The Activation Geometry Program: Twelve Papers on the Mathematical Structure of Neural Network Representations

Jeremy McEntire[1]

March 2026

Abstract

Over twelve papers, we have developed a systematic account of how neural network activation spaces encode, separate, and mix domain-specific and structural information. The program began with a practical question—can training compute be reduced by exploiting trajectory structure?—and arrived at a theoretical result: the *terminal measurement limit*, which establishes that linear interventions at the output layer cannot invert nonlinear mixing at intermediate layers. This paper traces the dependency graph across the series, identifies three falsifiable predictions that the framework makes for future work, and delineates where the framework breaks.

1 The Dependency Graph

The twelve papers form a directed acyclic graph of implications. Each paper's central finding serves as a premise for subsequent work.

1.1 Foundation: Training Has Exploitable Structure (Papers I–II)

Paper I (Leap+Verify) demonstrates that training trajectories contain detectable stable and chaotic regimes via activation fingerprint similarity, and that weight states can be predicted during stable regimes with 60–90% acceptance at 7B scale.

Paper II (Ensemble Collapse) shows that independently initialized training runs converge to the same activation manifold (cosine similarity > 0.999), establishing that the representational geometry is determined by the data and architecture, not initialization.

Implication: The activation space has a stable, reproducible geometry that can be studied as a mathematical object rather than a training artifact.

1.2 Geometry: Activations Encode Separable Structure (Papers III–V)

Paper III (Constellation Composition) demonstrates that domain-specialist models can be composed at the parameter level using activation fingerprints as a geometric index,

[1] Correspondence: jmc@cageandmirror.com

achieving 93.3% cross-domain win rate. The key mechanism is two-stage normalization; joint normalization achieves 0%.

Paper IV (Structural Transfer) establishes the central decomposition: INLP separates domain content from structural patterns (hierarchical, causal, constraint, evidential) with double dissociation. Shape survival 98.7–100% across four scales.

Paper V (Capability Manifold Surveillance) repurposes the geometric structure as a security primitive. Systematic distillation detection achieves AUC 1.000 at 1.5B+ scale; combined detection (systematic and randomized attacks) achieves AUC 0.856–0.905, declining with model scale.

Implication: Domain and structure occupy orthogonal subspaces. Interventions can target one without disturbing the other—in principle.

1.3 Theory: When Does Noise Help? (Paper VI)

Paper VI (Communicative Variance) proves five sufficient conditions (C1–C5) under which stochastic resonance produces net-beneficial effects. C1 (suboptimality) is the gate: noise can only help when the baseline system is suboptimal.

Implication: The SR effects observed in Papers III, VII, and VIII are instances of a general theorem, not isolated phenomena.

1.4 Applications: The Forward Pass Under Perturbation (Papers VII–VIII)

Paper VII (GenAI Is Socially Awkward) applies the framework to social cognition: RLHF suppresses pragmatic inference at 7B, and SR partially rescues it, confirming that instruction tuning reduces noise tolerance that social reasoning depends on.

Paper VIII (Shaped Noise Injection) attempts domain-selective entropy control via noise shaped to INLP directions. Three applications (domain precision, loop breaking, soft guardrails) are tested across ten experimental phases. The central negative result: cross-domain selectivity is fundamentally limited. The response matrix R is invertible but predicted-vs-actual correlation ≈ 0. This establishes the *terminal measurement limit.*

Implication: Linear interventions at the output layer cannot control nonlinear mixing at intermediate layers. The geometry that INLP discovers is real but the intervention path goes through a nonlinear bottleneck.

1.5 Resolution: Where the Nonlinearity Lives (Papers IX–XII)

Methodological note: Papers VIII, IX, and XII each use different selectivity metrics — entropy change ratios, z-score normalized response, and KL divergence ratios, respectively. All three capture the same intuition (self-domain effect relative to cross-domain bleed) but are not directly comparable in magnitude. Cross-paper comparisons in this section are qualitative unless otherwise noted.

Paper IX (Layer-Resolved Response Tensor) maps selectivity at nine sampled transformer layers in Qwen-2.5 7B. Result: **weak Outcome A**. Mean selectivity peaks at layer 10 ($\bar{s} = 0.57$) and declines toward both input and terminal layers. No layer achieves $\bar{s} > 1.0$. Domain asymmetry is sigma-dependent: at $\sigma = 0.05$, code and science INLP directions produce positive selectivity; at $\sigma = 0.2$, medical and legal directions become selective while code/science invert. This reversal complicates any description of "inherently selective" domains — the pattern depends on perturbation magnitude, not just geometry. The terminal measurement limit from Paper VIII extends across layers: no single layer achieves high mean selectivity, though the formal generalization is limited by the fact that Paper IX measures forward selectivity only (not R^{-1} inversion at intermediate layers).

Paper X (Spectral Geometry of the Forward Pass) computes Jacobian-vector products for INLP directions through consecutive layer pairs. Both mechanistic hypotheses are rejected: INLP directions are **not** preferentially amplified (INLP/random amplification ratio = 0.991 across all layers), and PCA-INLP alignment peaks at terminal layers (mean $|\cos\theta| = 0.038$ at layer 27), not intermediate layers where selectivity peaks. The forward pass operates as an isotropic amplifier, treating INLP directions identically to random directions. This confirms the concentration barrier (Paper XI) as the sole constraint: without spectral selectivity in the Jacobian, only the geometric overlap k/d_{eff} determines achievable domain selectivity.

Paper XI (The Concentration Barrier) proves that the INLP variance fraction $\leq k/d_{\text{eff}}$ via Cauchy-Schwarz, where d_{eff} is the participation ratio (effective dimensionality) at the injection layer. This bounds how much activation variance any fixed k-dimensional subspace can capture — a geometric quantity, not directly the behavioral selectivity measured in Papers VIII and IX. Result: the bound holds at all 28 layers. Last-token d_{eff} ranges from 4.7 to 26.0 (mean 19.2), giving bounds of 1.4–7.6. Observed INLP variance fraction ranges 1.28%–12.46%, always below the bound. A critical methodological finding: mean-pooled d_{eff} collapses to 1.0 at layers 3–25 (anisotropy artifact), making the bound vacuous; the correct metric is last-token d_{eff}, matching the pooling method used to compute INLP directions in Paper IV.

Paper XII (SR Channel Capacity) measures information-theoretic limits on domain-specific stochastic resonance. All four domains exhibit inverted-U KL divergence profiles, confirming SR in information space. Total channel capacity is large ($\sim$15 bits) but domain-agnostic. The domain-specific differential — self-domain KL minus mean cross-domain KL — is +1.3 bits (medical), +1.9 bits (legal), −1.0 bits (code), −1.5 bits (science). An approximate bound motivated by the concentration barrier, $C_{\text{domain}} \lesssim \log_2(1 + k^2/d_{\text{eff}}) \approx 2.24$ bits, is consistent with the observed ceiling on domain-specific information, though the bound is heuristic rather than rigorously derived. Domain asymmetry is reproduced in information space: medical and legal are weakly information-selective (IS = 1.09–1.13); code and science are information-anti-selective (IS = 0.91–0.93), consistent with the sigma-dependent pattern from Paper IX at $\sigma = 0.2$.

2 Three Predictions

The framework makes predictions that have not yet been tested:

2.1 Prediction 1: Mid-Layer Injection Achieves Better Selectivity

Paper IX confirmed weak Outcome A: selectivity peaks at layer 10 ($\bar{s} = 0.57$) vs. layer 27 ($\bar{s} = -0.10$). The selectivity improvement is real but modest — $\bar{s} = 0.57$ is a factor of six better than terminal but still below 1.0. The **falsifiable prediction**: on a generation quality benchmark, mid-layer (layer 10) shaped noise should produce measurably more domain-targeted effects than terminal-layer noise, but the absolute selectivity will remain bounded by $k/d_{\text{eff}} \approx 1.8$. The concentration barrier, not the injection layer, is the binding constraint.

2.2 Prediction 2: d_{eff} Predicts Variance Fraction Ceiling

Paper XI's concentration barrier theorem proves that the INLP variance fraction $\leq k/d_{\text{eff}}$. This is a bound on geometric capture, not directly on the behavioral selectivity (z-score ratio) measured in Paper IX. The numerical proximity of Paper IX's peak selectivity values (2.0–2.1) to Paper XI's bound ($\sim$1.8) is suggestive but may be coincidental — the two quantities measure different things. The **falsifiable prediction**: for any model, the INLP variance fraction at any layer is bounded by k/d_{eff} measured at that layer. Whether this geometric constraint *implies* a behavioral selectivity ceiling remains an open derivation. Testable: measure both variance fraction and behavioral selectivity at multiple scales (3B, 7B, 14B) and determine whether the empirical correlation persists.

2.3 Prediction 3: The Selectivity Barrier Is Architecture-Dependent

The concentration barrier depends on how the forward pass distributes activation variance. Architectures with different mixing patterns (e.g., state-space models with linear recurrence, mixture-of-experts with sparse routing, or models with linear attention) should show different

d_{eff} profiles and therefore different selectivity ceilings. The domain asymmetry from Paper IX (code/science selective, medical/legal anti-selective) may also be architecture-dependent. Testable: repeat Papers IX and XI on Mamba, Mixtral, or a model with linear attention.

3 Where the Framework Breaks

3.1 The Terminal Measurement Limit

The core negative result: classification exploits any separable signal, but intervention requires causal signal specific to the target. INLP finds directions that separate domains in representation space, but those directions are not the causal pathways through which domain-specific computation flows. The forward pass transforms INLP directions nonlinearly, scattering perturbation energy across domains.

Paper X makes this precise: the layer Jacobian's amplification spectrum is isotropic (INLP/random ratio = 0.991). The forward pass does not "know" which directions are INLP directions. It amplifies all perturbations equally, with a terminal spike at layers 22–27 that explains Paper VIII's large but unselective effects. Paper XII quantifies the consequence: $\sim$15 bits of perturbation information, but only $\sim$1–2 bits are domain-specific.

3.2 The Legal Direction as Shared Substrate

At 7B, the legal INLP direction has near-zero self-effect (-0.4%) but strong cross-domain bleed (-10.7% on medical). The R^{-1} optimal weight assigns zero self-weight to legal. Legal reasoning is not a domain—it is the *intersection* of domains, a shared formality substrate. This suggests the 4-domain INLP decomposition is not capturing independent features but a mixture of domain-specific and domain-shared components.

3.3 The Concentration Barrier

Paper XI makes this precise. In $d = 3584$ dimensions, any $k = 9$ directions capture a variance fraction bounded by k/d_{eff}. With last-token $d_{\text{eff}} \approx 20$, the bound is approximately 1.8. The INLP directions achieve perfect orthogonality ($\cos < 10^{-9}$), but this is a geometric tautology in high dimensions, not an informational achievement.

The bound is on variance fraction (how much of activation variance lies in the INLP subspace), not directly on behavioral selectivity (the entropy change ratios measured in Papers VIII and IX). The empirical proximity of Paper IX's peak selectivity (2.0–2.1) to Paper XI's bound ($\sim$1.8) is suggestive but not a formal test — the quantities are defined differently. What Paper XI establishes rigorously is that the *geometric footprint* of domain-specific directions is small relative to the full activation space, and Paper X confirms that the forward pass provides no spectral mechanism to amplify that footprint. Together, these constrain the space of possible interventions without deriving a closed-form behavioral selectivity bound.

A methodological subtlety: the bound is vacuous ($k/d_{\text{eff}} = 36$) under mean-pooled activations, where d_{eff} collapses to 1.0 at layers 3+ due to anisotropy. The theorem is only informative when d_{eff} is measured under the same pooling method used to compute the INLP directions (last-token). This sensitivity to pooling method is itself a finding about how activation geometry depends on the measurement frame.

4 Conclusion

The activation geometry program demonstrates that neural network representations have rich, reproducible, and mathematically characterizable structure. The structure supports practical applications (model composition, security monitoring, training acceleration) but also imposes fundamental limits on the precision of geometric interventions. The terminal measurement limit and the concentration barrier are not engineering failures—they are properties of high-dimensional computation.

Paper IX answered the first question: operating at intermediate layers improves selectivity modestly (layer 10 vs. layer 27) but cannot overcome the concentration barrier established in Paper XI. Paper X answered the second: the layer Jacobian provides no spectral shortcut — INLP directions are amplified identically to random directions (ratio 0.991), and PCA alignment is dissociated from selectivity. Paper XII answered the third: domain-specific information content is bounded at ∼2 bits, consistent with the concentration barrier, while total perturbation information (∼15 bits) is domain-agnostic. The inverted-U KL profile confirms SR in information space.

All four results converge on a single conclusion: domain-selective intervention via shaped noise is constrained by the concentration barrier (k/d_{eff} bounds the geometric footprint of domain directions), operating through an isotropic forward pass that provides no directional leverage. The formal connection from geometric constraint to behavioral selectivity ceiling remains an open derivation, but the empirical pattern is consistent: where the geometry is tight, selectivity is low.

This paper synthesizes findings from Papers I–XII in the Activation Geometry series (1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11; 12).

References

[1] McEntire, J. (2026). Paper I: Leap+Verify: Regime-Adaptive Speculative Weight Prediction. *arXiv:2602.19580*.

[2] McEntire, J. (2026). Paper II: Training Once Is Enough: Activation Fingerprint Convergence. *SSRN*.

[3] McEntire, J. (2026). Paper III: Constellation-Indexed Model Composition.

[4] McEntire, J. (2026). Paper IV: Structural Transfer via Activation Space Decomposition.

[5] McEntire, J. (2026). Paper V: Capability Manifold Surveillance.

[6] McEntire, J. (2026). Paper VI: The Source of Creation Is Dysfunction: Communicative Variance.

[7] McEntire, J. (2026). Paper VII: GenAI Is Socially Awkward.

[8] McEntire, J. (2026). Paper VIII: Shaped Noise Injection at Inference Time.

[9] McEntire, J. (2026). Paper IX: Layer-Resolved Response Tensor. [This series]

[10] McEntire, J. (2026). Paper X: Spectral Geometry of the Forward Pass. [This series]

[11] McEntire, J. (2026). Paper XI: The Concentration Barrier. [This series]

[12] McEntire, J. (2026). Paper XII: Channel Capacity of Domain-Specific SR. [This series]