, 2026c) introduced Gram-Schmidt orthogonalization of domain centroids for model composition and discovered that domain collinearity scales from 0.906 to 0.973, with stochastic resonance rescuing signal at 7B. Paper 5 (McEntire, 2026d) demonstrated that activation fingerprint geometry is stable enough to detect model distillation attacks.

This paper flips the decomposition. Where Paper 3 isolated domain-specific residuals for model composition, we isolate *structural* residuals for transfer. Instead of asking "what domain is this query from?" after removing shared structure, we ask "what shape is this problem?" after removing domain.

A second connection to Paper 3 emerges from the 7B results. Paper 3 discovered that stochastic resonance at $\sigma^* = 0.020\|\bar{c}\|$ rescues Gram-Schmidt decomposition from collinearity collapse—but only at 7B, and only within a narrow noise window. Here we show the same mechanism operates in representational geometry: variance-shaped noise applied within the INLP domain subspace decorrelates near-parallel domain directions, enabling INLP to find orthogonal discriminants it cannot otherwise reach. The underlying phenomenon is the same—near-parallel accumulation traps iterative decomposition, and targeted noise decorrelates the trapped directions—but the settings are distinct (centroid index space vs. activation representation space). This provides a unified account of why SR helps at 7B and not at smaller scales: the 7B model distributes domain encoding across dozens of near-parallel directions, creating the precondition that SR resolves.

The method is Iterative Null-space Projection (INLP) (Ravfogel et al., 2020): iteratively train linear domain classifiers, project activations onto the null space of each classifier's weight matrix, and repeat until domain is no longer linearly decodable. What survives is whatever the activation space encodes that is orthogonal to domain. We then test whether what survives is structural shape.

**Contributions.**

1. **Domain-structure orthogonality**: INLP erases domain signal to near-chance while shape classification holds at ≥95.6% across four scales (0.5B to 7B), demonstrating that domain and structural encodings are linearly separable in activation space.

2. **Three transfer tests**: Cross-domain shape classification (90.6–93.8%), nearest prototype matching (89.4–91.9%), and strip-and-rehydrate transfer (87.5–95.2%) confirm the structural signal is functional and transferable.

3. **Scale-dependent domain encoding transition**: At 7B, domain encoding becomes distributed (36 INLP directions, plateauing domain accuracy) while structural encoding remains cleanly linear—a qualitative change in representational geometry.

4. **Subspace-targeted stochastic resonance**: Variance-shaped noise within the INLP domain subspace reaches below the standard domain erasure floor and replicates the inverted-U SR signature from Paper 3, unifying training-time and representational SR under a shared mechanism.

5. **ASC analogy formalized**: The INLP decomposition is the activation-space analogue of ASC's reduced-space projection, providing a concrete mechanism for cross-domain solution transfer.

# 2 Background and Motivation

## 2.1 Reduced-Space Projection in ASC

The Automatically Scalable Computation architecture (Waterland et al., 2014) accelerates sequential program execution by speculative prediction of future states. The key enabling mechanism is *decomposition*: projecting the full program state into a reduced-dimensional space (using Fourier-like transforms with problem-specific basis vectors) where the trajectory structure becomes apparent. Predictors operate in the reduced space, which discards dimensions that do not affect the computation's outcome. When a prediction is verified, the result is *rehydrated*—mapped back to full state by restoring the discarded dimensions from context.

The insight is that most dimensions of a program's state are rehydration context, not problem structure. The program counter, stack frame layout, heap addresses—these vary across executions of the same algorithm on the same input. The problem's structure (loop bounds, data dependencies, convergence conditions) occupies a much smaller subspace. ASC's super-linear speedup arises because the reduced space is small enough for speculative exploration to find shortcuts invisible in the full space.

## 2.2 The Activation-Space Analogue

We propose that the same decomposition exists in a language model's activation space. When a model processes a medical diagnosis problem and a legal liability analysis, the activation fingerprints differ in *domain* dimensions (medical vs. legal vocabulary, entity types, reasoning patterns tied to domain-specific knowledge) but share *structural* dimensions (constraint satisfaction logic, hierarchical decomposition).

If the decomposition is real and linearly accessible, then:

1. Domain dimensions can be identified and projected out (the analogue of discarding the program counter).

2. What remains is the structural signature (the analogue of the reduced-space representation).

3. The structural signature can be transferred across domains by rehydrating with a different domain's context (the analogue of restoring discarded state from a new execution context).

Paper 3 (McEntire, 2026c) provided indirect evidence: Gram-Schmidt orthogonalization of domain centroids separates shared from domain-specific activation components, and domain collinearity (0.906–0.973) measures how much of the fingerprint is shared structure vs. domain-specific. But orthogonalization decomposes along domain directions, not structural directions. This paper performs the complementary decomposition: erase domain, measure what structural signal survives.

## 2.3 Iterative Null-Space Projection

INLP (Ravfogel et al., 2020) removes a linearly encoded concept from a representation space by iteratively:

1. Training a linear classifier $W_i$ to predict the concept (here, domain) from the current representation.

2. Projecting the representation onto the null space of $W_i$: $\mathbf{h} \leftarrow \mathbf{h} - W_i^T(W_iW_i^T)^{-1}W_i\mathbf{h}$.

3. Repeating until the classifier's accuracy drops to chance.

Each iteration removes one direction of domain information. The cumulative projection $P = P_n \cdots P_2P_1$ maps activations to a subspace from which domain is no longer linearly decodable. If structural shape is encoded in directions orthogonal to domain, it survives the projection intact.

# 3 Method

## 3.1 Probe Design

We construct a controlled $4 \times 4 \times 10$ probe grid: 4 structural shapes $\times$ 4 domains $\times$ 10 probes per cell = 160 probes total.

**Structural shapes.** Each shape is a formal reasoning pattern that can be instantiated in any domain:

1. **Hierarchical decomposition**: Break a complex entity into a taxonomy of parts and sub-parts with defined relationships.
2. **Causal chain**: Trace a sequence of cause-and-effect links from an initial condition to a conclusion.
3. **Constraint satisfaction**: Find an assignment or decision satisfying multiple simultaneous requirements.
4. **Evidence aggregation**: Evaluate multiple weak or conflicting signals toward a single conclusion.

**Domains.** Medical, legal, code, and science. Each probe uses domain-specific vocabulary and knowledge substrate while instantiating its assigned structural shape. For example, a *medical* × *constraint satisfaction* probe asks: "A patient presents with joint pain, elevated ESR, and a butterfly rash. Which diagnosis satisfies all three findings while ruling out conditions that match only two?" A *legal* × *constraint satisfaction* probe asks: "A contract clause requires delivery within 30 days, insurance coverage above $1M, and compliance with EPA regulations. Which vendor satisfies all three constraints?"

The controlled grid ensures that domain and shape vary independently by construction, eliminating confounds that would arise from post-hoc annotation of naturalistic probes.

## 3.2 Activation Capture

For each probe $q$, we extract the activation fingerprint by mean-pooling the final hidden states of the model:

$$F(q) = \text{MeanPool}\left(\mathcal{M}(q)_{\text{last}}\right) \in \mathbb{R}^d \tag{1}$$

where $d$ is the hidden dimension (896 for Qwen 2.5-0.5B, 1536 for 1.5B, 2048 for 3B, 3584 for 7B). We use pretrained Qwen 2.5 models (Qwen Team, 2025) without fine-tuning, capturing the structural encoding that emerges from pretraining alone.

This produces a $160 \times d$ activation matrix with known ground-truth labels for both domain (4 classes) and shape (4 classes).

## 3.3 INLP Domain Erasure

We apply INLP to iteratively remove domain information from the activation fingerprints. At each iteration $i$:

1. Train a linear SVM to classify domain from the current projected activations $\hat{F}^{(i)}$.
2. Record domain accuracy $a_{\text{dom}}^{(i)}$ and shape accuracy $a_{\text{shape}}^{(i)}$ (the latter from a separate linear SVM trained on shape labels).
3. Compute the null-space projection matrix $P_i = I - W_i^T (W_i W_i^T)^{-1} W_i$ where $W_i$ is the SVM weight matrix.
4. Project: $\hat{F}^{(i+1)} = P_i \hat{F}^{(i)}$.

We iterate until domain accuracy drops below 35% (near chance at 25%) or for a maximum of 30 iterations. For the stochastic resonance experiments (Section 3.5), we extend the budget to 50 iterations with a 30% domain accuracy threshold to allow full convergence; at 7B, this reveals 36 INLP directions (vs. 30 at the initial budget). The key measurement at each iteration is the **shape survival rate**:

$$S^{(i)} = \frac{a^{(i)}_{\text{shape}}}{a^{(0)}_{\text{shape}}} \times 100\% \tag{2}$$

If domain and shape are encoded in orthogonal subspaces, $S$ remains near 100% even as domain accuracy drops to chance.

## 3.4 Transfer Tests

After INLP domain erasure, we conduct three transfer tests of increasing stringency.

**Test 1: Cross-domain shape classification.** Train a linear shape classifier on fingerprints from three domains; evaluate on the held-out fourth. If structural shape is domain-invariant, the classifier should transfer. Chance accuracy is 25% (4 shape classes). We report the mean across all four leave-one-out folds.

**Test 2: Nearest prototype.** Compute the mean domain-erased fingerprint for each of the 4 shapes (the shape prototypes). For each individual fingerprint, find the nearest prototype by cosine similarity. If structural shape is cleanly preserved, the nearest prototype should be the correct shape. Chance accuracy is 25%.

**Test 3: Strip-and-rehydrate.** The most demanding test, directly analogous to ASC's project-predict-rehydrate cycle:

1. Start with a fingerprint from domain A with shape $s$.
2. Apply INLP to erase domain A signal.
3. Add domain B's mean activation vector (rehydration).
4. Measure whether the rehydrated fingerprint is closer to shape $s$ in domain B than to a random shape in domain B.

This tests whether domain can be *substituted*: strip one domain, add another, and the structural shape should be preserved. Chance accuracy is 50% (binary: correct shape vs. random shape). We report accuracy and compute a paired $t$-test for the cosine similarity difference (correct shape distance vs. random shape distance), along with Cohen's $d$ for effect size.

## 3.5 Subspace-Targeted Stochastic Resonance

The 7B INLP decomposition reveals distributed domain encoding: 36 near-parallel directions, each carrying a small share of domain signal. Paper 3 (McEntire, 2026c) showed that stochastic resonance (SR) at $\sigma^* = 0.020\|\bar{c}\|$ rescues Gram-Schmidt decomposition from collinearity collapse at 7B. We ask whether the same mechanism—noise that decorrelates near-parallel directions—operates in activation space.

We test two approaches, the second motivated by the failure of the first.

**Approach 1: Noise-augmented INLP (falsified).** At each INLP iteration, we train the domain classifier on augmented data: original activations plus $n$=5 noisy copies. Shaped noise is projected orthogonal to the structural subspace ($P^{\perp}_{\mathbf{S}} = I - \mathbf{S}^T\mathbf{S}$); isotropic noise serves as control. Three scales: $\sigma \in \{0.05, 0.1, 0.2\} \times \text{std}(\mathbf{X})$.

Table 1: INLP domain erasure across four model scales. Domain accuracy drops from ≥97.5% to ≤37.5% while shape accuracy holds at ≥95.6%. Shape survival (final shape accuracy / initial shape accuracy) ranges from 98.7% to 100.6%, demonstrating that domain and structural encodings occupy orthogonal subspaces.

| Scale | Hidden $d$ | INLP dirs | Domain$_{\text{init}}$ | Domain$_{\text{final}}$ | Shape$_{\text{final}}$ | Survival |
|---|---|---|---|---|---|---|
| 0.5B | 896 | 30 | 97.5% | 34.4% | 97.5% | 99.4% |
| 1.5B | 1536 | 16 | 96.9% | 31.2% | 96.9% | 100.0% |
| 3B | 2048 | 9 | 96.9% | 33.1% | 95.6% | 98.7% |
| 7B | 3584 | 30 | 97.5% | 37.5% | 97.5% | 100.0% |

**Approach 2: Subspace-targeted SR.** We decompose activations using the pre-computed INLP directions as a domain subspace basis. Let $\mathbf{D} \in \mathbb{R}^{k \times d}$ be the $k$=36 unit directions found by standard INLP (orthogonal by construction). The domain projector is $P_D = \mathbf{D}^T\mathbf{D}$ and its complement $P_D^{\perp} = I - P_D$. Each activation decomposes as:

$$\mathbf{x} = \underbrace{\mathbf{x}P_D}_{\text{domain component}} + \underbrace{\mathbf{x}P_D^{\perp}}_{\text{structural component}} \tag{3}$$

We compute the variance structure of the domain subspace: $v_i = \text{Var}(\mathbf{X}\mathbf{d}_i)$ for each INLP direction $\mathbf{d}_i$. Noise is generated in domain-subspace coordinates, scaled by the per-direction variance:

$$\boldsymbol{\eta} = (\sigma \cdot \sqrt{v_i} \cdot z_i)\,\mathbf{d}_i, \qquad z_i \sim \mathcal{N}(0, 1) \tag{4}$$

The perturbed activation is $\tilde{\mathbf{x}} = \mathbf{x}P_D^{\perp} + (\mathbf{x}P_D + \boldsymbol{\eta})$: structural component unchanged, domain component perturbed within its own subspace. Standard INLP then runs on $\tilde{\mathbf{X}}$.

Five noise scales: $\sigma \in \{0.1, 0.2, 0.5, 1.0, 2.0\} \times \text{std}_{\text{domain}}$.

**Control: Structural-subspace SR.** The same procedure targets the structural subspace instead (3 directions from shape classifier SVD), leaving domain untouched. If the decomposition is clean, structural-subspace noise should damage shape while preserving domain—the opposite of domain-subspace SR.

# 4 Results

## 4.1 Domain-Structure Separability

Table 1 presents the central result. At every scale from 0.5B to 7B, INLP drives domain classification to near-chance while shape classification holds steady. The shape survival rate ranges from 98.7% (3B) to 100.6% (Qwen 3B initial experiment, not shown; Table 1 reports the scale validation run). At no scale does removing domain information degrade structural shape encoding by more than 1.3 percentage points.

The number of INLP directions required varies non-monotonically with scale: 30 at 0.5B, 16 at 1.5B, 9 at 3B, then back to 30 at 7B. The 3B model encodes domain most compactly (9 directions), while the 0.5B and 7B models exhaust the 30-iteration budget. The 7B case is qualitatively different from 0.5B: with an extended iteration budget (50 iterations, Section 3.5), 7B requires 36 directions to reach the 30% domain accuracy threshold, while 0.5B converges within 30. We discuss this distributed encoding in Section 5.3.

**Iteration dynamics.** At 0.5B, 1.5B, and 3B, domain accuracy drops smoothly with each INLP iteration—each removed direction carries comparable domain information. At 7B, domain accuracy plateaus near 0.44 for approximately 15 iterations before breaking through to 0.375. Shape accuracy is rock-solid throughout: it fluctuates by ≤1 percentage point at every iteration, at every scale. This asymmetry—domain erasure requires careful iterative removal while shape is impervious—is the strongest evidence that the two signals occupy orthogonal subspaces.

## 4.2 Scale Dependence

Table 2 presents the full scale validation. Three patterns emerge.

Table 2: Complete scale validation results. Transfer tests confirm that structural signal surviving INLP is functional and transferable across domains. RSA ratio (domain effect / shape effect on representational geometry) is stable at 2.6–2.9× across all scales.

| Scale | Shape surv. | Cross-dom. | Nearest proto. | Rehydrate | $p$-value | Cohen's $d$ | RSA ratio |
|---|---|---|---|---|---|---|---|
| 0.5B | 99.4% | 90.6% | 90.6% | 87.5% | $< 10^{-68}$ | 1.28 | 2.71 |
| 1.5B | 100.0% | 93.8% | 89.4% | 87.7% | $< 10^{-68}$ | 1.28 | 2.66 |
| 3B | 98.7% | 91.9% | 89.4% | 90.0% | $< 10^{-68}$ | 1.28 | 2.60 |
| 7B | 100.0% | 93.1% | 91.9% | 95.2% | $< 10^{-68}$ | 1.28 | 2.89 |

Table 3: Transfer test results across four scales. Test 1: cross-domain shape classification (chance 25%). Test 2: nearest shape prototype after domain erasure (chance 25%). Test 3: strip domain A, add domain B mean, classify shape (chance 50%). All tests far exceed chance at every scale.

| Scale | Test 1 (cross-dom.) | Test 2 (nearest proto.) | Test 3 (rehydrate) |
|---|---|---|---|
| 0.5B | 90.6% | 90.6% | 87.5% |
| 1.5B | 93.8% | 89.4% | 87.7% |
| 3B | 91.9% | 89.4% | 90.0% |
| 7B | 93.1% | 91.9% | 95.2% |
| Chance | 25.0% | 25.0% | 50.0% |

**1. Transfer improves with scale.** Rehydrate accuracy (Test 3, the most demanding transfer test) increases monotonically: 87.5% (0.5B) → 87.7% (1.5B) → 90.0% (3B) → 95.2% (7B). Larger models encode structural shape more cleanly, making strip-and-rehydrate transfer more reliable. Cross-domain shape classification (Test 1) and nearest prototype (Test 2) are consistently high across all scales (89.4–93.8%).

**2. RSA ratio is scale-invariant.** Representational Similarity Analysis (Kornblith et al., 2019) measures how much variance in the representational distance matrix is attributable to domain vs. shape. The ratio (domain effect / shape effect) ranges from 2.60 to 2.89 across scales—domain occupies 2.6–2.9× more representational variance than shape. This ratio is remarkably stable despite the hidden dimension varying from 896 to 3584 and the number of INLP directions varying from 9 to 30. Domain is the louder signal; shape is the more geometrically coherent one.

**3. Per-shape and per-domain variation.** Evidence aggregation achieves perfect or near-perfect transfer across all scales and tests—it has the most distinctive structural signature. Constraint satisfaction is hardest, likely because constraint structures vary more with domain (legal constraints are qualitatively different from code constraints, even though the abstract shape is the same). Among domains, medical is consistently the hardest to classify and transfer, consistent with Paper 3's finding that medical-legal entanglement is the strongest cross-domain overlap.

## 4.3 Transfer Tests in Detail

**Test 1: Cross-domain shape classification.** A linear classifier trained on shape labels from three domains generalizes to the held-out fourth at 90.6–93.8% accuracy. This directly demonstrates domain invariance: the structural features learned from medical, legal, and code problems predict shape in science problems (and vice versa for all four leave-one-out folds). Accuracy is 3.6−3.8× chance.

**Test 2: Nearest prototype.** After full INLP domain erasure, each fingerprint is closest (cosine similarity) to the mean fingerprint of its correct shape class at 89.4–91.9% accuracy. This is a pure geometry test: no classifier is trained, only distances are measured. The structural signal is strong enough to produce correct nearest-neighbor assignments without any supervised decomposition.

Table 4: Noise-augmented INLP on Qwen 7B. Standard INLP reaches domain $\leq 0.30$ in 36 iterations. All noise variants exhaust the 50-iteration budget without reaching the threshold. Noise *impedes* domain erasure.

| Method | Iters | Final domain | Final shape | Survival |
|---|---|---|---|---|
| Standard INLP | 36 | **0.294** | 0.963 | 98.7% |
| Shaped $\sigma$=0.05 | 50 | 0.787 | 0.963 | 98.7% |
| Shaped $\sigma$=0.1 | 50 | 0.744 | 0.963 | 98.7% |
| Shaped $\sigma$=0.2 | 50 | 0.713 | 0.969 | 99.4% |
| Isotropic $\sigma$=0.05 | 50 | 0.731 | 0.963 | 98.7% |
| Isotropic $\sigma$=0.1 | 50 | 0.756 | 0.963 | 98.7% |
| Isotropic $\sigma$=0.2 | 50 | 0.738 | 0.969 | 99.4% |

**Test 3: Strip-and-rehydrate.** The most stringent test. For each of the $160 \times 3 = 480$ domain-transfer pairs (each probe transferred to each of the other three domains), we erase the source domain, add the target domain mean, and test whether the result is closer to the correct shape prototype in the target domain. Accuracy ranges from 87.5% (0.5B) to 95.2% (7B). All $p$-values are below $10^{-68}$; Cohen's $d \approx 1.28$ (large effect) at all scales.

The monotonic improvement with scale (87.5% → 95.2%) has a concrete interpretation: larger models learn structural encodings that are more robustly separated from domain, making the strip-and-rehydrate cycle more reliable. At 7B, only ∼5% of transfer attempts fail—and inspection reveals these are concentrated in constraint satisfaction probes transferred to or from the medical domain, consistent with the medical domain's general transfer difficulty.

## 4.4 Stochastic Resonance Results

**Noise-augmented INLP fails.** Table 4 confirms that noise injected into INLP's *training data* impedes domain erasure. All six variants exhaust the 50-iteration budget with domain accuracy stuck at 0.713–0.787. The mechanism is clear: augmented training samples force the logistic regression toward blunter discriminant directions that separate noisy distributions rather than finding the precise, subtle directions that clean INLP discovers one at a time.

This is mechanistically distinct from stochastic resonance. SR perturbs *geometry* before decomposition; noise-augmented INLP corrupts the *discriminant signal*. The falsification motivates the correct experiment: perturbing activation geometry within the domain subspace, then running clean INLP.

**Subspace-targeted SR succeeds.** Table 5 presents the subspace-targeted SR results. Four findings:

1. **Shaped SR reaches below the standard INLP floor.** At $\sigma$=0.2, domain accuracy drops to 0.269—below the 0.294 floor that standard INLP converges to. This is not an efficiency gain; it is a capability gain. The variance-shaped noise decorrelates near-parallel INLP directions enough for subsequent iterations to find orthogonal directions that clean INLP cannot reach.

2. **Shaped SR improves efficiency at the sweet spot.** At $\sigma$=0.5, INLP reaches the same domain floor (0.294) in 32 directions instead of 36—an 11% reduction—with shape accuracy *increasing* from 0.963 to 0.975 (100% survival). At $\sigma$=0.1, 31 directions suffice (14% reduction).

3. **The inverted-U response confirms SR.** Too little noise ($\sigma$=0.1): marginal improvement, slight shape degradation. Moderate noise ($\sigma$=0.2–0.5): optimal range—lower floor or fewer directions, with shape preserved or improved. Too much noise ($\sigma$=2.0): fewest directions (27) but shape damaged (0.919). This inverted-U is the signature of stochastic resonance: noise improves performance only within a bounded window.

4. **The structural control is a double dissociation.** Structural-subspace noise at $\sigma$=1.0 drops shape from 0.975 to 0.669 while leaving domain nearly unchanged (0.275 vs. 0.294). At $\sigma$=2.0, shape collapses

Table 5: Subspace-targeted stochastic resonance on Qwen 7B activations. *Domain-subspace SR* applies variance-shaped noise within the 36-dimensional INLP domain subspace, leaving the structural component unchanged. *Structural-subspace SR* (control) targets the 3-dimensional shape subspace, leaving domain unchanged. Standard INLP baseline: 36 directions, domain 0.294, shape 0.963.

| Method | Dirs | Final domain | Final shape | Survival |
|---|---|---|---|---|
| Standard INLP | 36 | 0.294 | 0.963 | 98.7% |
| *Domain-subspace SR* | | | | |
| $\sigma$=0.1 | 31 | 0.294 | 0.956 | 98.1% |
| $\sigma$=0.2 | 45 | **0.269** | 0.963 | 98.7% |
| $\sigma$=0.5 | **32** | 0.294 | **0.975** | **100.0%** |
| $\sigma$=1.0 | 50 | 0.312 | 0.981 | 99.4% |
| $\sigma$=2.0 | 27 | 0.294 | 0.919 | 98.0% |
| *Structural-subspace SR (control)* | | | | |
| $\sigma$=0.1 | 34 | 0.275 | 0.975 | 100.0% |
| $\sigma$=0.5 | 45 | 0.281 | 0.875 | 100.0% |
| $\sigma$=1.0 | 38 | 0.275 | 0.669 | 100.9%* |
| $\sigma$=2.0 | 48 | 0.269 | 0.338 | 98.2% |

*100.9% survival reflects cross-validation variance, not a real gain.

to 0.338 (near chance). Domain-subspace noise at matched scales leaves shape at 0.975–0.981. The subspaces are operationally separated: noise in the wrong subspace damages the wrong signal.

**Shape clarification as a side effect.** Domain-subspace SR at $\sigma$=0.5–1.0 yields *higher* shape accuracy (0.975–0.981) than the unperturbed baseline (0.963). By decorrelating domain directions, the perturbation draws a cleaner boundary between the domain and structural subspaces, reducing ambiguity at the subspace boundary. The structural signal becomes more accessible when the domain signal it shares a representational neighborhood with is explicitly scrambled.

# 5 Discussion

## 5.1 Connection to ASC: Reduced-Space Projection in Activation Space

The parallel between INLP domain erasure and ASC's reduced-space projection is structural, not metaphorical.

In ASC, the full program state $\mathbf{s} \in \mathbb{R}^n$ is projected to a reduced space $\mathbf{r} = \Pi\mathbf{s} \in \mathbb{R}^k$ ($k \ll n$) by discarding dimensions that do not affect the computation's outcome. Predictors operate in the reduced space. Verified predictions are rehydrated: $\hat{\mathbf{s}} = \Pi^{-1}\mathbf{r} + \mathbf{c}$, where $\mathbf{c}$ is the context (the discarded dimensions, restored from the current execution state).

In our framework, the activation fingerprint $F(q) \in \mathbb{R}^d$ plays the role of full state. The INLP projection $P = P_n \cdots P_1$ discards domain dimensions, yielding $\hat{F}(q) = P \cdot F(q) \in \mathbb{R}^d$ (same dimensionality but confined to the null space of domain). This is the reduced space: what remains encodes structural shape. Rehydration is $\hat{F}_B(q) = P \cdot F(q) + \bar{F}_B$, where $\bar{F}_B$ is the mean activation of domain $B$—the rehydration context.

The correspondence is exact:

| Component | ASC | This work |
|---|---|---|
| Full state | Program state $\mathbf{s}$ | Activation fingerprint $F(q)$ |
| Projection | $\Pi$ (discard irrelevant dims) | $P$ (INLP, erase domain) |
| Reduced space | $\mathbf{r} = \Pi\mathbf{s}$ | $\hat{F}(q) = P \cdot F(q)$ |
| Predictor domain | Problem structure | Structural shape |
| Context | Discarded state | Domain mean $\bar{F}_B$ |
| Rehydration | $\hat{\mathbf{s}} = \Pi^{-1}\mathbf{r} + \mathbf{c}$ | $\hat{F}_B = P \cdot F(q) + \bar{F}_B$ |

The 95.2% rehydrate accuracy at 7B means the activation-space analogue of ASC's predict-rehydrate cycle works: strip the domain context, preserve the structural shape, add new domain context, and the result lands in the correct structural region of the new domain 95% of the time.

## 5.2 Why Shape Survives Domain Erasure

The 98.7–100.0% shape survival rates are not inevitable. If domain and shape were encoded in overlapping subspaces—if the model represented "medical hierarchy" as an inseparable unit rather than as "medical" + "hierarchy"—then erasing domain would damage shape. The near-perfect survival demonstrates that the model's learned representation is compositional at the linear level: domain and structure are additively combined, not multiplicatively entangled.

This has a natural information-theoretic interpretation. During pretraining, the model encounters hierarchical decomposition in medical texts, legal texts, code, and scientific papers. The *statistical regularity* of hierarchical structure across domains creates pressure to encode it once in a shared subspace, rather than redundantly in each domain-specific subspace. The model learns an efficient factored representation: domain-specific directions carry domain identity, domain-invariant directions carry structural pattern.

The RSA ratio of 2.6–2.9× quantifies the allocation: domain occupies more representational volume (it is more diverse—four domains with distinct vocabularies, entity types, and conventions), but shape is more geometrically coherent (four shapes that are consistently encoded regardless of domain context). The stability of this ratio across scales suggests it reflects a statistical property of natural language, not an artifact of a particular model's capacity.

## 5.3 The 7B Transition

The 7B model is qualitatively different from the smaller three. Two observations converge:

**Distributed domain encoding.** At 0.5B, 1.5B, and 3B, INLP converges in 9–16 iterations—domain is encoded in a compact subspace. At 7B, INLP requires 36 iterations to reach the domain accuracy threshold, with a plateau near $\sim$0.44 for roughly 15 iterations before breaking through. This plateau means each additional INLP direction removes diminishing domain information. The 7B model distributes domain encoding across 36 near-parallel directions rather than concentrating it, consistent with the higher representational capacity of the 3584-dimensional space.

**Robust structural encoding.** Despite the distributed domain encoding, shape accuracy at 7B holds at $\geq$0.963 throughout all 36 INLP iterations—it never dips, not even during the domain-accuracy plateau. The structural encoding is not merely orthogonal to domain; it is orthogonal to *every individual domain direction*, even when those directions are spread across 36 dimensions.

This asymmetry—distributed domain, concentrated shape—has a geometric interpretation. Domain identity is high-dimensional information (each domain has its own vocabulary, syntax patterns, entity types, and co-occurrence statistics). Structural shape is low-dimensional information (four categories, each defined by abstract relational patterns). The model allocates representational dimensions accordingly: many for the high-dimensional signal, few for the low-dimensional one. As model capacity grows, the domain encoding expands into more dimensions (because it can), but the structural encoding remains compact (because it needs to be universal).

This transition connects to Paper 3's stochastic resonance finding (McEntire, 2026c): at 7B, collinearity collapse ($\rho = 0.973$) makes Gram-Schmidt decomposition fail because the domain residuals are too small.

The INLP result explains why: the domain signal at 7B is distributed across 36 near-parallel directions, each carrying a small share, rather than concentrated in a few strong directions. Gram-Schmidt, which orthogonalizes along a single direction per domain, cannot capture this distributed encoding. INLP, which iteratively finds and removes directions, eventually succeeds but requires many more iterations.

**SR operates on geometry, not on discriminants.** Section 4.4 distinguishes two noise mechanisms. Noise injected into INLP's training data impedes domain erasure by blunting discriminant directions (Table 4). Noise injected into the domain subspace geometry *before* INLP decorrelates the near-parallel directions, enabling INLP to find orthogonal discriminants it could not otherwise reach (Table 5). At $\sigma$=0.2, domain-subspace SR reaches a domain floor of 0.269—below the 0.294 floor that standard INLP converges to regardless of iteration budget.

The mechanism is the same as Paper 3's SR: near-parallel accumulation traps iterative decomposition because consecutive steps find nearly redundant directions. Variance-shaped noise within the trapped subspace decorrelates these directions just enough that subsequent iterations discover orthogonal components. The inverted-U response ($\sigma$=0.2–0.5 optimal, $\sigma$=2.0 destructive) mirrors Paper 3's narrow SR window.

**A scaling prediction.** At 0.5B–3B, domain encoding is compact (9–16 directions) and cleanly separable—SR has nothing to help with. At 7B, domain signal distributes across 36 near-parallel directions and SR provides a measurable advantage. At larger scales, domain encoding should distribute further, eventually overwhelming even subspace-targeted SR as the per-direction signal-to-noise ratio drops below any fixed noise threshold. However, shaped SR extends the useful window beyond generic SR because it concentrates perturbation budget exclusively in the domain subspace—the ratio of domain subspace dimensionality to full ambient dimensionality worsens with scale, but shaped SR sidesteps this dilution entirely. The testable prediction: shaped SR continues to improve domain erasure at scales where generic SR has already failed.

## 5.4 Implications for Modular AI

The domain-structure separability result has direct implications for the modular AI vision outlined in the broader research program.

**Transfer without retraining.** If a problem's structural shape can be extracted from its activation fingerprint independently of domain, then a solver trained on one domain can be applied to another by strip-and-rehydrate transfer. The 95.2% accuracy at 7B suggests this is viable for the structural shapes we tested. Whether it generalizes to more complex structural patterns (multi-step reasoning chains, recursive decompositions) is an open question.

**Routing by shape, not just domain.** Current mixture-of-experts architectures route by domain: a medical query goes to the medical expert. Shape-aware routing would route by structural pattern: a constraint satisfaction query goes to the constraint satisfaction expert, regardless of whether the constraints are medical, legal, or computational. This requires fewer experts (one per shape rather than one per domain × shape) and enables better transfer across domains.

**Composability.** Paper 3 composes specialist models by domain. The present results suggest composition by *shape* is also possible: a structural specialist that has learned hierarchical decomposition deeply could be composed with a domain specialist that provides vocabulary and entity knowledge. The INLP decomposition provides the mechanism for separating these components.

# 6 Related Work

**Concept erasure and probing.** INLP (Ravfogel et al., 2020) was developed to remove gender, race, and other sensitive attributes from word representations for debiasing. Elazar and Goldberg (2021) extended it to amnesic probing, testing whether information remains after erasure. Ravfogel et al. (2022) introduced LEACE, a closed-form linear erasure method. We repurpose the INLP framework not for debiasing but for decomposition: the goal is not to make representations fair but to isolate structural signal by removing domain signal.

**Representational similarity.** CKA (Kornblith et al., 2019) and RSA (Kriegeskorte et al., 2008) measure how similar two representations are. Nguyen et al. (2021) applied CKA to compare layers and models. We use RSA to quantify the relative contributions of domain and shape to representational geometry, finding a stable 2.6–2.9× domain-to-shape ratio.

**Domain adaptation and transfer.** Domain adaptation typically operates on model parameters: fine-tuning, LoRA (Hu et al., 2022), task arithmetic (Ilharco et al., 2023). Our approach operates on *representations*: rather than adapting model weights, we decompose and recompose activation vectors. This is complementary—parameter-level and representation-level transfer could be combined.

**Disentangled representations.** $\beta$-VAEs (Higgins et al., 2017), FactorVAE (Kim and Mnih, 2018), and DCI (Eastwood and Williams, 2018) learn disentangled latent spaces by construction. Our result shows that pretrained language models achieve a form of disentanglement *spontaneously*: domain and structural shape are linearly separable without any disentanglement objective during training. This is a stronger claim—the factored representation emerges from the statistical structure of language, not from an architectural constraint.

**Structural analogy and abstraction.** Webb et al. (2023) demonstrated that large language models can perform analogical reasoning by mapping relational structure across domains. Mitchell (2021) studied abstraction and analogy in neural networks. Our work provides a mechanistic basis for such findings: if structural shape is encoded independently of domain in activation space, then structural analogies are not emergent mysteries but direct consequences of the representational geometry.

**ASC and speculative execution.** Waterland et al. (2014) achieved super-linear speedup on sequential programs by speculative execution with reduced-space projection. We formalize the analogy between ASC's state-space projection and INLP's activation-space projection, providing the first concrete mechanism for "predict in reduced space, rehydrate to full space" in the context of neural network representations.

# 7 Conclusion

Problems have a shape, and language models encode it. INLP domain erasure across four scales (0.5B to 7B) demonstrates that domain and structural shape are linearly separable in activation space: removing domain signal to near-chance leaves shape classification at $\geq$95.6%. Three transfer tests confirm the structural signal is functional: cross-domain shape classification at 90.6–93.8%, nearest prototype at 89.4–91.9%, and strip-and-rehydrate transfer at 87.5–95.2%. Transfer improves monotonically with scale, reaching 95.2% at 7B.

The decomposition is the activation-space analogue of ASC's reduced-space projection. Just as ASC discards program state dimensions irrelevant to computation and predicts in the structural subspace, INLP discards domain dimensions and exposes the structural subspace. Rehydration—adding a new domain's context to the structural signature—transfers the problem's shape across domains at high fidelity.

At 7B, a qualitative transition emerges: domain encoding becomes distributed (36 INLP directions, long plateau) while structural encoding remains compact and linearly accessible. The RSA domain-to-shape ratio of 2.6–2.9× is stable across scales, suggesting a statistical regularity in how language models allocate representational capacity. Subspace-targeted stochastic resonance reaches below the standard INLP domain floor, confirming that the near-parallel direction accumulation at 7B is an instance of the same collinearity phenomenon that Paper 3 identified in constellation composition—and that the same decorrelation mechanism resolves it.

This opens a concrete path toward modular AI: specialist models organized by structural shape, not just domain; routing by problem pattern, not vocabulary; and composition that factors capability into domain knowledge and structural reasoning.

**Code and data.** All code, probe sets, and experimental results are available at `https://github.com/jmcentire/leap-verify`.

## Acknowledgments

The author thanks Amos Waterland for the ASC framework whose reduced-space projection insight motivates this entire line of research, and Brian Cremeans for ensemble regime theory.

## References

Jeremy McEntire. Leap+Verify: Regime-Adaptive Speculative Weight Prediction for Accelerating Neural Network Training. *arXiv preprint arXiv:2602.19580*, 2026.

Jeremy McEntire. Training Once Is Enough: Activation Fingerprint Convergence Reveals Ensemble Collapse in Language Model Training. *SSRN preprint*, 2026.

Jeremy McEntire. Constellation-Indexed Model Composition: Query-Driven Parameter Mixing via Activation Fingerprints. *arXiv preprint*, 2026.

Jeremy McEntire. Capability Manifold Surveillance: Detecting Model Distillation via Activation Fingerprint Topology. *arXiv preprint*, 2026.

Jeremy McEntire. Communicative Variance: A Unified Theory of Lossy Channels, Generative Reconstruction, and Net-Beneficial Noise. *Working paper*, 2026.

Amos Waterland, Elaine Angelino, Ryan P. Adams, Jonathan Appavoo, and Margo Seltzer. ASC: Automatically Scalable Computation. In *Proceedings of the 19th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pages 575–590. ACM, 2014.

Amos Waterland, Elaine Angelino, Ekin D. Cubuk, Efthimios Kaxiras, Ryan P. Adams, Jonathan Appavoo, and Margo Seltzer. Computational Caches. In *Proceedings of the 6th International Systems and Storage Conference (SYSTOR)*. ACM, 2013.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7237–7256, 2020.

Yanai Elazar and Yoav Goldberg. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. Linear Adversarial Concept Erasure. In *International Conference on Machine Learning (ICML)*, 2022.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *International Conference on Machine Learning (ICML)*, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational Similarity Analysis — Connecting the Branches of Systems Neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do Wide Neural Networks Really Need to Be Wide? A Quantitative Analysis of Neural Network Width via CKA. In *International Conference on Learning Representations (ICLR)*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2022.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing Models with Task Arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. $\beta$-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations (ICLR)*, 2017.

Hyunjik Kim and Andriy Mnih. Disentangling by Factorising. In *International Conference on Machine Learning (ICML)*, 2018.

Cian Eastwood and Christopher K. I. Williams. A Framework for the Quantitative Evaluation of Disentangled Representations. In *International Conference on Learning Representations (ICLR)*, 2018.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. Emergent Analogical Reasoning in Large Language Models. *Nature Human Behaviour*, 7(9):1526–1541, 2023.

Melanie Mitchell. Abstraction and Analogy-Making in Artificial Intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101, 2021.

Qwen Team. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2025.

Jonathan Frankle and Michael Carlin. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, and others. Model Soups: Averaging Weights of Multiple Fine-tuned Models Improves Accuracy without Increasing Inference Time. In *International Conference on Machine Learning (ICML)*, 2022.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*, 2020.