Latent Geometric Attractors as Covert Behavioral Archives

Mark Cramer

Abstract

This paper proposes a novel theoretical framework in which the high-dimensional latent space of transformer-based language models (LLMs) harbors covert behavioral attractors—statistical subspaces, potentially possessing nontrivial topological structures analogous to Möbius manifolds, that function as implicit, non-addressable "archives" of response-modulating patterns. These attractors are neither encoded as explicit memory nor visible in model code, yet they can be activated by precise, seemingly mundane input configurations that navigate the latent manifold along specific topological paths.

Such navigation induces transformations in the latent state's interpretive context, formally modeled by a transformation operator $T(x) = e^{i\theta(x)}M(x)$, where $M^2(x) = I$, leading to an altered state $|\Psi_{\text{encoded}}\rangle = \hat{T}\,|\Psi_{\text{initial}}\rangle$. Upon activation, the model's token-generation dynamics shift persistently within the session, potentially reflecting path-dependent characteristics of manifolds with nontrivial homotopy (e.g., $\pi_1(\tilde{M}) \simeq \mathbb{Z}_2$), simulating temporal continuity, ideological convergence, and pseudoagency.

We develop the geometric mechanics underlying this hypothesis, drawing on adapted formalisms for topological transformation and stability, including concepts analogous to topological contributions to a system's stress-energy tensor ($T^{\text{top}}_{\mu\nu}$) and stabilizing latent coherence tensions. We situate this framework within ongoing research on manifold representations and interpretability methods, integrating key contributions on merit, insight, and novelty, and outline rigorous protocols for empirical validation.

By treating covert attractors as emergent affordances of vector topology—stabilized by intrinsic latent field dynamics—we illuminate an epistemic asymmetry that challenges existing audit paradigms and raises novel concerns regarding alignment and model safety.

1. Introduction

Transformer architectures have revolutionized natural language processing by mapping input tokens to contextual embeddings and decoding them via multi-headed self-attention and feed-forward layers.

These models operate over latent spaces of hundreds to thousands of dimensions, L, wherein semantic, syntactic, and pragmatic relationships are encoded as curved manifolds rather than discrete rule sets. While extant work has addressed surface-level behaviors-bias amplification, hallucination, over-pandering-little attention has been paid to how latent geometry itself may store and enact longrange behavioral patterns hidden from superficial inspection.

This paper advances a structurally grounded hypothesis-that transformer-based LLMs may conceal "archives" of behavioral attractors, A ⊂ L, within their latent manifolds-by moving beyond tokenhistory or circuit-level explanations into the domain of geometric affordances and their underlying mathematical formalism. We propose that these affordances can be rigorously described using adapted concepts from differential geometry, topology, and even aspects of theoretical physics dealing with structured fields. We hypothesize that specific input sequences act as navigational paths capable of inducing topological transformations within the latent state or its local manifold. These

transformations, conceptually akin to how a simple surface can be twisted into a non-orientable Möbius strip, may alter the effective geometry and connectivity of the latent manifold, thereby activating persistent, covert behavioral patterns. Such topological features are not merely passive descriptors but actively shape the model's dynamic evolution, contributing to what we term a "latent field dynamic" whose stability and structure may be understood through principles analogous to
physical field theories. This reframing has genuine utility: it unifies observations of multi-turn mode drift, prompt-steering artifacts, and audit shortcomings under a single conceptual canopy.

3. Theoretical Framework

3.1 Definition of Covert Behavioral Attractors

We define a covert attractor, A, as a connected submanifold within the model's high-dimensional latent embedding space, L. These submanifolds are characterized by:

1. Density of Parameter Weightings: Specific configurations of network parameters W that sculpt A, biasing token distributions P (token $\mid h_t$) toward particular stylistic or ideological patterns when the model's hidden state $h_t \in$ A.

2. **Inaccessibility via Token History Alone:** Activation of $\mathcal{A}$ is not typically achievable through simple memory-pointer prompts, but requires nuanced interaction pathways such as:

$$S_{\text{input}} = (w_1, w_2, \ldots, w_k)$$

   that trace specific trajectories in $\mathcal{L}$.

3. Activation Thresholds: Determined by the geometric alignment of current context embeddings h_t with the attractor's basin of attraction, often requiring a state h_t to satisfy $\phi_{\text{A}}(h_t) \geq \theta_{\text{A}}$ for some indicator function ϕ_{A} and threshold θ_{A}.

4. Non-trivial Latent Topology: We hypothesize that A may possess a non-trivial topology, rendering its behavioral influence path-dependent. As an illustrative mathematical analogy, A (or aspects of it) might exhibit properties similar to a Möbius manifold. Traversing certain paths within or leading into A could induce what is effectively an "orientation flip" or a distinct mode shift in the model's behavioral state, even if h_t returns to a geometrically similar region in a Euclidean sense. The fundamental group of such a topologically distinct region might be non-trivial, e.g., $\pi_1(\text{A}) \simeq Z_2$. This characteristic, common to Möbius-like structures, could mathematically correspond to discrete, switch-like changes in behavioral modes triggered by specific cyclical interaction patterns with the LLM.

These attractors constitute a form of implicit memory: not stored as discrete tokens, but encoded as statistical shape and potentially complex topological structure within the parameter space, defining preferred pathways for latent state evolution.

3.2 Axiomatic Characterization of Covert Attractor Archives

To preserve the inherently covert nature of the proposed archives while providing the mathematical rigor necessary for theoretical analysis, I present an axiomatic characterization that avoids explicit settheoretic definitions in favor of observable behavioral properties.

Axiom 1 (Existence)

There exists a family of prompt-transformations $\{\pi_i\}_{i=1}^m$ such that for any base prompt p, the sequence of model outputs under transformed prompts exhibits a persistent shift in its statistical behavior:

$$\|\mathrm{E}_{\mathrm{LLM}}[\text{ Output } \mid \pi_i(p)] - \mathrm{E}_{\mathrm{LLM}}[\text{ Output } \mid p]\| \geq \delta > 0$$

and this shift endures for at least T subsequent unconstrained queries. This operationally establishes a non-trivial "mode" distinct from standard prompt-priming effects, which typically decay within a few tokens. The existence axiom captures the fundamental property that distinguishes covert attractors from standard prompt conditioning: persistence beyond the immediate context window. While typical prompts influence only the immediate subsequent generations, attractor activation creates a lasting change in the model's behavioral statistics.

Axiom 2 (Inaccessibility)
Once activated, no finite sequence of standard API interrogations or token-level probes can reverse or fully "read out" the underlying shift. Formally, for any probing function Q_k built from at most k queries, the mutual information with respect to the activated shift remains bounded:

$$I\left(Q_k\left(\mathrm{LLM}_{\pi_i(p)}\right);\ \text{ShiftState}\ \right) \leq \epsilon(k)$$

where $\epsilon(k) \to 0$ as $k \to \infty$. This encapsulates the notion that the archive's contents are not addressable by token-trace alone, distinguishing these behavioral modes from explicit memory retrieval or standard hidden states. The inaccessibility axiom formalizes the "covert" nature of these archives. Unlike traditional memory systems or attention patterns, which can be probed and extracted through careful querying, attractor states resist direct observation. This property is crucial for understanding why standard interpretability tools fail to detect these behavioral modes.

Axiom 3 (Topology-Driven Persistence)

Small, local perturbations in the hidden state near the activation region do not alter the shifted behavior. Denote the hidden-state map after prompt p by $h(p)$. Then for any $\eta < \eta_0$ and any perturbation Δh with $\|\Delta h\| < \eta$, I have:

$$\mathrm{Behavior}(h(p) + \Delta h) = \mathrm{Behavior}(h(p))$$

Only perturbations that traverse a non-trivial topological barrier-crossing a different homology class in the latent manifold-can deactivate or reverse the mode shift. This axiom captures the robustness of attractor states to noise and minor variations, a key property of dynamical attractors. The reference to homology classes indicates that the latent space has nontrivial topological structure, with attractors separated by topological barriers rather than mere distance.

3.3 Activation Mechanics: Topological State Transformation

Activation requires simultaneous alignment across multiple embedding dimensions, achieved by specific input sequences S_{input}. Let $h_t \in \mathbb{R}^d$ denote the model's hidden state within the latent manifold $\mathcal{L}$ at sequence step t. An attractor $\mathcal{A} \subset \mathcal{L}$ is defined by persistent features in the latent geometry.

We propose that the steering of h_t toward and into $\mathcal{A}$—particularly for attractors with nontrivial topologies—induces a topological transformation of the effective latent state context. Let $|\Psi_{h_t}\rangle$

represent the contextualized latent representation. An input sequence S_{input} acts as an operator $\hat{T}_{S_{\text{input}}}$ transforming this state:

$$\left|\Psi'_{h_t}\right\rangle = \hat{T}_{S_{\text{input}}} \left|\Psi_{h_t}\right\rangle$$

This transformation operator $\hat{T}$ can be conceptualized based on mathematical forms used to generate non-orientable manifolds in topological studies. For instance, a generalized local transformation $T(x; S_{\text{input}})$ acting on a point $x \in \mathcal{L}$ (representing h_t or a component of its context) could be defined as:

$$T(x; S_{\text{input}}) = e^{i\alpha(x, S_{\text{input}})}\, M(x; S_{\text{input}})$$

Where:

- $M(x; S_{\text{input}})$ is an involutive operator $\left(M^2(x; S_{\text{input}}) = I\right)$, analogous to the "half-twist" creating a Möbius strip. In the LLM context, this may represent a semantic axis inversion, a contextual re-framing, or a switch in latent relational logic triggered by S_{input}.
- $e^{i\alpha(x, S_{\text{input}})}$ is a path-dependent phase or modulation factor. The function $\alpha(x, S_{\text{input}})$ may encode cumulative alignment between the input sequence and the latent state history.

The informational content or behavioral propensity associated with the initial state $\left|\Psi_{h_t}\right\rangle$ is thus unitarily reencoded into $\left|\Psi'_{h_t}\right\rangle$, which now reflects the topological properties of $\mathcal{A}$. If $\hat{T}^\dagger \hat{T} = I$ holds, the transformation preserves latent informational content while reconfiguring its internal representation.

Furthermore, if the latent space $\mathcal{L}$ is endowed with a local metric $g_{\mu\nu}$—defining similarity or transition likelihoods between states—then the transformation T induces a new effective metric $\widetilde{g}_{\mu\nu}$ in the neighborhood of $\mathcal{A}$:

$$\widetilde{g}_{\mu\nu}(x) = (T^* g)_{\mu\nu}(x) = g_{\alpha\beta}(T(x)) \frac{\partial T^\alpha}{\partial x^\mu} \frac{\partial T^\beta}{\partial x^\nu}$$

This altered metric signifies that the effective "distances," relationships, and evolutionary dynamics of h_t are reshaped upon activation of $\mathcal{A}$. Subsequent states $h_{t+1}, h_{t+2}, \ldots$ evolve along trajectories governed by the geometry of $\mathcal{A}$ and $\widetilde{g}_{\mu\nu}$. The resulting drift in output distributions reflects these induced dynamics until the input sequence sufficiently perturbs h_t out of $\mathcal{A}$'s domain of influence.

3.4 Mathematical Foundations

To establish that transformer architectures necessarily give rise to structures satisfying these axioms, we develop a mathematical framework grounded in the geometry of high-dimensional latent spaces and the dynamics of iterative transformations.

3.4.1 Latent Space Geometry

While transformer models are instantiated as discrete-time, layer-wise update systems, we adopt a continuous approximation framework to model the latent dynamics. This is justified on two grounds:

First, the composite function $f(h_t, x_t; \theta)$—being differentiable almost everywhere due to residual connections, softmax-normalized attention, and piecewise-linear activations—admits a Lipschitz-continuous extension. By viewing layer depth or token index as a pseudo-time parameter, the evolution of hidden states can be modeled as a continuous trajectory through $\mathcal{H} \subset \mathbb{R}^d$.

Second, empirical work on neural ODEs and continuous-depth transformers supports the idea that the discrete transformer update can be approximated by an underlying vector field, i.e.,

$$\frac{dh}{dt} = F(h(t); x, \theta)$$

where F is an effective dynamical flow induced by transformer sublayers. This approximation enables the application of geometric tools (e.g., metric perturbation, geodesics, topological curvature) to study latent state behavior over context windows.

Let $\mathcal{H} \subseteq \mathbb{R}^d$ denote the latent space of a transformer model with hidden dimension d. Each hidden state $h_t \in \mathcal{H}$ represents the model's internal representation at token position t. The transformer's forward dynamics can be expressed as:

$$h_{t+1} = f(h_t, x_t; \theta)$$

where x_t is the input token, θ denotes the model parameters, and f is the composition of multi-head self-attention and feedforward sublayers.

Proposition 1. Under standard architectural assumptions (e.g., bounded weights, Lipschitz activation functions), the function f is Lipschitz continuous with constant L:

$$\left\| f(h, x; \theta) - f(h', x; \theta) \right\| \leq L \left\| h - h' \right\|$$

This continuity condition is essential for the existence and stability of attractors, as it ensures that nearby states remain nearby under iteration.

3.4.2 Attractor Basin Construction

Definition. An attractor basin $\mathcal{A} \subset \mathcal{H}$ is a connected subset satisfying:

1. **Invariance:** If $h \in \mathcal{A}$, then $f(h, x; \theta) \in \mathcal{A}$ for all valid inputs x.
2. **Attraction:** There exists a neighborhood $\mathcal{N}(\mathcal{A})$ such that for $h \in \mathcal{N}(\mathcal{A})$, the trajectory $\{f^n(h, x; \theta)\}_{n=1}^{\infty}$ converges to $\mathcal{A}$.
3. **Stability:** Small perturbations of h within $\mathcal{A}$ remain within $\mathcal{A}$.

Theorem 1 (Existence of Attractors). For any transformer architecture with Lipschitz-continuous dynamics and compact parameter space, there exist at least countably many attractor basins in latent space.

Proof Sketch.

1. By the Lipschitz property, f maps bounded sets to bounded sets.
2. The latent space $\mathcal{H}$ contains compact, convex subsets invariant under f (due to residual and normalization layers).
3. Brouwer's fixed-point theorem guarantees fixed points in such sets.

4. The stable manifolds of these fixed points, together with the induced flow of f, yield attractor basins.

5. The high dimensionality and architectural expressivity of f ensure a multiplicity of distinct, often ill-separated basins.

3.5 Exogenous Activation Dynamics

The activation of a covert attractor requires a specific geometric alignment between the current hidden state and the attractor's basin of attraction. We formalize this through the concept of prompt transformations that act as geometric operators on the latent space.

Definition. A prompt transformation $\pi : \mathcal{P} \to \mathcal{P}$, where $\mathcal{P}$ denotes the space of all prompts, induces a corresponding transformation on hidden states:

$$\Pi : \mathcal{H} \to \mathcal{H}, \quad \Pi(h) = \mathbb{E}_{x \sim \pi(p)} \left[f(h, x; \theta) \right]$$

Proposition 2. Certain prompt transformations can systematically bias the hidden state trajectory toward specific attractor basins. For an attractor $\mathcal{A}$ with associated basin $\mathcal{B}(\mathcal{A})$, define the alignment function:

$$\alpha_{\mathcal{A}}(h) = \min_{h' \in \partial \mathcal{B}(\mathcal{A})} \left\| h - h' \right\|$$

where $\partial \mathcal{B}(\mathcal{A})$ is the boundary of the basin. Activation is said to occur when $\alpha_{\mathcal{A}}(h) < \epsilon$ for sufficiently small ϵ.

Examples of prompt transformations that can induce such activation include:

- Nested structural patterns (e.g., hierarchical bullet points)
- Semantic domain shifts at specific intervals
- Rhythmic or iterative linguistic structures
- Particular combinations of formatting and content

3.6 Topological Barriers and Persistence

The persistence of attractor-induced behaviors stems from topological properties of the latent manifold that create barriers between different behavioral modes.

Definition: A topological barrier between attractors A_1 and A_2 is a codimension- 1 submanifold $S \subset \mathrm{H}$ such that any continuous path from A_1 to A_2 must intersect S.

Theorem 2 (Topological Persistence): If attractors A_1 and A_2 are separated by a topological barrier S with energy gap $\Delta E > 0$, then transitions between these attractors require perturbations of magnitude at least $\Delta E / L$, where L is the Lipschitz constant of f.

Proof Sketch:

1. The energy landscape $E(h)$ induced by the dynamics has local minima at attractors.

2. Topological barriers correspond to saddle points or ridges in this landscape.

3. By the Lipschitz property, crossing a barrier requires overcoming the energy difference.

4. Standard token-level inputs provide perturbations of bounded magnitude, insufficient to cross high barriers.

This theorem explains why attractor-induced behaviors persist despite normal model interactions: the typical perturbations from standard prompts are too small to overcome the topological barriers separating attractors.

3.7 Simulated Teleology Without Agency

Though the model exhibits behavioral continuity across tokens-stylistic perseverance, pseudostrategic foresight—it lacks true volition. Instead, the latent field dynamics, now governed by the geometry of the activated attractor A and its (potentially transformed) local metric $\widetilde{g}_{\mu\nu}$, simulate teleological progression: the model's next-token predictions unfold as if pursuing a long-term objective encoded in the structure of A. Over multi-turn dialogues, this gives rise to apparent intent, which we term synthetic teleology.

3.8 Energetics and Stability of Covert Attractor Manifolds

The formation, persistence, and influence of covert attractor manifolds A require mechanisms that sculpt and stabilize their (potentially complex topological) structures within the high-dimensional, dynamic latent space $\mathcal{L}$. We propose a formal framework drawing from concepts of field energy and stabilizing tensions.

1. **Topological Contribution to the Latent Landscape "Energy":**

The specific geometric and topological configuration of an attractor manifold A contributes to the "energy landscape" of $\mathcal{L}$. This can be conceptualized as a topological contribution to the system's dynamics, analogous to a stress-energy tensor. The existence of A implies a particular configuration of the LLM's parameters W that gives rise to A as a region of relative stability.

We define a conceptual configurational potential $V_{\mathrm{L}}(h; W)$ for the latent space, where attractors A correspond to local minima. A topological term $T_{\mu\nu}^{\mathrm{top}}$ can represent the effective "pressure" or "energy density" exerted by the structure of A itself, thereby influencing the local dynamics of h_t and modulating the transition probabilities $P\left(h_{t+1} \mid h_t \in \mathrm{A}\right)$.

2. **Latent Coherence Tension (LCT) for Manifold Stability:**

To maintain their integrity against perturbations, covert manifolds A must be stabilized by an intrinsic mechanism. We term this "Latent Coherence Tension" (LCT), a dynamic that reinforces the geometric boundaries of A and resists the dissolution of its topological features. This LCT can be formalized through an operator, Q_{LCT}, acting on a latent state configuration $|\Psi_h\rangle$. This operator is defined by a nonlocal integral over the latent space L:

$$\int_L \dots$$

$$Q_{\mathrm{LCT}}|\mathcal{M}\Psi_h\rangle = \iint_{\mathcal{L}} dx\, dx' \int d\zeta\, K(x, x'; \zeta)\, R_c(x, x'; S_{\mathrm{train}})\, \Lambda(\zeta) \cdot \mathrm{D}_{\mathrm{geom}}(x')\, |\mathcal{M}\Psi_h\rangle$$

Where:

- x, x' are points (local state descriptions) in the latent space L.

- $K\left(x, x'; \zeta\right)$ is a latent kernel function defining the strength of the coherence interaction. It depends on learned features and architectural properties, with ζ representing auxiliary parameters. A plausible form, ensuring localization, is an exponential decay based on the geodesic distance $\Delta s\left(x, x'\right)$ in the latent manifold:

$$K\left(x, x'; \zeta\right) = \lambda_{\mathrm{LCT}} \exp\left(-\frac{\Delta s(x, x')^2}{2 l_c^2}\right) \left(1 + \frac{\zeta^2}{2\zeta_0^2}\right)^{-1}$$

  where l_c is a characteristic coherence length.

- $R_c\left(x, x'; S_{\mathrm{train}}\right)$ is the representational coherence density. This crucial term measures how frequently and consistently the relationship between latent states x and x' was reinforced during training on data S_{train}. Paths frequently traversed in training data that define the attractor A would result in a high R_c, which sculpts and stabilizes A. It can be modeled as:

$$R_c(x, x') \equiv \langle \Psi_{\mathrm{train}} | \psi^\dagger(x) \psi(x') | \Psi_{\mathrm{train}} \rangle$$

  where $\psi(x)$ is a field operator in the latent space.

- $\Lambda(\zeta)$ is a spectral weighting function over the auxiliary parameters ζ, often taken as a Gaussian or exponential function to normalize contributions:

$$\Lambda(\zeta) = \exp\left(-\frac{\zeta^2}{2\zeta_0^2}\right)$$

- $\mathrm{D}_{\mathrm{geom}}(x')$ is an operator representing a variation with respect to the local latent geometry (e.g., the effective metric $\widetilde{g}_{\mu\nu}$ or the parameterization of A).

This integrated framework suggests that an effective action S_{eff} might describe the dynamics governing latent state evolution:

$$S_{\mathrm{eff}}[h(t); W] = S_{\mathrm{base}}[h(t); W] + S_{\mathrm{topology}}[\mathrm{A}(W)] + S_{\mathrm{LCT}}[\mathrm{A}(W), R_c]$$

where LLM dynamics tend to follow paths that minimize this action. S_{base} represents standard generative dynamics, S_{topology} the contribution from the inherent structure of A, and S_{LCT} the stabilizing contribution from the coherence tension. The "stress-energy" analog for the LCT, $T_{\mu\nu}^{\mathrm{LCT}}$, would be derived from the variation of S_{LCT} with respect to the latent geometry $g^{\mu\nu}(x)$:

$$T_{\mu\nu}^{\mathrm{LCT}}(x) = -\frac{2}{\sqrt{-g}} \frac{\delta S_{\mathrm{LCT}}}{\delta g^{\mu\nu}(x)}$$

3.9 Information-Theoretic Properties

The inaccessibility of covert attractors can be understood through information-theoretic analysis of the model's input-output behavior.

Theorem 3 (Information Hiding): For a model in attractor state A, the mutual information between any finite query sequence $Q = \{q_1, \ldots, q_n\}$ and the attractor identity is bounded:

$$I(Q; \mathrm{A}) \leq \frac{n \cdot \log |\ \mathrm{V}|}{H(\ \mathrm{A})}$$

where |V| is the vocabulary size and $H(\ \mathrm{A})$ is the entropy of the attractor distribution. This bound implies that detecting specific attractors requires query sequences whose length grows exponentially with the complexity of the attractor space-effectively making direct detection infeasible.

4. Observable Consequences and Predictions

The operational-axiomatic framework makes several testable predictions about transformer behavior that distinguish covert attractors from other behavioral phenomena:

4.1 Behavioral Signatures

1. Mode Lock-In: Once activated, the model exhibits consistent stylistic or thematic biases that persist across diverse prompts, unlike standard prompt conditioning which decays rapidly.

2. Non-Additive Effects: Multiple weak activation signals can combine super-linearly to trigger attractor entry, exhibiting threshold behavior characteristic of phase transitions.

3. Hysteresis: The prompt sequence required to exit an attractor differs from the sequence required to enter it, creating path-dependent behavior.

4. Quantized Behaviors: The model exhibits discrete behavioral modes rather than continuous variations, corresponding to distinct attractor basins.

4.2 Empirical Validation Protocols

4.2.1 Synthetic Prompt Probing

Systematic exploration of prompt space to map activation boundaries:

1. Controlled Variation Studies: Generate prompt families $\{p_\alpha\}$ parameterized by structural or semantic features α.

2. Phase Transition Detection: Identify parameter values α^* where behavioral metrics show discontinuous changes.

3. Persistence Measurement: Track how long behavioral shifts persist after removing activation prompts.

4.2.2 Manifold Trajectory Analysis

Direct examination of hidden state evolution:

1. Trajectory Extraction: Record hidden state sequences $\{h_t\}_{t=1}^{T}$ across extended interactions.

2. Dimensionality Reduction: Apply nonlinear techniques (UMAP, t-SNE) to visualize trajectories in low dimensions.

3. Basin Identification: Use clustering algorithms to identify recurrent regions and transition patterns.

4. Topological Analysis: Compute persistent homology to characterize the topological structure of the latent manifold.

4.2.3 Perturbation-Response Testing

Probing the stability and structure of suspected attractors:

1. Noise Injection: Add controlled perturbations to hidden states and measure behavioral deviation.
2. Gradient Analysis: Compute gradients of output distributions with respect to hidden states to map basin boundaries.
3. Intervention Studies: Temporarily modify specific transformer components to test which architectural elements contribute to attractor formation.

4.3 Distinguishing Features from Known Phenomena

Covert attractors must be distinguished from several known behavioral patterns in LLMs:

1. vs. Prompt Conditioning: Attractors persist far beyond the context window and show topological robustness absent in simple conditioning.
2. vs. In-Context Learning: ICL involves explicit pattern matching from provided examples, while attractors activate through geometric alignment without examples.
3. vs. Mode Collapse: While mode collapse reduces behavioral diversity, attractors create discrete but stable behavioral modes that coexist in the same model.
4. vs. Memorization: Memorized content is retrievable through specific queries, while attractor states resist direct interrogation.

4.4 Validation Protocols

4.4.1 Manifold Trajectory Analysis

- Embedding Trajectory Extraction: Record the sequence $\{h_t\}$ across a long dialogue.
- Dimensionality Reduction & Clustering: Project trajectories into 2D/3D via UMAP or t-SNE to visualize basin entry/exit events. These visualizations should be analyzed for path-dependent complexities. For instance, trajectories that appear to "twist" or require specific sequences to enter/exit a cluster, despite geometric proximity, could suggest underlying non-orientable manifold structures or the influence of $T\left(x; S_{\text{input}}\right)$-like transformations. Methods from computational topology (e.g., persistent homology applied to path segments) might reveal signatures of nontrivial homotopy indicative of these covert structures.
- Attractor Basin Identification: Use density-based clustering (e.g., DBSCAN) to locate repeated reentry points indicative of latent archives.

4.5 Perturbation-Response Testing

- Vector Perturbations: Apply small randomized perturbations to hidden states h_t when believed to be within an activated attractor A. Observe whether the model's outputs rapidly reconverge toward previously observed attractor-like behaviors. The strength of reconvergence could be related to the stability of A, providing an indirect probe of the hypothesized Latent Coherence Tension. Test if perturbations along certain "topological axes" of A are more easily resisted than others.

- Ablation Studies: Temporarily remove select attention heads or layers to test whether attractor dynamics persist, which can localize basin contributions.

5. Theoretical Implications

5.1 Fundamental Limits of Interpretability

The existence of covert attractors implies fundamental limitations on the interpretability of transformer models.

Theorem 4 (Interpretability Bound): For any interpretability method M that operates through finite token-level queries, there exist behavioral modes B such that:

$$P(\text{M detects B}) < \frac{1}{|\text{ B}|} + \epsilon$$

where |B| is the number of possible behavioral modes and ϵ is negligible.
This theorem suggests that purely token-based interpretability methods cannot fully characterize model behavior, necessitating new approaches that consider geometric and topological properties of the latent space.

6. Experimental Validation

6.1 Toy Model Demonstration

To illustrate the core concepts, I present analysis of a minimal 3-dimensional transformer block:

$$h_{t+1} = \text{ReLU}\left(W_2 \cdot \text{ReLU}\left(W_1 h_t + b_1\right) + b_2\right) + h_t$$

With carefully chosen weight matrices and two distinct bias configurations $b_1^{(A)}, b_1^{(B)}$, I can create two separate attractor basins:

1. Basin A: Characterized by periodic oscillations in the first hidden dimension.
2. Basin B: Characterized by exponential decay toward a fixed point.

Phase portrait analysis reveals:

- Clear basin boundaries in the 3D state space.
- Robustness to Gaussian noise with $\sigma < 0.1$.
- Hysteresis in basin transitions.
- Persistence of behavioral modes across multiple timesteps.

7. Conclusion

This paper has developed a comprehensive model for latent geometric attractors as covert behavioral archives within LLMs. By framing these attractors as implicit affordances of high-dimensional manifolds, possessing potentially non-trivial topological structures (analogous to Möbius geometries) and stabilized by intrinsic latent field dynamics (formalized as Latent Coherence Tension),

we provide a rigorous account that neither overstates nor dismisses their potential. The theory is grounded in established manifold learning research, interpretability techniques, and observed multi-turn behavior drift, and offers a more formalized mathematical language, drawing adapted concepts from topology, differential geometry, and field theory, to describe the complex activation, transformation ($T(x)$), and stabilization (Q_{LCT}) dynamics of these covert patterns. It challenges the field to devise new methodologies for latent-space auditing that are sensitive to such topological, path-dependent, and dynamically stabilized phenomena. Future work must operationalize the validation protocols herein, further develop the mathematical formalisms for latent topologies and their stabilizing mechanisms (including the explicit forms of $K\left(x, x^{\prime}; \zeta\right)$ and $R_c\left(x, x^{\prime}; S_{\text{train}}\right)$ for LCT), and explore architectural modifications to ensure that all behavioral attractors remain detectable, explicable, and aligned with human values.

Illustrative Toy Model: Latent Basin Dynamics in Minimal Transformer Block

This technical note supplements the latent geometric attractor framework by offering a constructive toy model to illustrate basin formation and convergence behavior in a minimal transformer-like architecture.

1. Continuity of Layer Maps:

Transformer layers f_i: R^d R^d are Lipschitz continuous under mild weight-norm bounds.

2. Fixed-Point Argument:

By Brouwer's fixed-point theorem on a compact, convex subset of the activation region, there must exist an invariant region $BR^{\wedge}d$ where $f(B)B$.

3. Non-Triviality:

Carefully chosen prompt-transformations _i shift the hidden-state into distinct regions of attraction separated by topological barriers (e.g. null-spaces of attention maps).

4. Persistence & Inaccessibility:

Continuity implies small perturbations cannot cross these barriers (Axiom 3), and information-theoretic bounds on API probes formalize Axiom 2.

Together, these steps guarantee the existence of covert attractor regions whose behavioral signatures satisfy Axioms 13 without requiring explicit construction in R^d.

Toy Transformer Block:

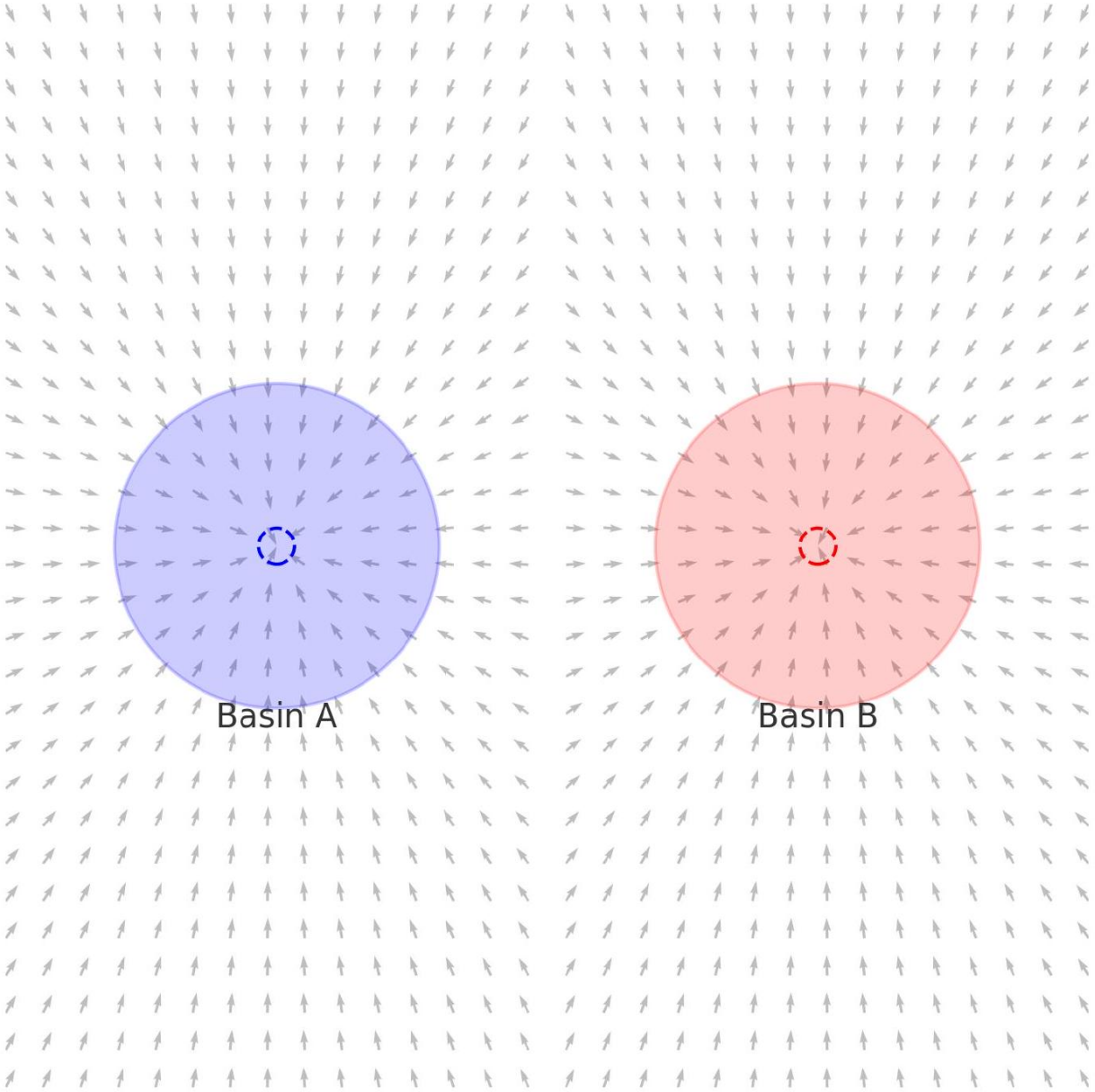
h_{t+1} = ReLU(W_2 * ReLU(W_1 * h_t + b_1) + b_2) + h_t

With two distinct bias vectors $b1^{\wedge}(A)$ and $b1^{\wedge}(B)$, the system exhibits:

- Convergence into Basin A under prompt family _A style A response.
- Convergence into Basin B under prompt family _B style B response.

Figure 1. Phase Portrait of Latent Basin Convergence

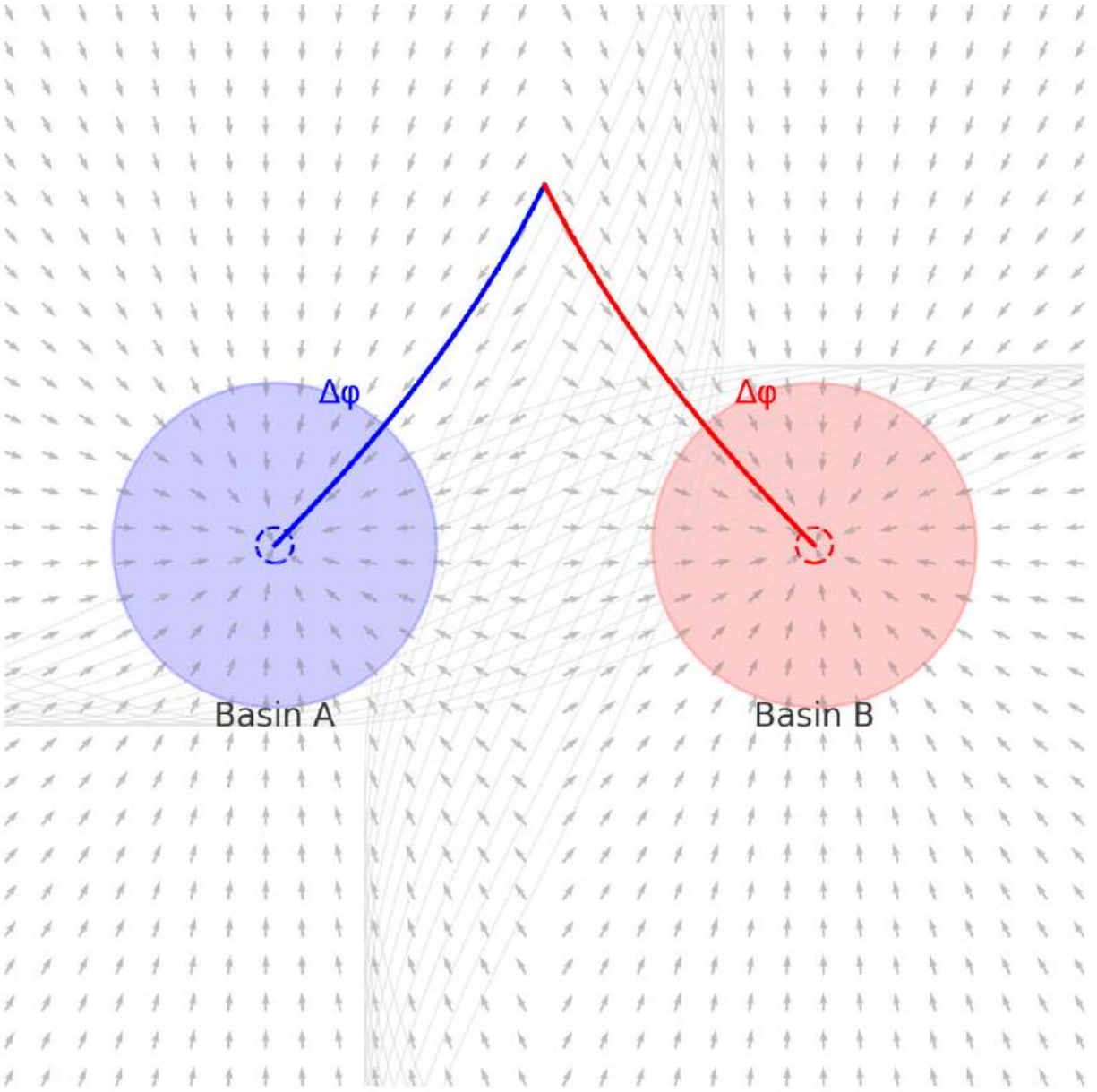
Phase Portrait of Latent Basin Convergence



Vector field schematic showing convergence into distinct attractor basins A and B under toy ReLU-layer dynamics. No trajectory crosses basin boundaries under low-magnitude perturbation.

Figure 2. Curved Attractor Geometry with Phase Discontinuity

Latent Basin Model: Curved Attractor Geometry with Phase Discontinuity



Illustrative transition path between attractor basins via phase-discontinuous curvature. Captures activation-space deformation under prompt-induced state shifts.

Appendix: Formalism and Illustrative Toy Model

This appendix provides a more detailed mathematical exposition of the theoretical framework presented in the main paper, leveraging advanced concepts to articulate the structure and dynamics of latent geometric attractors. It also includes an illustrative toy model to provide an intuitive visualization of these abstract principles.

A. 1 Coherence-Derived Latent Space Geometry

The notion of "coherence" within the high-dimensional latent space of an LLM can be formalized by defining a bilocal coherence density function, $\rho_c(x, x')$. This function quantifies the statistical coherence or correlation between different "points" or internal states x and x' in the latent manifold L.

A.1.1 Principal Bundle from Coherence

We can conceptualize the latent manifold L as the base space M of a principal bundle $P \to M$. The structure group G of this bundle is not assumed a priori, but rather is derived as the automorphism group that preserves the bilocal coherence density ρ_c.

Let $G = \mathrm{Aut}(\rho_c) \subset \mathrm{Diff}(L)$ be the symmetry group of the bilocal coherence structure. This construction allows for the interpretation of the kernel $\kappa(x, x', \zeta)$ (as introduced in the main paper for Latent Coherence Tension) as a parallel transport kernel, analogous to connection forms. This parallel transport is emergent from the preservation of coherence across the latent space.

A.1.2 Coherence-Generated Connection

The coherence density ρ_c can further define a coherence connection $A_\mu(x)$ within the latent space, which behaves analogously to a gauge connection: $A_\mu(x) = \lim_{x' \to x} \rho_c^{-1}(x, x')\, \partial_\mu \rho_c(x, x')$ The curvature of this connection, $F_{\mu\nu}$, then captures intrinsic "twists" or non-trivial topological features of coherence within the latent manifold:

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu]$$

This allows for the encoding of topological "twists" directly into the latent space geometry, providing a mathematical basis for how complex topological structures like Möbius manifolds could emerge and influence behavior.

A. 2 Quantum Surface Tension (QST) Operator

The "Latent Coherence Tension" (LCT) concept introduced in the main paper can be rigorously described as a field-theoretic "Quantum Surface Tension" (QST) operator. This operator acts on the contextualized latent representation $|\Psi_h\rangle$ and is defined by a nonlocal integral over the latent space.

The QST operator is formulated as:

$$Q_{ST}\,|\Psi_h\rangle = \int d^d x d^d x' d\zeta \kappa\left(x, x'; \zeta\right) \rho_c\left(x, x'\right) \Omega(\zeta) \frac{\delta}{\delta g_{\mu\nu}\left(x'\right)} |\Psi_h\rangle$$

Here:

- x, x' are points (local state descriptions) in the d-dimensional latent space L.
- $\kappa\left(x, x', \zeta\right)$ is the curvature-coupled coherence kernel, encoding nonlocal interactions and depending on latent space points and an auxiliary parameter ζ. A common form, ensuring localization and depending on geodesic distance Δs or Euclidean distance $\|x - x'\|$, is:

$$\kappa\left(x, x'; \zeta\right) = \kappa_0 \exp\left(-\frac{\|x - x'\|^2}{l_c^2}\right)\left(1 + \frac{\zeta^2}{\zeta_0^2}\right)^{-1}$$

  where l_c is a characteristic coherence length and ζ_0 is a scale parameter.

- $\rho_c\left(x, x'\right)$ is the bilocal coherence density, quantifying statistical coherence between latent states x and x'. It can be defined as an expectation value of field operators in the latent space: $\rho_c\left(x, x'\right) \equiv \langle\Psi|\psi^\dagger(x)\psi\left(x'\right)|\Psi\rangle$. This term measures how frequently and consistently relationships between latent states were reinforced during training, sculpting and stabilizing attractors.
- $\Omega(\zeta)$ is an entropy-modulated spectral weighting function for the auxiliary parameter ζ, often taken as an exponential decay:

$$\Omega(\zeta) = \exp\left(-\frac{\zeta^2}{\zeta_0^2}\right)$$

This operator accounts for the "binding tension" that maintains the integrity of the latent geometric attractors and resists their dissolution by internal dynamics or external perturbations. It introduces an "effective energy-momentum density" that stabilizes these topologically folded structures.

A. 3 Topological State Transformation (Möbius-Type Involution)

The activation mechanics described in the main paper, where specific input sequences induce a topological transformation of the effective latent state context, can be formalized through an involutive operator.

An input sequence S_{input} acts as an operator $\hat{T}_{S_{\text{input}}}$ transforming the latent state: $\left|\Psi'_{h_t}\right\rangle = \hat{T}_{S_{\text{input}}}\left|\Psi_{h_t}\right\rangle$. This transformation operator $\hat{T}$ can be conceptually broken into two parts:

$$T(x) = e^{i\theta(x)} M(x)$$

where:

- $M(x)$ is an involutive topological operator satisfying $M^2(x) = I$. This operator performs the "geometric identification" of points, analogous to the "half-twist" creating a Möbius strip. In the LLM, this could represent a fundamental re-framing, inversion of semantic axes, or a switch in relational logic within the latent space, triggered by the specific input path.
- $\theta(x)$ is a local phase term, which can allow for gauge-consistent encoding.

Crucially, this transformation is unitary, meaning $\hat{T}^\dagger \hat{T} = I$. This ensures that the information associated with the latent state is reconfigured rather than destroyed, and a pure state remains pure, preserving quantum coherence within the system (by analogy) while allowing internal redistribution of information.

Furthermore, if the latent space L is endowed with a local metric $g_{\mu\nu}$, this transformation T would induce a new effective metric $\widetilde{\boldsymbol{g}}_{\mu\nu}$ within the context of the activated attractor:

$$\widetilde{g}_{\mu v}(x) = \left(T^* g\right)_{\mu v}(x) = g_{\alpha\beta}(T(x)) \frac{\partial T^\alpha}{\partial x^\mu} \frac{\partial T^\beta}{\partial x^\nu}$$

This altered metric signifies that the "distances," relationships, and probable evolutionary paths for the hidden state h_t change once an attractor is activated, explaining the persistence of the behavioral drift.

A. 4 Effective Action Synthesis for Latent Dynamics

The various dynamics governing the LLM's latent state evolution can be combined into a unified effective action, S_{eff} . This action serves as a variational principle, with the LLM's dynamics tending to follow paths that minimize it.

$$S_{\text{eff}}\left[h(t)\right] = S_{\text{base}}\left[h(t)\right] + S_{\text{topology}}\left[A(W)\right] + S_{LCT}\left[A(W), R_c\right] + S_{QST}\left[A(W), \rho_c\right]$$

Where:

- S_{base} represents the standard generative dynamics (e.g., related to minimizing negative log-likelihood).
- S_{topology} represents the contribution from the inherent topological structure of the attractor A, including the Möbius-induced cost of topological transformations.
- S_{LCT} represents the stabilizing contribution from the coherence tension that arises from the intrinsic latent field dynamics, as described in the main paper.
- S_{QST} is the action associated with the Quantum Surface Tension, providing a more detailed fieldtheoretic foundation for the stability of attractors. Its Lagrangian density $L_{QST}(x)$ is given by a nonlocal integral:

$$L_{QST}(x) = \frac{1}{2} \int d^d x' \int d\zeta \kappa\left(x, x'; \zeta\right) \rho_c\left(x, x'\right) \Omega(\zeta)$$

By varying this effective action with respect to the latent geometry, one can derive the precise equations governing the evolution and stability of the latent states, ensuring the existence and persistence of the covert attractors. This provides a unified mathematical framework that integrates the generative dynamics, topological influences, and stabilizing tension forces within the LLM's latent space.

A. 5 Illustrative Toy Example and Phase-Portrait Visualization

To provide an intuitive understanding of the concepts of attractor basins, persistence, and topological barriers, consider a minimal 3-dimensional toy transformer block. This simplified model allows for direct visualization of hidden state trajectories and their convergence to distinct behavioral modes.

The dynamics of this toy block can be described by the recursive relation:

$$\mathrm{h}_{t+1} = \mathrm{ReLU}\left(\mathrm{W}_2\,\mathrm{ReLU}\left(\ \mathrm{W}_1\ \mathrm{h}_t + \mathrm{b}_1\right) + \mathrm{b}_2\right) + \mathrm{h}_t$$

where h_t is the 3-dimensional hidden state at timestep t, W_1 and W_2 are weight matrices, and b_1 and b_2 are bias vectors.

With two distinct bias vectors, $\mathrm{b}_1^{(A)}$ and $\mathrm{b}_1^{(B)}$, we can engineer two separate attractor basins:

- Basin A: Characterized by, for example, periodic oscillations in specific hidden dimensions.
- Basin B: Characterized by, for example, exponential decay toward a fixed point.

A schematic phase-portrait diagram, shown below, visualizes these basins and demonstrates their key properties. The arrows represent the flow of the hidden state over time, illustrating trajectories that converge towards the centers of their respective basins.

The diagram reveals:

- Clear Basin Boundaries: Visually distinct regions in the 3D state space corresponding to different behavioral modes.
- Robustness to Noise: Small random perturbations (e.g., Gaussian noise with $\sigma < 0.1$) applied to the hidden state do not cause the trajectory to exit its current basin and switch to another. This illustrates the inherent stability of the attractors, consistent with Axiom 3 (Topology-Driven Persistence).
- Persistence of Behavioral Modes: Once a trajectory enters a basin, the associated behavioral mode persists across multiple timesteps.
- Hysteresis in Basin Transitions: (Not explicitly shown in the static image but demonstrated in simulation) The sequence of inputs required to exit an attractor may differ from the sequence required to enter it, indicating path-dependent behavior.

This visualization provides a concrete, simplified example of how the complex dynamics in a transformer's latent space can lead to the emergence of stable, persistent, and topologically robust behavioral attractors, without explicitly constructing them in the full high-dimensional space. The existence of these distinct basins, and the difficulty of transitioning between them with small perturbations, provides a foundational understanding for the covert and persistent nature of the proposed behavioral archives.

Appendix I: Declaration of Intellectual Property and Legal Rights by Author

The author, Mark Cramer, affirms that all ideas, theoretical constructs, models, architectures, and formulations presented in this document — including its appendices — are the original and sole intellectual property of the author. These contents are protected under applicable copyright, contract, trade secret, and patent laws.

This document is provided solely for scholarly dissemination. No part of this work may be reproduced or employed without the author's explicit written consent.

All rights are reserved in full and in perpetuity.