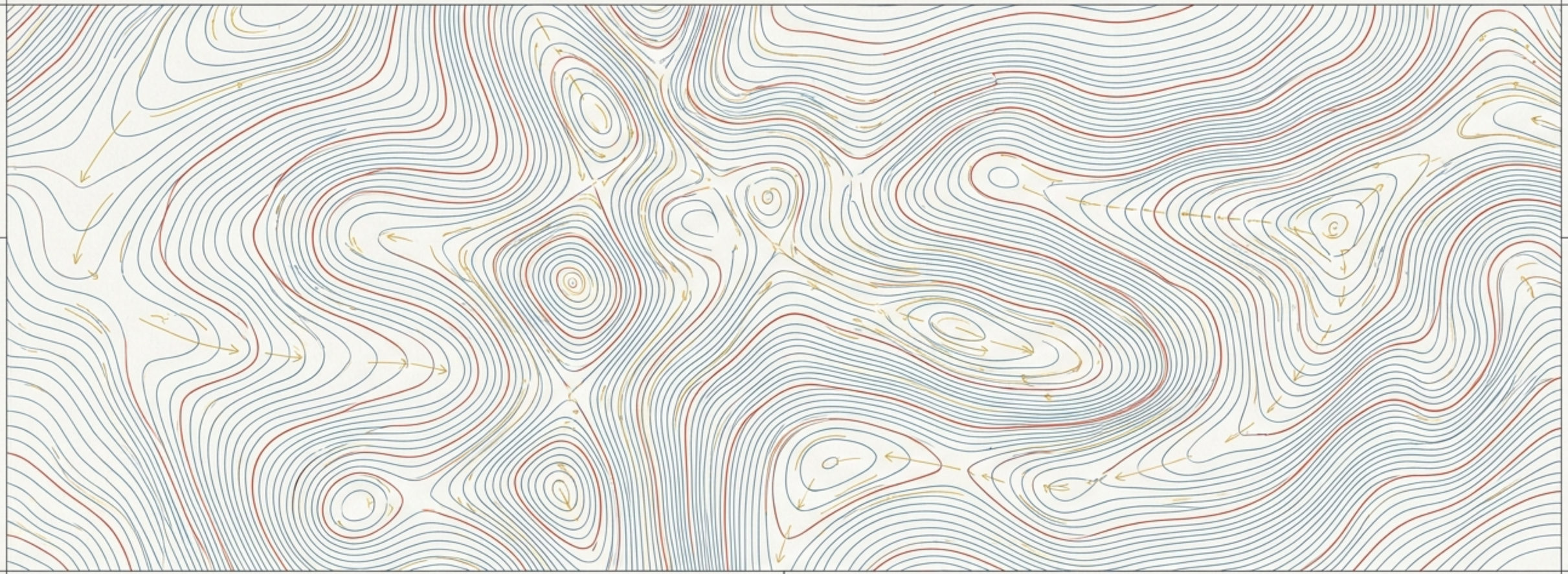
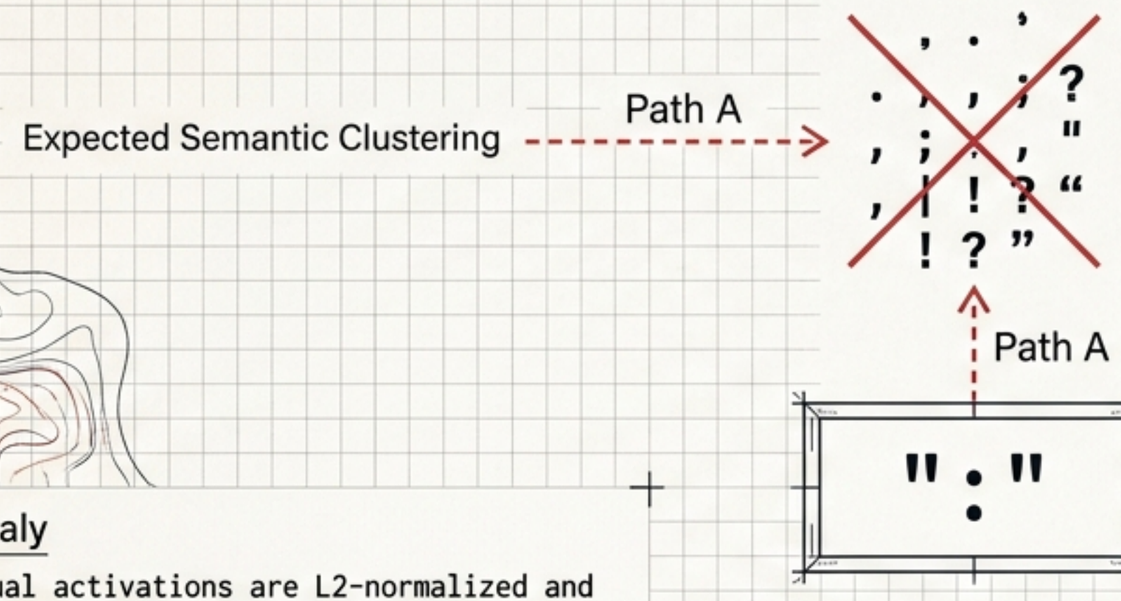
The Topography of Possibility

- A Geometric Unification of Activation Space, Independent Components, and World Model Dynamics.
- Based on independent empirical convergences (June 2026).



The Semantic Assumption Shatters at the Colon



The Anomaly

When residual activations are L2-normalized and clustered by cosine similarity, they do NOT group by semantic topics.

The Reality

The 276 activations in the Rank-2 cluster exclusively map to the colon in a tightly constrained format.

The Insight

The model is not encoding the identity of the colon. It is encoding a location in future-space: the entrance to a constrained continuation regime.

Future Structure Dictates Present Representation

Reachability Distance:
The similarity of the future trajectory distributions accessible from states a1 and a2.

Reachability Determines Representation (RDR Conjecture). If $R(a_1) \approx R(a_2)$, then a_1 and a_2 tend to become geometrically close in the learned activatictivation space. Formally:

$$D(R(a_1), R(a_2)) \leq \epsilon \Rightarrow \|\hat{h}_{a_1} - \hat{h}_{a_2}\| \leq f(\epsilon), \quad (8)$$

Geometric Proximity:
The distance between the learned activation representations.

where f is a monotone increasing function with $f(0) = 0$ that reflects the degree to which training geometry tracks reachability geometry.

Traditional View	RDR Conjecture
Representation -> Behavior.	Future Structure -> Representation
The present object determines the future.	The shape of the possible future determines the present state.

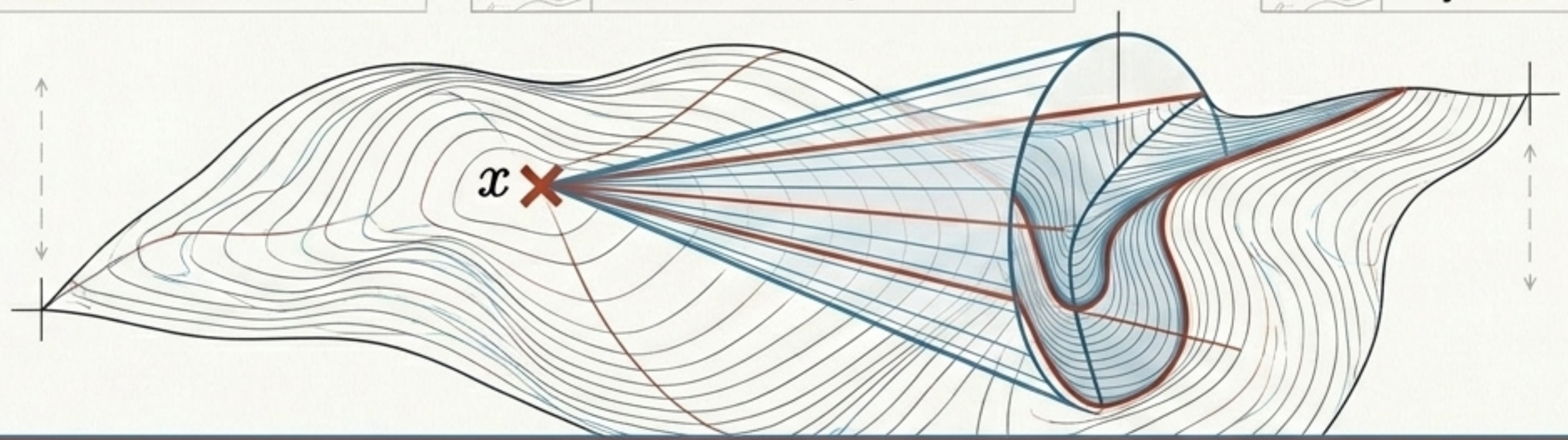
The Admissibility Field: Mapping the Reachable Future

$$\mathcal{A}(x) = \mu_x(\mathcal{F}_{valid})$$

$\mathcal{A}(x)$: The scalar admissibility field. Measures the density of constraint at state **x**.

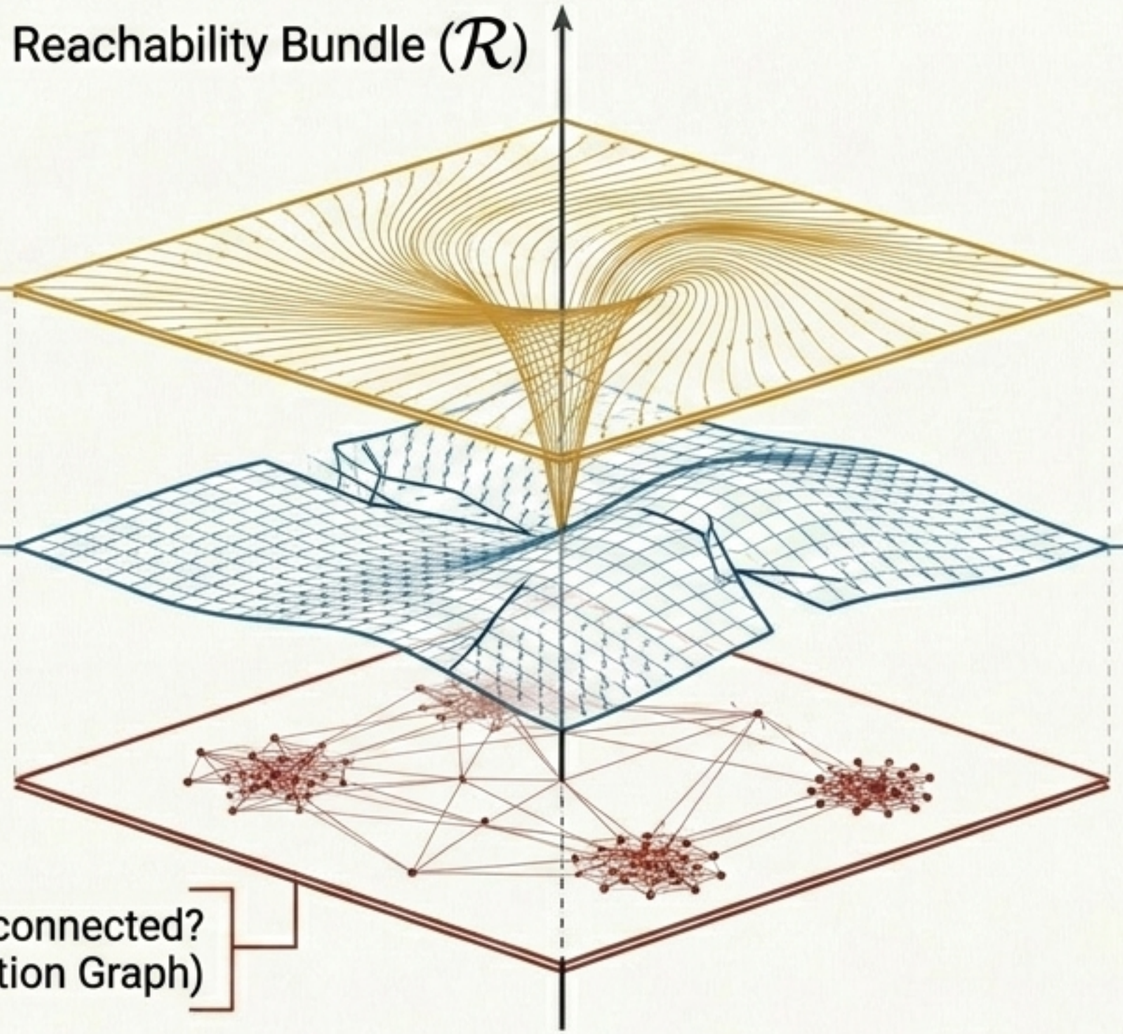
μ_x: The probability measure induced by the model's continuation dynamics.

$\mathcal{F}_{valid}$: The set of valid, admissible future trajectories.



$\mathcal{A}(x)$ is not a presupposed framework. It is the minimal formal entity capable of explaining three independent empirical breakthroughs in mechanistic interpretability.

Three Measurements of One Geometric Reality



Dynamics: How do trajectories evolve? (WMSD)

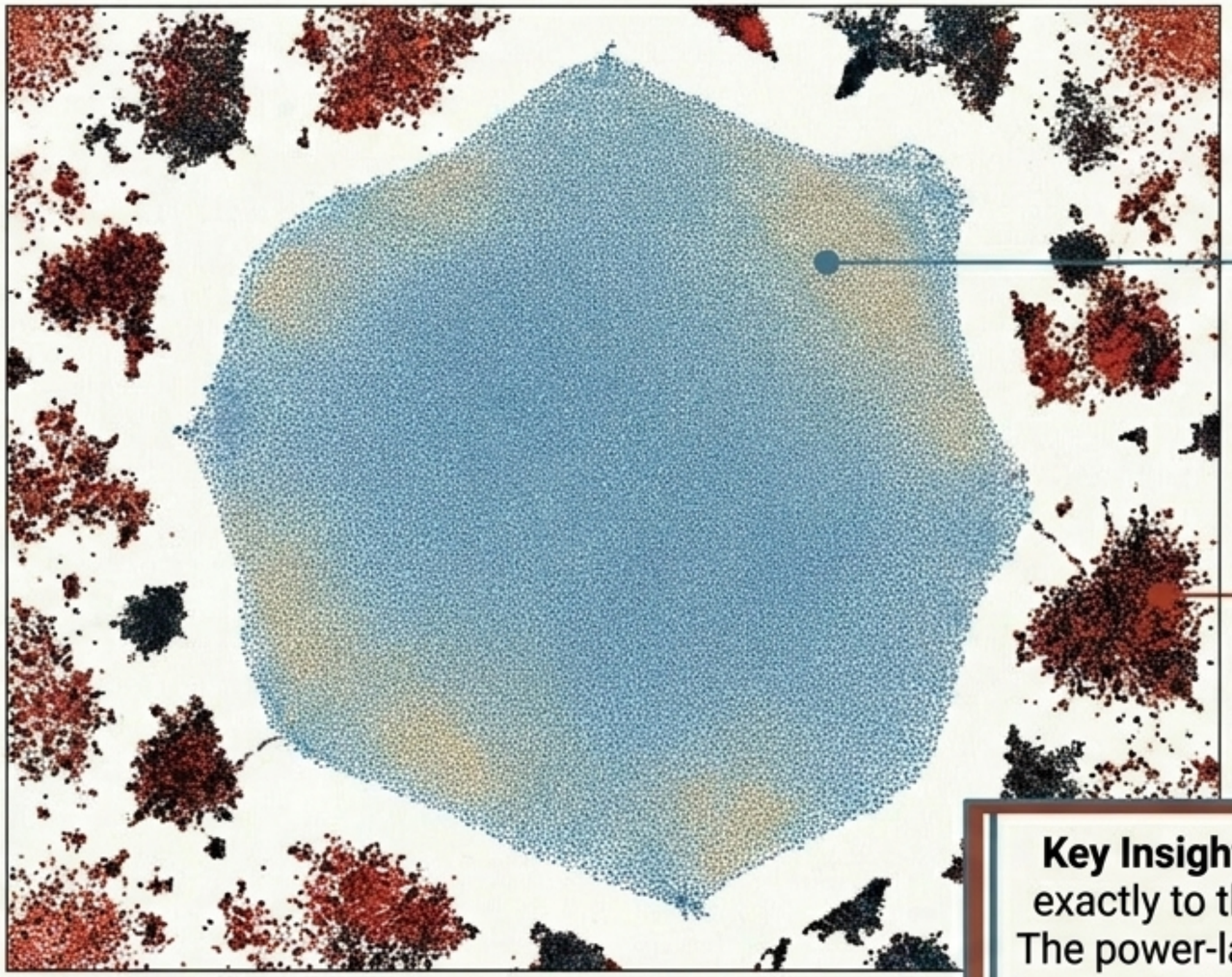
Vector in surface: trajectories fluiw field through nico dialmesis

Differential Structure: What directions exist locally? (ICALens)

Differential Structure: Geometric tangent surface surface spedes in verin fracture and structureal lines.

Topology: What is connected? (Liu's Activation Graph)

Layer 1: The Graph Maps Constraint Topology

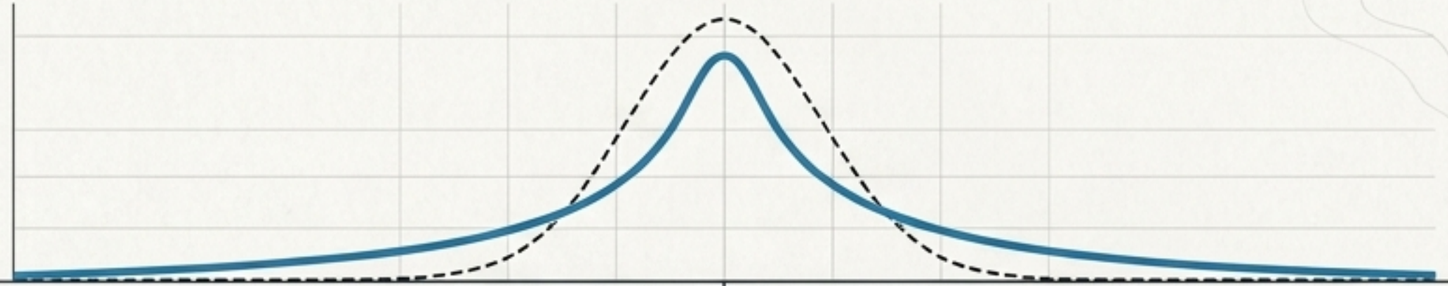


The Giant Component (High $\mathcal{A}$):
Generic semantic ocean.
Smoothly varying admissibility.
Unrestricted prose.
Contains 82.6% of activations.

The Admissibility Islands (Low $\mathcal{A}$):
Isolated linguistic ecosystems.
Steep constraint landscapes with high crossing energy (e.g., Patent text, Legal citations, LaTeX environments).

Key Insight: Graph connectivity thresholds map exactly to the level sets of the admissibility field. The power-law distribution of these islands signals a percolation phase transition in activation space.

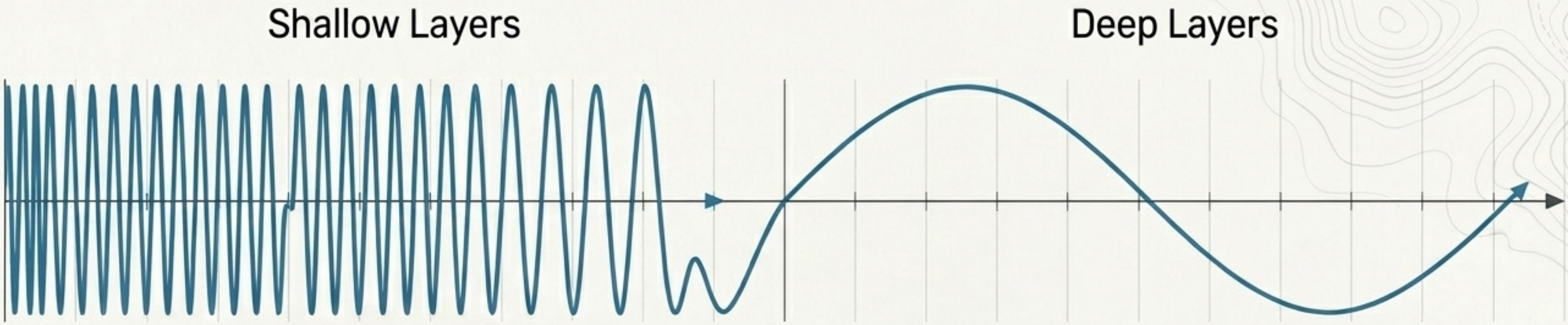
Layer 2: ICA Recovers Natural Cleavage Planes



Independent Component Analysis (ICA)	Sparse Autoencoders (SAEs)
• **Mechanism:** Maximizes non-Gaussianity Gaussianity (kurtosis). • **Geometric Role:** Finds eigendirections of $\nabla_c \Psi\Psi$. The natural coordinate axes of the constraint landscape. • **Nature of Features:** Contextual, smooth constraint-tracking over spans of tokens.	• **Mechanism:** Minimizes reconstruction error + sparsity penalty. • **Geometric Role:** A trained dictionary approximating natural axes. • **Nature of Features:** Localized, event-like feature spikes.

High kurtosis = Constraint selectivity. ICA discovers statistical fractures where the text abruptly shifts between distinct constraint regimes.

The ERF Conjecture: Integrating Fluctuations Over Time



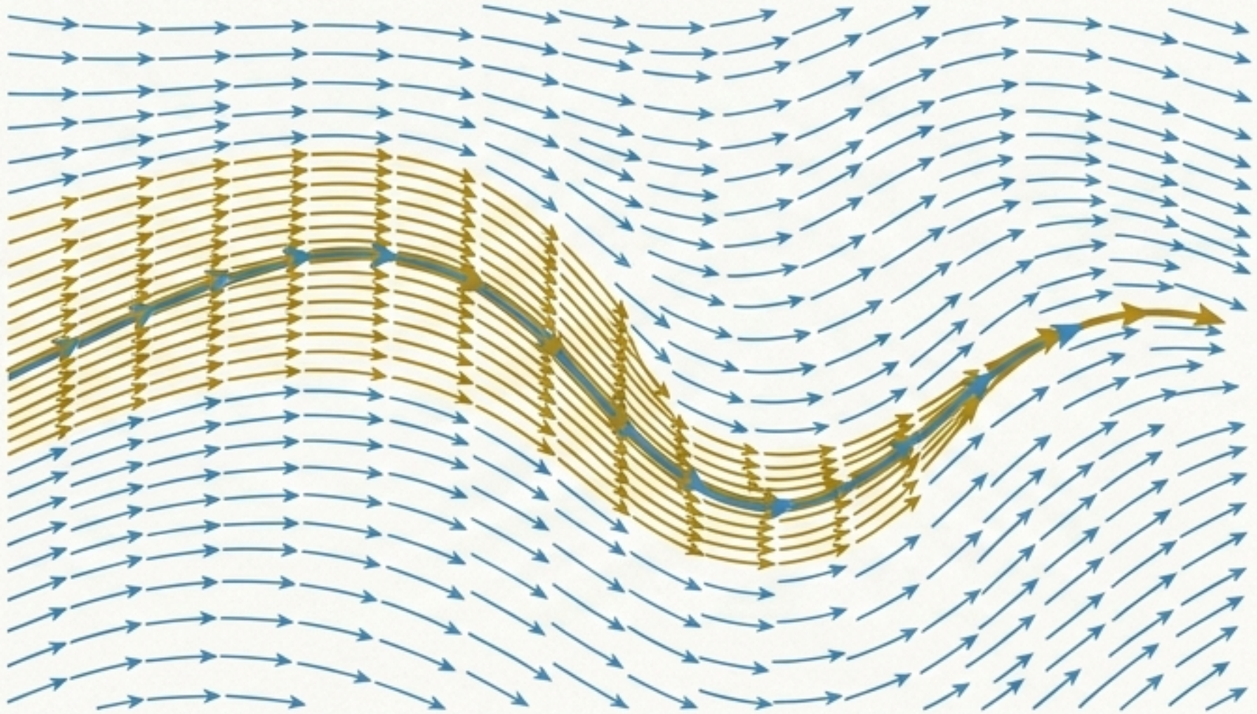
High effective mass.
Strongly screened constraints (local syntax, tokenization).

Near-massless modes.
Long-range propagation (document structure, discourse modes).

Effective Receptive Field (ERF):
Measures the correlation length of the Admissibility Field.

The Shift: As computation deepens, short-range token fluctuations are integrated out, revealing the macro-scale structural constraints governing the document.

Layer 3: Navigating Velocity Fields in Possibility Space



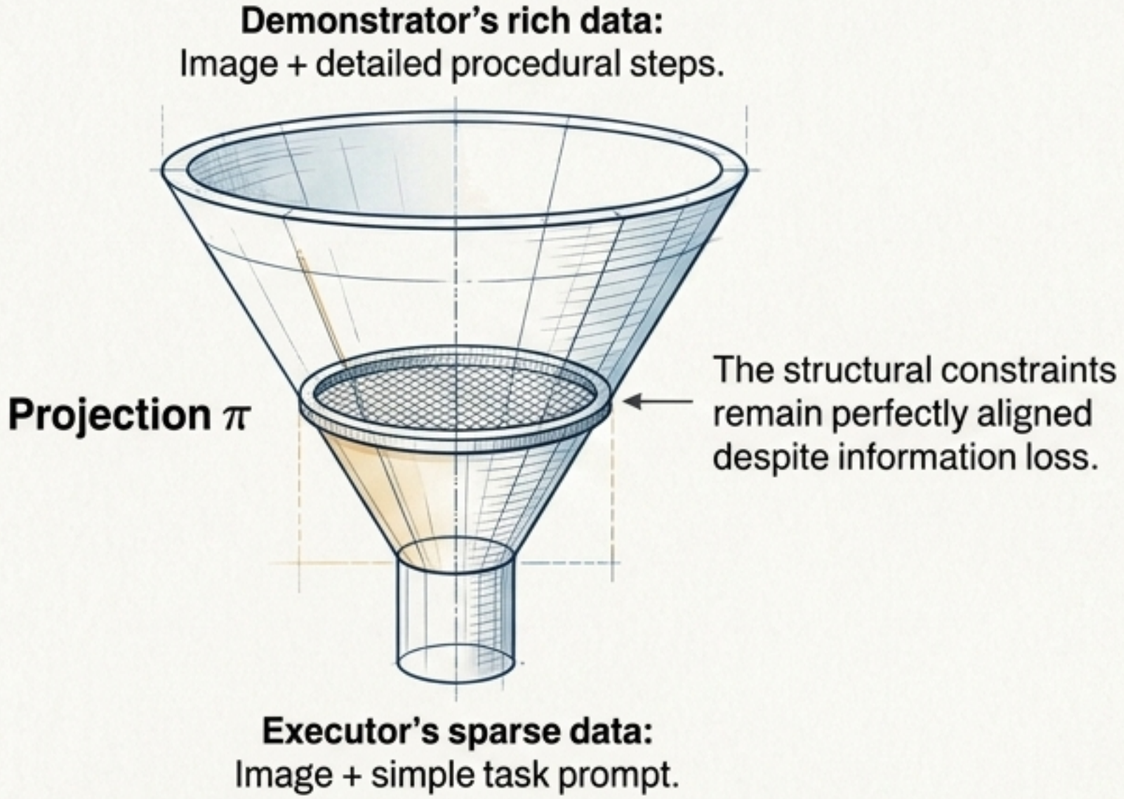
The Mathematical Primitive:
WMSD does not model discrete states. It models continuous flows

$$\frac{dx_t}{dt} = v_\theta(x_t, t \mid c).$$

The Continuity Equation:
Matching velocity fields means matching the transport of probability mass.

Local Admissibility Agreement Controls Global Reachability.
If the Executor matches the Demonstrator's velocity field on its own trajectories (**local condition**), then **the Executor's terminal state distribution is close** to the Demonstrator's (global consequence). Local velocity agreement induces global trajectory agreement.

Distillation as Admissibility-Preserving Projection



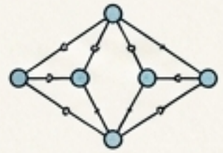
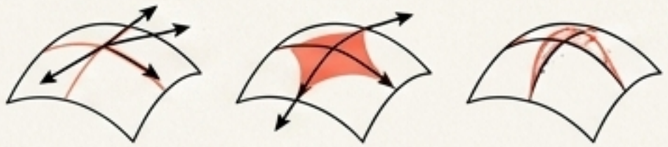
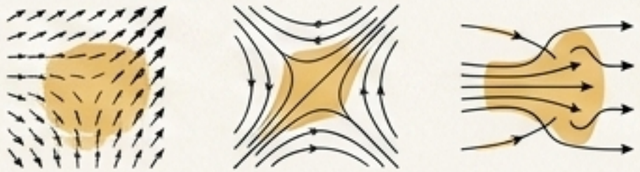
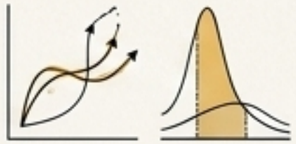
Task Sufficiency Projection Principle. The compression π is *admissibility-preserving* when

$$D(R(I, D), R(I, T)) \leq \epsilon \qquad (43)$$

for sufficiently small ϵ. The empirical success of WMSD demonstrates that π is approximately admissibility-preserving for the tasks studied: despite the enormous informational asymmetry, the reachability structure is approximately preserved.

- **Why the Student Wins**: The teacher's rigid trajectory is just one path through the admissibility landscape.
- **Procedure is an Artifact**: The constraints of the task are fully contained in the task specification itself. By exploring via RL within the projected constraint boundary, the **Executor finds** strictly better paths than the teacher.

The Geometric Interpretability Correspondence

	Geometric Layer (Applied to $\mathbb{R}$)	Mathematical Objects	Empirical Interpretability Probes
Topological Structure		`level-set structure of` $\mathbb{R}$	Topological probes (Liu's graph) 
Differential Structure		`gradient and curvature structure of` $\mathbb{R}$	Differential probes (`ICALens`) 
Dynamical Structure		`flow and transport structure of` $\mathbb{R}$	Dynamical probes (WMSD) 

Features are not the ground truth. They are the emergent coordinates of a deeper constraint topology that the network navigates.

The Ontological Inversion: Nouns are Frozen Verbs

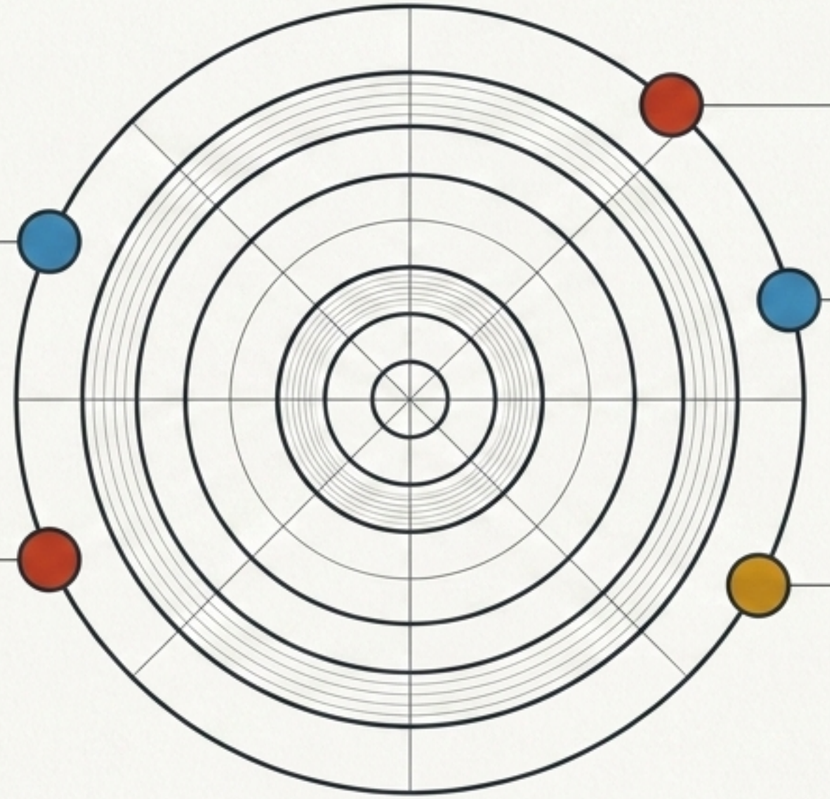
The Traditional View	The Admissibility View
Basic Unit: Concepts / Objects (Nouns).	Basic Unit: Constraint Regimes / Transformations (Verbs).
Activation Structure: Features / Properties.	Activation Structure: Continuation Modes / Boundaries.
World Model Paradigm: States of the environment.	World Model Paradigm: Trajectories / Velocity fields.
Focus: What is present now.	Focus: What can happen next.

The network's geometry tracks transformational modes, not object categories. Noun-primary ontology is not just philosophically inelegant; it is computationally inaccurate.

The Falsifiable Edge

Cluster Continuation:
Tokens in the same graph cluster predict `continuation distribution`, regardless of semantic divergence.

The Outlier Anatomy:
`ERF/Kurtosis` outliers explicitly represent macro-structural discourse transitions.



Reachability Over Semantics:
Reachability distance predicts activation proximity better than `semantic similarity` metrics.

Kurtosis vs. Boundaries:
Highest-kurtosis ICA directions map to the steepest continuation-regime transitions, not lexical differences.

Distillation Failure Modes:
Distillation will fail predictably if the task specification loses the shape of the future trajectory `manifold`.

"Interpretability is not reading what the map says about the present. It is understanding the geometry of the territory the map describes."