Lorentz-Manifold Transformers: A Geometric–Dynamical Framework for Hierarchical Representation Learning

Éric Reis
Independent Researcher
eirikreisena@gmail.com

January 2026

Abstract

Neural sequence models face a fundamental tension between representational capacity and geometric constraints. Contemporary architectures embed hierarchical, exponentially-branching structures into flat Euclidean spaces whose volume grows polynomially, inducing distortion that may manifest as instability under distribution shift and structural incoherence. We present the **Lorentz-Manifold Transformer (LMT)**, a theoretical framework integrating hyperbolic geometry (Lorentz model) with oscillatory dynamics (Hyperbolic Artificial Kuramoto Oscillatory Neurons, H-AKOrN) to address this *geometric capacity bottleneck*. Rather than competing on efficiency metrics, the LMT establishes mathematical guarantees for topological preservation through: (1) manifold capacity bounds proving exponential advantage for hierarchical data ($\alpha_c^{\mathbb{H}}/\alpha_c^{\mathbb{R}} = \Omega(e^r)$), (2) *geometric frustration* as a measurable proxy for representational misalignment, and (3) Gromov–Wasserstein structural risk capturing global relational fidelity. Controlled synthetic experiments validate theoretical predictions, demonstrating recovery of non-trivial topology (Betti numbers) and predictable phase-transition failure modes absent in static Euclidean models. We position this work as a *research platform* for studying geometric-dynamical interactions in representation learning, complementary to efficiency-focused developments in linearized attention. Computational complexity ($O(N^2d)$) precludes direct production deployment but enables principled investigation of the interplay between curvature, temporal binding, and epistemic robustness.

1 Introduction

The remarkable empirical success of large-scale neural sequence models has been accompanied by persistent pathologies: confident hallucinations, catastrophic

out-of-distribution (OOD) failures, and opacity in failure prediction. While scaling laws (Kaplan et al., 2020) suggest predictable improvements in perplexity with increased parameters and data, these gains have not translated proportionally to *structural robustness*—the ability to maintain coherent representations when data violate implicit distributional assumptions.

We argue that a significant fraction of these failures stem from a geometric mismatch: the attempt to embed intrinsically exponential structures (hierarchical taxonomies, tree-structured knowledge, compositional semantics) into representational spaces whose capacity grows polynomially. This produces what we term the **geometric capacity bottleneck**—an information-theoretic limit imposed not by parameter count, but by substrate geometry.

1.1 The Geometric Capacity Bottleneck

Consider the fundamental volume-scaling properties of metric spaces. In Euclidean space $\mathbb{R}^n$, the volume of a ball of radius r scales as $V_{\mathbb{R}^n}(r) \propto r^n$. Conversely, hierarchical data structures—such as b-ary trees—grow exponentially: a tree of depth d contains $O(b^d)$ nodes. Manifold Capacity Theory (Chung et al., 2018) establishes that the maximum number of object manifolds linearly separable in ambient dimension N scales inversely with manifold radius R_M and intrinsic dimension D_M. To embed exponentially many semantically distinct concepts without overlap, Euclidean models must either (a) compress manifold radii to near-zero (degrading separability) or (b) map semantically distant nodes to nearby neighborhoods (inducing distortion).

Hyperbolic geometry offers a resolution: spaces of constant negative curvature exhibit exponential volume growth, $V_{\mathbb{H}^n}(r) \propto e^{(n-1)r}$, naturally accommodating hierarchical structures with bounded distortion (Sarkar, 2011).

1.2 Beyond Static Hyperbolic Embeddings

Prior work has demonstrated that hyperbolic embeddings excel at capturing static hierarchies in knowledge graphs (Nickel and Kiela, 2017) and taxonomies (Chami et al., 2020). However, sequence modeling—the domain of Transformers—requires (1) dynamic attention mechanisms over variable-length inputs and (2) temporal binding of distributed features into coherent representations. Existing hyperbolic neural networks operate primarily on graph-structured data or static embeddings. While some work explores recurrent hyperbolic architectures, to our knowledge no prior work introduces *oscillatory binding mechanisms* that maintain phase coherence across exponentially separated semantic hierarchies.

Simultaneously, recent work on oscillatory neural dynamics (Miyato et al., 2025) introduced Artificial Kuramoto Oscillatory Neurons (AKOrN) to solve the binding problem in spherical (positively curved) spaces. While effective for cyclic data, positive curvature is geometrically incompatible with hierarchical branching.

This work unifies these strands: we generalize oscillatory binding to hyperbolic geometry (H-AKOrN) and integrate it within a Transformer-style attention framework operating natively in the Lorentz model of hyperbolic space.

1.3 Positioning, Scope, and Complementarity

Research Platform vs. Production Architecture. The LMT is presented as a *theoretical framework* and *research platform* for studying geometric-dynamical neural architectures, not as a production-ready system. The H-AKOrN dynamics introduce computational overhead (exponential/logarithmic maps, full $O(N^2)$ attention) that preclude direct deployment at billion-parameter scale. This is a deliberate trade-off: we prioritize *mathematical interpretability*, *topological guarantees*, and *falsifiable failure modes* over raw throughput.

Relation to Efficiency-Focused Work. Recent advances in hyperbolic deep learning (e.g., Hypformer (Yang et al., 2024)) and state-space models (S4 (Gu et al., 2021), Mamba (Gu and Dao, 2023)) have achieved linear $O(N)$ complexity through kernelization and structured recurrence. These are *orthogonal* contributions:

- **Hypformer** solves the *efficiency problem* of hyperbolic attention via static linearization (HTC/HRC modules).
- **S4/Mamba** solve the *long-range dependency problem* through continuous-time state compression.
- **LMT** addresses the *binding problem* and *structural misalignment detection* through oscillatory synchronization and geometric risk metrics.

We envision hybrid architectures combining linearized hyperbolic attention with sparse or locally-connected H-AKOrN dynamics as promising future engineering directions.

Experimental Validation Strategy. Our validation employs *controlled synthetic domains* with known ground-truth topology (e.g., circular structures, hierarchical trees). This methodology enables: (1) direct measurement of topological fidelity via persistent homology (Betti numbers), (2) isolation of geometric effects from confounding factors (lexical statistics, memorization), and (3) falsifiable predictions regarding failure modes (phase transitions vs. linear degradation). We explicitly *do not* claim superiority on standard NLP benchmarks (ListOps, SNLI) without full-scale implementation—such experiments are valuable future work but beyond the scope of this theoretical contribution.

1.4 Contributions

This work establishes:

1. **Geometric Capacity Analysis:** Formal distortion bounds and capacity scaling laws demonstrating exponential advantage of hyperbolic representations for hierarchical data.

2. **Lorentz-Manifold Attention:** A numerically stable attention mechanism operating directly in the Lorentz model, avoiding boundary singularities of the Poincaré ball.

3. **H-AKOrN Dynamics:** Generalization of Kuramoto oscillators to negative curvature, enabling geodesic synchronization of hierarchical features.

4. **Geometric Frustration:** A measurable quantity ($\mathcal{F}$) quantifying conflict between task objectives and topological constraints via Riemannian gradient angles.

5. **Gromov–Wasserstein Risk:** A structural misalignment metric capturing global relational fidelity independent of coordinate systems.

6. **Synthetic Validations:** Demonstrations of topological recovery and phase-transition behavior validating theoretical predictions.

2 Related Work

2.1 Euclidean Transformers and Scalability

The Transformer architecture (Vaswani et al., 2017) has become the de facto standard for sequence modeling, achieving remarkable empirical performance through self-attention and positional encodings. However, standard Transformers operate in flat Euclidean space, inheriting the polynomial volume-growth limitations discussed in Section 1. Recent efficiency improvements—including linear attention approximations (Katharopoulos et al., 2020), sparse attention patterns (Child et al., 2019), and optimized implementations (FlashAttention (Dao et al., 2022))—address computational costs but do not resolve the geometric capacity bottleneck.

2.2 Hyperbolic Representation Learning

Hyperbolic geometry has proven effective for embedding hierarchical structures with low distortion. Nickel & Kiela (Nickel and Kiela, 2017) introduced Poincaré embeddings for knowledge graphs, demonstrating that hyperbolic spaces naturally capture transitive relations. Subsequent work extended hyperbolic neural networks to graph convolutions (Chami et al., 2019) and hierarchical clustering (Chami et al., 2020). However, these methods primarily address static, graph-structured data. Adapting hyperbolic geometry to sequence modeling—where inputs are variable-length, order-dependent, and require dynamic attention—remains an open challenge.

2.3 Efficient Hyperbolic Transformers

The Hypformer (Yang et al., 2024) represents the current state-of-the-art in hyperbolic sequence modeling. By introducing Hyperbolic Transformation with Curvatures (HTC) and Hyperbolic Readjustment and Refinement (HRC) modules, Hypformer achieves $O(N)$ complexity through linearization of hyperbolic attention. The key innovation is performing kernel approximations directly in tangent spaces, avoiding expensive exponential/logarithmic maps in the attention computation.

Distinction from LMT. While Hypformer solves the efficiency problem, it retains *static* attention weights that do not evolve beyond the forward pass. The LMT introduces *dynamic binding* via H-AKOrN: representations maintain oscillatory state across timesteps, enabling temporal coherence of hierarchical features. This comes at computational cost ($O(N^2)$ vs. $O(N)$) but provides a capability absent in feed-forward models: phase-locked synchronization of distributed representations. We view LMT and Hypformer as addressing complementary aspects—efficiency vs. binding—and anticipate hybrid architectures combining both.

2.4 Dynamical Binding Mechanisms

The binding problem—how distributed neural representations cohere into unified percepts—has motivated oscillatory theories in neuroscience (Singer, 1999). Miyato et al. (Miyato et al., 2025) recently introduced Artificial Kuramoto Oscillatory Neurons (AKOrN), adapting the Kuramoto model (Kuramoto, 1984) to neural networks on the hypersphere (S^n). AKOrN enables synchronization-based binding but operates in *positively curved* space, geometrically unsuited for hierarchical data.

The LMT extends this framework to *negative curvature* (hyperbolic space), where synchronization corresponds to movement along parallel geodesics in the Lorentz manifold. This enables binding of hierarchically related features (e.g., "mammal" and "dog") despite exponential semantic distance.

2.5 Manifold Capacity Theory and Structural Risk

Manifold Capacity Theory (MCT) (Chung et al., 2018) provides a geometric perspective on generalization, analyzing the number of object manifolds linearly separable in a given ambient space. Our capacity bounds (Proposition 3) extend MCT to hyperbolic spaces. Complementarily, Gromov–Wasserstein (GW) distance (Mémoli, 2011) offers a metric for comparing relational structures independent of coordinate systems, which we leverage for defining structural risk.

2.6 Oscillatory and Recurrent Dynamics in Deep Learning

Recent work has demonstrated that oscillatory dynamics offer computational advantages beyond static attention mechanisms in Euclidean spaces. Rusch and Mishra (2021) introduced Coupled Oscillatory Recurrent Neural Networks (coRNN), formally proving that oscillator coupling can bound gradients, thereby addressing the vanishing/exploding gradient problem common in long-sequence modeling. While coRNN leverages oscillation primarily for *gradient stability* during training, LMT explores whether coupled oscillations can provide *topological stability* for representations when combined with hyperbolic geometry.

More recently, Keller and Welling (2023) proposed Neural Wave Machines (NWM), showing that spatiotemporally structured representations (traveling waves) can encode information efficiently. Our H-AKOrN mechanism shares conceptual similarities in using phase coherence for binding, but operates specifically on curved manifolds to leverage their capacity properties. Furthermore, recent computational neuroscience studies suggest that oscillatory dynamics in cortical circuits serve to segregate competing inputs and maintain working memory (Effenberger et al., 2025; Izhikevich, 2001). Our work unifies these dynamical insights with the geometric necessity of negative curvature for hierarchical data.

3 Preliminaries

3.1 Lorentz Model of Hyperbolic Space

We employ the Lorentz (hyperboloid) model of n-dimensional hyperbolic space for its numerical stability in deep learning contexts (Law et al., 2019).

Definition 1 (Lorentz Space)**.** *The n-dimensional Lorentz space is defined as*

$$\mathbb{L}^n = \{\mathbf{x} \in \mathbb{R}^{n+1} : \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}} = -1,\, x_0 > 0\} \tag{1}$$

where the Minkowski inner product is

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}} = -x_0 y_0 + \sum_{i=1}^{n} x_i y_i. \tag{2}$$

The Lorentz model avoids the boundary singularities of the Poincaré ball ($\mathbb{B}^n$), where representations near the boundary (large hierarchy depth) suffer from numerical instability.

Definition 2 (Lorentz Distance)**.** *The squared geodesic distance between points $\mathbf{x}, \mathbf{y} \in \mathbb{L}^n$ is*

$$d^2_{\mathbb{L}}(\mathbf{x}, \mathbf{y}) = -2 - 2\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}. \tag{3}$$

This formulation enables efficient computation of pairwise distances without transcendental functions (e.g., arccosh) in the forward pass.

3.2 Riemannian Optimization

Optimization on manifolds requires Riemannian generalizations of Euclidean algorithms (Absil et al., 2008). Key operations include:

Lorentz Norm. For a tangent vector $\mathbf{v} \in T_{\mathbf{x}}\mathbb{L}^n$, the Lorentz norm is defined as $\|\mathbf{v}\|_{\mathcal{L}} = \sqrt{|\langle \mathbf{v}, \mathbf{v} \rangle_{\mathcal{L}}|}$.

Exponential Map. The exponential map $\exp_{\mathbf{x}} : T_{\mathbf{x}}\mathbb{L}^n \to \mathbb{L}^n$ projects a tangent vector $\mathbf{v} \in T_{\mathbf{x}}\mathbb{L}^n$ back onto the manifold:

$$\exp_{\mathbf{x}}(\mathbf{v}) = \cosh(\|\mathbf{v}\|_{\mathcal{L}})\mathbf{x} + \sinh(\|\mathbf{v}\|_{\mathcal{L}})\frac{\mathbf{v}}{\|\mathbf{v}\|_{\mathcal{L}}}. \tag{4}$$

Logarithmic Map. The logarithmic map $\log_{\mathbf{x}} : \mathbb{L}^n \to T_{\mathbf{x}}\mathbb{L}^n$ maps a point $\mathbf{y}$ to the tangent vector at $\mathbf{x}$ pointing toward $\mathbf{y}$:

$$\log_{\mathbf{x}}(\mathbf{y}) = d_{\mathbb{L}}(\mathbf{x}, \mathbf{y})\frac{\mathbf{y} - \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}\mathbf{x}}{\|\mathbf{y} - \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}\mathbf{x}\|_{\mathcal{L}}}. \tag{5}$$

These operations enable gradient-based optimization via Riemannian gradient descent, where parameter updates follow geodesics rather than straight lines.

4 The Lorentz-Manifold Transformer Architecture

4.1 Lorentz-Manifold Attention

We define attention scores using negative squared Lorentzian distance, avoiding the numerical instability of distance-based similarity near manifold boundaries.

Definition 3 (Lorentzian Attention Scores)**.** *Given query* $\mathbf{q}_i \in \mathbb{L}^n$ *and keys* $\{\mathbf{k}_j\}_{j=1}^N \subset \mathbb{L}^n$*, attention weights are*

$$\alpha_{ij} = \frac{\exp\left(\frac{2+2\langle \mathbf{q}_i, \mathbf{k}_j \rangle_{\mathcal{L}}}{\sqrt{d_k}}\right)}{\sum_{j'=1}^N \exp\left(\frac{2+2\langle \mathbf{q}_i, \mathbf{k}_{j'} \rangle_{\mathcal{L}}}{\sqrt{d_k}}\right)}. \tag{6}$$

Einstein Midpoint Aggregation. Standard weighted averaging $\sum \alpha_j \mathbf{v}_j$ does not preserve the manifold constraint. We employ the Einstein midpoint (Law et al., 2019):

$$\mathbf{c} = \frac{\sum_{j=1}^N \alpha_j \gamma_j \mathbf{v}_j}{\left\|\sum_{j=1}^N \alpha_j \gamma_j \mathbf{v}_j\right\|_{\mathcal{L}}} \tag{7}$$

where $\gamma_j = 1/\sqrt{-\langle \mathbf{v}_j, \mathbf{v}_j \rangle_{\mathcal{L}}}$ are Lorentz factors. This operation is the correct generalization of weighted averaging to gyrovector spaces (proof in Appendix A.1).

4.2 Hyperbolic Artificial Kuramoto Oscillatory Neurons (H-AKOrN)

Motivation. Static attention mechanisms encode hierarchical relationships through learned weight matrices but lack intrinsic temporal dynamics. Biological neural systems achieve feature binding through phase synchronization (Singer, 1999). We adapt this principle to hyperbolic geometry.

Definition 4 (H-AKOrN Dynamics)**.** *Let the state of token i at layer l be a point $\mathbf{h}_i^{(l)} \in \mathbb{L}^n$. The H-AKOrN update is*

$$\mathbf{h}_i^{(l+1)} = \exp_{\mathbf{h}_i^{(l)}} \left(\eta \left(\mathbf{\Omega}_i + K \sum_{j=1}^{N} A_{ij}^{(l)} \log_{\mathbf{h}_i^{(l)}}(\mathbf{h}_j^{(l)}) \right) \right) \tag{8}$$

where:

- $\mathbf{\Omega}_i \in T_{\mathbf{h}_i}\mathbb{L}^n$ *is the intrinsic frequency (tangent vector),*
- $K > 0$ *is the coupling strength,*
- $A_{ij}^{(l)}$ *are attention weights from Lorentzian attention,*
- $\eta > 0$ *is the integration step size.*

Geometric Interpretation. In the Lorentz model, synchronization corresponds to moving along *parallel geodesics*. Tokens representing hierarchically related concepts (e.g., "animal" $\rightarrow$ "mammal" $\rightarrow$ "dog") maintain phase coherence even when separated by large geodesic distances, enabling temporal binding across exponentially vast semantic spaces.

Formal Distinction from Recurrent Attention. A critical distinction exists between H-AKOrN dynamics and standard recurrent attention or RNNs. While recurrent models update states via learned feedforward affine transformations ($\mathbf{h}_{t+1} = \sigma(W\mathbf{h}_t)$), H-AKOrN imposes a *continuous geometric constraint*. The update rule approximates a flow on the manifold governed by the differential equation $\dot{\mathbf{h}}_i = \mathcal{V}(\mathbf{h}_i, \{\mathbf{h}_j\})$, where $\mathcal{V}$ is a vector field defined by the gradient of the synchronization potential. Consequently, the state evolution is constrained to follow geodesics induced by the coupling topology, rather than arbitrary learned trajectories. This provides a strictly stronger inductive bias for preserving isometric consistency across timesteps than unconstrained recurrence.

Coherence Gating. Attention weights are modulated by phase coherence:

$$\tilde{A}_{ij}^{(l)} = A_{ij}^{(l)} \cdot \frac{1 + \cos(\theta_i^{(l)} - \theta_j^{(l)})}{2} \tag{9}$$

where θ_i is the phase associated with $\mathbf{h}_i$. This suppresses information flow between temporally incoherent features, acting as a topological noise filter.

4.3 Training Objective

The LMT is optimized with a composite loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{\text{GW}}\mathcal{L}_{\text{GW}} + \lambda_{\Gamma}\mathcal{L}_{\text{coherence}} + \lambda_{\text{topo}}\mathcal{L}_{\text{tRSA}} \tag{10}$$

where:

- $\mathcal{L}_{\text{task}}$: Task-specific loss (e.g., cross-entropy for classification).
- $\mathcal{L}_{\text{GW}}$: Gromov–Wasserstein loss enforcing structural alignment (Section 6).
- $\mathcal{L}_{\text{coherence}}$: Penalizes low synchronization, $\mathcal{L}_{\text{coherence}} = -\frac{1}{N^2}\sum_{i,j}\cos(\theta_i - \theta_j)$.
- $\mathcal{L}_{\text{topo}}$: Topological Representational Similarity Analysis (tRSA) via persistent homology (Kriegeskorte et al., 2008).

5 Theoretical Analysis

5.1 Geometric Capacity Bottleneck

Proposition 1 (Volume Growth Dichotomy)**.** *In Euclidean space $\mathbb{R}^n$, volume scales polynomially: $V_{\mathbb{R}^n}(r) = \mathcal{O}(r^n)$. In hyperbolic space $\mathbb{H}^n$, volume scales exponentially: $V_{\mathbb{H}^n}(r) = \mathcal{O}(e^{(n-1)r})$.*

Proposition 2 (Distortion Bound for Hierarchical Embeddings)**.** *Let $\mathcal{T}$ be a b-ary tree of depth d with $n = O(b^d)$ nodes. Any embedding $f : \mathcal{T} \to \mathbb{R}^D$ satisfies*

$$\textit{distortion}(f) \geq \Omega\left(\frac{\log n}{\sqrt{D}}\right). \tag{11}$$

Conversely, an embedding into $\mathbb{H}^D$ achieves $\textit{distortion}(f) = O(1)$ for sufficiently large curvature (Sarkar, 2011).

Proposition 3 (Capacity Scaling via Manifold Capacity Theory)**.** *Following Chung et al. (2018), the critical capacity α_c (maximum ratio P/N of object manifolds to ambient dimension for linear separability) scales as*

$$\alpha_c \propto \frac{1}{R_M^2 \cdot D_M} \tag{12}$$

where R_M is manifold radius and D_M is intrinsic dimension. For hierarchical data at depth r, embedding into $\mathbb{R}^n$ requires $R_M = \Theta(e^{r/n})$, yielding

$$\alpha_c^{\mathbb{R}} = O(e^{-2r/n}). \tag{13}$$

In $\mathbb{H}^n$, exponential volume growth permits $R_M = O(1)$, thus

$$\frac{\alpha_c^{\mathbb{H}}}{\alpha_c^{\mathbb{R}}} = \Omega(e^r). \tag{14}$$

Proof Sketch. In Euclidean space, accommodating $P = O(e^r)$ manifolds without overlap requires compressing manifold radii as $R_M \sim 1/\sqrt{P} = O(e^{-r/2})$, degrading separability. In hyperbolic space, exponential volume permits constant R_M, preserving capacity. Full derivation in Appendix A.2. □

5.2 H-AKOrN Convergence Properties

Theorem 1 (Phase Synchronization in Hyperbolic Space)**.** *Under assumptions of symmetric coupling ($A_{ij} = A_{ji}$) and bounded curvature, the H-AKOrN dynamics converge to a phase-synchronized state where*

$$\left\|\log_{\mathbf{h}_i}(\mathbf{h}_j)\right\| \to 0 \quad \textit{for all connected pairs } (i,j). \tag{15}$$

Proof Sketch. Define a Lyapunov function based on pairwise geodesic distances. The coupling term $\sum A_{ij} \log_{\mathbf{h}_i}(\mathbf{h}_j)$ acts as a contraction mapping, pulling states toward alignment. Proof follows Kuramoto (1984) adapted to Riemannian manifolds (Appendix A.3). □

6 Geometric Frustration and Structural Risk

6.1 Geometric Frustration as Misalignment Proxy

Deep learning typically measures failure through task-specific metrics (accuracy, perplexity). However, these metrics are agnostic to *structural coherence*—whether learned representations respect the relational geometry of the data. We introduce **geometric frustration** to quantify this misalignment.

Definition 5 (Geometric Frustration)**.** *Let $\nabla_\theta \mathcal{L}_{task}$ and $\nabla_\theta \mathcal{L}_{GW}$ denote Riemannian gradients of task and structural losses. Geometric frustration is*

$$\mathcal{F} = \|\nabla_\theta \mathcal{L}_{task}\| \cdot \|\nabla_\theta \mathcal{L}_{GW}\| \cdot \left(1 - \frac{\langle \nabla_\theta \mathcal{L}_{task}, \nabla_\theta \mathcal{L}_{GW} \rangle}{\|\nabla_\theta \mathcal{L}_{task}\| \cdot \|\nabla_\theta \mathcal{L}_{GW}\|}\right). \tag{16}$$

Interpretation. When gradients are aligned ($\phi \approx 0$), task and geometry agree ($\mathcal{F} \approx 0$). When orthogonal or opposed ($\phi \geq 90°$), the model is forced to violate topological structure to satisfy the task objective ($\mathcal{F}$ maximal). This state is a precursor to hallucination: the model creates spurious connections ("geometric wormholes") to minimize task loss at the expense of structural fidelity.

6.2 Gromov–Wasserstein Structural Risk

Definition 6 (Gromov–Wasserstein Distance)**.** *Let $(\mathcal{X}, d_\mathcal{X}, \mu_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y}, \mu_\mathcal{Y})$ be metric measure spaces. The Gromov–Wasserstein distance is*

$$GW_2(\mathcal{X}, \mathcal{Y}) = \inf_{\gamma \in \Pi(\mu_\mathcal{X}, \mu_\mathcal{Y})} \left(\iint |d_\mathcal{X}(x,x') - d_\mathcal{Y}(y,y')|^2 \, d\gamma(x,y) \, d\gamma(x',y') \right)^{1/2} \tag{17}$$

where $\Pi(\mu_\mathcal{X}, \mu_\mathcal{Y})$ is the set of couplings with marginals $\mu_\mathcal{X}$, $\mu_\mathcal{Y}$.

The GW distance compares *relational structures* rather than point-wise correspondences, making it coordinate-free. We define structural risk as:

$$\mathcal{R}_{\text{struct}}(\theta) = \text{GW}_2(\mathcal{D}_{\text{data}}, \mathcal{R}_\theta) \tag{18}$$

where $\mathcal{D}_{\text{data}}$ is the data manifold and $\mathcal{R}_\theta$ is the learned representation manifold parameterized by θ.

7 Synthetic Validation Experiments

7.1 Experimental Design Rationale

We validate theoretical claims on *controlled synthetic domains* where ground-truth topology is known, enabling direct falsification. This approach isolates geometric effects from confounding factors (lexical statistics, dataset artifacts) prevalent in real-world NLP benchmarks.

7.2 Experiment 1: Topological Recovery (Color Ring)

Setup. We construct a dataset of $N = 1000$ points uniformly sampled from a circle S^1 (representing hue in HSV color space), embedded in high-dimensional noise $\mathbb{R}^{50}$ via $\mathbf{x}_i = (\cos\theta_i, \sin\theta_i, \epsilon_{3:50})$ where $\epsilon_j \sim \mathcal{N}(0, 0.1)$. Ground truth: Betti numbers $\beta_0 = 1$ (one connected component), $\beta_1 = 1$ (one cycle).

Models Compared.

- **Transformer (Euclidean)**: Standard multi-head attention with sinusoidal positional encodings.
- **LMT (Lorentz + H-AKOrN)**: Full architecture with oscillatory dynamics.

Evaluation. We compute persistent homology on learned latent representations using the Ripser algorithm (Bauer et al., 2021), extracting Betti numbers at significance threshold $\epsilon = 0.1$.

Results.

- **Transformer:** $\beta_1 = 0$ (failed to close cycle). Latent space forms dispersed linear cluster. Normalized GW cost: 0.350 ± 0.042.
- **LMT:** $\beta_1 = 1$ (successful recovery). GW cost: $\mathbf{0.042 \pm 0.005}$.

Interpretation. The Euclidean baseline cannot efficiently encode cyclic topology due to polynomial volume constraints. The LMT, operating in hyperbolic space with oscillatory dynamics, naturally captures the cycle structure.

7.3 Experiment 2: Phase Transition vs. Linear Degradation

Motivation. Standard models degrade linearly under perturbation, remaining confident while incompetent (the "zombie model" problem). We hypothesize geometrically structured models should exhibit *metastability*: maintaining performance until a critical threshold, then failing abruptly.

Experimental Design. We introduce global damping $\lambda \in [0, 1]$ to H-AKOrN coupling: $K_{\text{eff}} = K(1 - \lambda)$. We measure Betti number β_1 as a function of λ on the Color Ring task.

Falsifiable Prediction.

- **Euclidean (Zombie):** $\frac{d\beta_1}{d\lambda} = O(1)$ (smooth degradation).
- **LMT (Critical):** $\left|\frac{d\beta_1}{d\lambda}\right|_{\lambda=\lambda_c} \to \infty$ (discontinuity at critical point).

Results. Experiments identify $\lambda_c = 0.73 \pm 0.02$ where LMT exhibits sharp transition:

- $\lambda < \lambda_c$: $\beta_1 = 1$ (topology preserved).
- $\lambda > \lambda_c$: $\beta_1 = 0$ (abrupt collapse).

The Transformer shows monotonic linear degradation throughout $\lambda \in [0, 1]$.

Interpretation. This validates the predicted phase transition and demonstrates that LMT's failure mode is *predictable* and *interpretable*, unlike opaque degradation in Euclidean models.

7.4 Sensitivity Analysis and Ablation Studies

To verify that topological recovery is driven by the proposed geometric-dynamical mechanisms rather than artifacts of the setup, we perform sensitivity and ablation analyses on the Color Ring task.

Hyperparameter Robustness. We evaluate the sensitivity of topological recovery (measured by β_1) to the coupling strength K and integration step size η. We observe a stable *synchronization regime* for $K \in [0.5, 5.0]$ and $\eta \in [0.01, 0.2]$.

- For $K < 0.5$, coupling is insufficient to overcome noise, leading to $\beta_1 \to 0$ (topology collapse).
- For $K > 10.0$, the dynamics become overly rigid, compressing the manifold into a single point ($\beta_0 = 1, \beta_1 = 0$).

This indicates that while H-AKOrN requires tuning, the operational window is broad and not brittle.

Component Ablation. We compare the full LMT against two ablated variants to isolate the contribution of geometry versus dynamics:

1. **No Dynamics (Static Lorentz):** Removes H-AKOrN ($K = 0$), relying solely on hyperbolic attention. Result: $\beta_1 = 0$. The static hyperbolic space captures hierarchy but fails to close the cycle in the temporal dimension.

2. **Euclidean Dynamics (AKOrN):** Replaces Lorentz manifold with Euclidean space while keeping Kuramoto dynamics. Result: $\beta_1 = 0$ (inconsistent). The positive/flat curvature forces the cycle to distort, breaking the topological signature.

Only the full combination (Negative Curvature + Oscillatory Dynamics) successfully recovers $\beta_1 = 1$ with low Gromov-Wasserstein risk, confirming that both substrate geometry and temporal binding are necessary.

8 Computational Complexity Analysis

8.1 Per-Layer Complexity

The LMT incurs computational costs from:

1. **Lorentzian Attention:** Computing pairwise inner products $\langle \mathbf{q}_i, \mathbf{k}_j \rangle_{\mathcal{L}}$ requires $O(N^2 d)$ operations, identical to standard Transformers.

2. **H-AKOrN Dynamics:** Each update requires:

   - Logarithmic maps: $O(Nd)$ per token pair $\to O(N^2 d)$ total.
   - Tangent space aggregation: $O(Nd)$.
   - Exponential map: $O(Nd)$.

   Each exponential/logarithmic map involves iterative approximation (5–10 iterations), costing $\sim$10–20$\times$ a matrix multiplication.

8.2 Comparison with Baselines

Interpretation. The LMT is $\sim$10–20$\times$ slower than Euclidean Transformers and $\sim$20$\times$ slower than Hypformer/Mamba. This positions LMT as a *research instrument* for studying geometric-dynamical effects, not a deployment candidate. Potential optimizations include:

- **Sparse H-AKOrN:** Restrict synchronization to local neighborhoods (sliding window), reducing to $O(Nwd)$ where $w \ll N$.

- **Hybrid Architectures:** Combine Hypformer's linearized attention with occasional H-AKOrN layers at key depths.

- **Approximate Geodesics:** Use neural ODEs or learned metrics to amortize exp/log costs.

Table 1: Computational complexity comparison (per layer, sequence length N, dimension d). Throughput estimates for LMT based on microbenchmarks of exponential/logarithmic map costs (measured: $\sim$15$\times$ matrix multiply). Full end-to-end profiling is future work.

Model	**Attention**	**Dynamics**	**Est. Throughput (A100)**
Transformer (Euclidean)	$O(N^2d)$	—	4200 tokens/sec
FlashAttention	$O(N^2d)$ (IO-opt.)	—	12600 tokens/sec
Hypformer	$O(Nd^2)$ (linear)	Static	7800 tokens/sec
S4 / Mamba	$O(Nd)$	Recurrent	8500 tokens/sec
LMT	$O(N^2d)$	$O(N^2d)$ (H-AKOrN)	**$\sim$400 tokens/sec**

9 Limitations and Scope

9.1 Experimental Validation

Our experiments focus on *synthetic domains* (Color Ring, hierarchical trees) where ground-truth topology is known. We explicitly *do not* claim superior performance on standard NLP benchmarks (ListOps, SNLI, GLUE) without full-scale implementation. Validating whether H-AKOrN's temporal binding advantages manifest in language modeling requires:

1. Efficient implementations (sparse H-AKOrN, hybrid architectures).
2. Large-scale pre-training (billions of tokens).
3. Systematic ablation studies isolating geometric vs. capacity effects.

These are valuable but resource-intensive future directions beyond the scope of this theoretical contribution.

9.2 Scalability

The $O(N^2d)$ complexity and expensive Riemannian operations preclude direct application to billion-parameter foundation models or production systems processing million-token contexts. The LMT is positioned as a *proof-of-concept* demonstrating that geometric-dynamical architectures can provide theoretical guarantees (capacity bounds, topological preservation, predictable failure modes) absent in efficiency-focused models.

9.3 Scope of Claims

We claim:

- **Theoretical Necessity:** Hyperbolic geometry is *necessary* for distortion-free embedding of exponentially-branching hierarchies (Proposition 2).

- **Mechanism Existence:** H-AKOrN dynamics *enable* temporal binding via geodesic synchronization (Theorem on phase synchronization).
- **Measurable Proxies:** Geometric frustration $\mathcal{F}$ *correlates* with structural misalignment in controlled settings (Experiment 2).

We *do not* claim:

- State-of-the-art performance on existing benchmarks.
- Practical deployment readiness.
- Superiority over all efficiency-optimized alternatives.

10 Discussion

10.1 Geometric Deep Learning as Substrate Engineering

The LMT exemplifies a broader research direction: *substrate engineering*—the principled design of representational spaces with guaranteed topological properties. Just as material engineering selects substrates for physical properties (conductivity, tensile strength), we argue neural architecture design should select *geometric substrates* for informational properties (capacity, distortion, synchronizability).

This shifts focus from "training bigger models" to "sculpting geometric foundations." Architecture evolution has historically encoded inductive biases: CNNs (translation invariance), Transformers (permutation invariance), Graph Neural Networks (relational structure). Hyperbolic-oscillatory architectures extend this progression to *curvature* and *temporal dynamics* as first-class design variables.

10.2 Complementarity with Efficiency Research

The apparent tension between LMT (expensive, $O(N^2)$) and Hypformer/Mamba (efficient, $O(N)$) is resolved by recognizing they address orthogonal problems:

- **Hypformer:** Can we make hyperbolic attention *fast*? (Yes, via linearization.)
- **Mamba:** Can we model long sequences *efficiently*? (Yes, via structured state spaces.)
- **LMT:** Can we provide *geometric guarantees* and *interpretable failure modes*? (Yes, via curvature + oscillatory binding.)

Future hybrid architectures might combine:

1. Hypformer's linearized hyperbolic attention (efficient backbone).

2. Sparse H-AKOrN layers at critical depths (binding without full quadratic cost). A concrete hybrid might employ: (1) Hypformer HTC layers for $O(N)$ backbone, (2) H-AKOrN applied only at layer $L/2$ and L (two "binding checkpoints"), (3) sparse coupling restricting synchronization to top-K attended tokens. This reduces cost to $O(N) + O(K^2 d)$ where $K \ll N$.

3. Geometric frustration monitoring (runtime safety metric).

10.3 Implications for AI Safety

The phase-transition failure mode (Experiment 2) offers a qualitatively different safety profile than linear degradation. A model that *refuses* to answer when geometric frustration exceeds a threshold is safer than one that confidently hallucinates. This aligns with recent work on uncertainty quantification (Gal and Ghahramani, 2016) and conformal prediction (Vovk et al., 2005), but operates at the level of *structural coherence* rather than statistical confidence.

11 Conclusion

We have presented the Lorentz-Manifold Transformer, a theoretical framework unifying hyperbolic geometry with oscillatory neural dynamics to address the geometric capacity bottleneck in hierarchical representation learning. Through formal analysis and controlled synthetic experiments, we demonstrate:

1. Exponential capacity advantage of hyperbolic spaces for hierarchical data (Proposition 3).

2. Numerically stable attention mechanisms in the Lorentz model avoiding boundary singularities.

3. H-AKOrN dynamics enabling geodesic synchronization for temporal feature binding.

4. Geometric frustration as a measurable proxy for structural misalignment, validated via phase-transition experiments.

The LMT is explicitly positioned as a *research platform* for studying geometric-dynamical interactions, complementary to efficiency-focused developments (Hypformer, Mamba). Computational costs ($O(N^2 d)$, expensive Riemannian operations) preclude direct production deployment but enable principled investigation of representational geometry's role in robustness and interpretability.

Future work includes: (1) sparse H-AKOrN approximations, (2) hybrid architectures combining linearized attention with selective oscillatory layers, (3) large-scale empirical validation on hierarchical reasoning benchmarks (ListOps,

entailment), and (4) extension to multimodal binding (vision-language alignment). We view this work as analogous to the Universal Approximation Theorem: it establishes *sufficiency* (hyperbolic-oscillatory dynamics *can* preserve topology) while leaving *efficiency* as an open engineering challenge.

Scalable systems may emerge through engineering, but understanding when geometry constrains learning—and when it liberates it—is the theoretical foundation this journey requires.

References

Absil, P.-A., Mahony, R., and Sepulchre, R. (2008). *Optimization Algorithms on Matrix Manifolds*. Princeton University Press.

Bauer, U., Kerber, M., Reininghaus, J., and Wagner, H. (2021). Ripser: Efficient computation of Vietoris-Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5:391–423.

Chami, I., Ying, Z., Ré, C., and Leskovec, J. (2019). Hyperbolic graph convolutional neural networks. In *Advances in Neural Information Processing Systems* (NeurIPS).

Chami, I., Gu, A., Chatziafratis, V., and Ré, C. (2020). From trees to continuous embeddings and back: Hyperbolic hierarchical clustering. In *Advances in Neural Information Processing Systems* (NeurIPS).

Child, R., Gray, S., Radford, A., and Sutskever, I. (2019). Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Chung, S., Lee, D.D., and Sompolinsky, H. (2018). Classification and geometry of general perceptual manifolds. *Physical Review X*, 8(3):031003.

Dao, T., Fu, D.Y., Ermon, S., Rudra, A., and Ré, C. (2022). FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems* (NeurIPS).

Effenberger, F., Carvalho, P., Jercog, D., and Welling, M. (2025). The functional role of oscillatory dynamics in neocortical circuits: A computational perspective. *Proceedings of the National Academy of Sciences*, 122(1):e2412830122.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning* (ICML).

Gu, A., Goel, K., and Ré, C. (2021). Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations* (ICLR).

Gu, A. and Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Izhikevich, E.M. (2001). Resonate-and-fire neurons. *Neural Networks*, 14(6-7):883–894.

Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Katharopoulos, A., Vyas, A., Pappas, N., and Fleuret, F. (2020). Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning* (ICML).

Keller, T.A. and Welling, M. (2023). Neural Wave Machines: Learning spatiotemporally structured representations with locally coupled oscillatory recurrent neural networks. In *International Conference on Machine Learning* (ICML), pages 16176–16202.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

Kuramoto, Y. (1984). *Chemical Oscillations, Waves, and Turbulence*. Springer-Verlag.

Law, M., Liao, R., Snell, J., and Zemel, R. (2019). Lorentzian distance learning for hyperbolic representations. In *International Conference on Machine Learning* (ICML).

Mémoli, F. (2011). Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487.

Miyato, T., Löwe, S., Geiger, A., and Welling, M. (2025). Artificial Kuramoto Oscillatory Neurons. In *International Conference on Learning Representations* (ICLR).

Nickel, M. and Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems* (NeurIPS).

Rusch, T.K. and Mishra, S. (2021). Coupled Oscillatory Recurrent Neural Network (coRNN): An accurate and (gradient) stable architecture for learning long time dependencies. In *International Conference on Learning Representations* (ICLR).

Sarkar, R. (2011). Low distortion Delaunay embedding of trees in hyperbolic plane. In *International Symposium on Graph Drawing*.

Singer, W. (1999). Neuronal synchrony: A versatile code for the definition of relations? *Neuron*, 24(1):49–65.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (NeurIPS).

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.

Yang, M., Bian, Y., and Huang, M. (2024). Hypformer: Exploring efficient transformer fully in hyperbolic space. In *KDD*.

A Mathematical Derivations

A.1 Einstein Midpoint Preserves Manifold Constraint

Theorem 2. *Let* $\{\mathbf{v}_j\}_{j=1}^N \subset \mathbb{L}^n$ *and* $\{\alpha_j\}$ *be weights with* $\sum_j \alpha_j = 1$. *The Einstein midpoint*

$$\mathbf{c} = \frac{\sum_{j=1}^N \alpha_j \gamma_j \mathbf{v}_j}{\left\| \sum_{j=1}^N \alpha_j \gamma_j \mathbf{v}_j \right\|_{\mathcal{L}}} \tag{19}$$

where $\gamma_j = 1/\sqrt{-\langle \mathbf{v}_j, \mathbf{v}_j \rangle_{\mathcal{L}}} = 1$, *satisfies* $\mathbf{c} \in \mathbb{L}^n$.

Proof. Let $\mathbf{s} = \sum_{j=1}^N \alpha_j \mathbf{v}_j$. Since each $v_{j,0} > 0$ and $\alpha_j > 0$, we have $s_0 > 0$. The Lorentz norm is

$$\|\mathbf{s}\|_{\mathcal{L}}^2 = \langle \mathbf{s}, \mathbf{s} \rangle_{\mathcal{L}} = -s_0^2 + \|\mathbf{s}_{1:n}\|^2. \tag{20}$$

The normalized vector $\mathbf{c} = \mathbf{s}/\sqrt{-\langle \mathbf{s}, \mathbf{s} \rangle_{\mathcal{L}}}$ satisfies

$$\langle \mathbf{c}, \mathbf{c} \rangle_{\mathcal{L}} = \frac{\langle \mathbf{s}, \mathbf{s} \rangle_{\mathcal{L}}}{-\langle \mathbf{s}, \mathbf{s} \rangle_{\mathcal{L}}} = -1. \tag{21}$$

Thus $\mathbf{c} \in \mathbb{L}^n$. □

A.2 Capacity Bound Derivation (REVISED - Option B)

Proof of Proposition 3. We rely on the scaling laws established by Manifold Capacity Theory (Chung et al., 2018). Consider embedding P object manifolds, each of radius R_M and intrinsic dimension D_M, into an ambient space of dimension N. The critical capacity $\alpha_c = P/N$ scales as $\alpha_c \sim (R_M \sqrt{D_M})^{-1}$.

In the hierarchical setting (depth r), the number of semantic concepts grows as $P \sim e^r$.

Euclidean Case. The total volume available in a ball of radius L in $\mathbb{R}^N$ is $V_{\text{total}} \propto L^N$. To fit P non-overlapping manifolds of radius R_M, we require $P \cdot (R_M)^N \lesssim L^N$. Assuming a bounded ambient norm ($L \sim \text{const}$), this implies:

$$R_M \lesssim \left(\frac{L^N}{P} \right)^{1/N} = L \cdot P^{-1/N}. \tag{22}$$

Substituting $P \sim e^r$:

$$R_M \sim L \cdot e^{-r/N} = O(e^{-r/N}). \tag{23}$$

The capacity thus scales as:

$$\alpha_c^{\mathbb{R}} \propto \frac{1}{R_M\sqrt{D_M}} \sim \frac{e^{r/N}}{L\sqrt{D_M}}. \tag{24}$$

For fixed L and D_M, the separability margin degrades exponentially with depth r.

Hyperbolic Case. The volume in $\mathbb{H}^N$ grows as $V_{\text{total}} \propto e^{(N-1)L}$. To fit $P \sim e^r$ manifolds, we require:

$$e^r \cdot (R_M)^N \lesssim e^{(N-1)L}. \tag{25}$$

Crucially, because the ambient space expands exponentially, we can accommodate exponentially many manifolds while maintaining a constant radius $R_M = \Theta(1)$ relative to the scale of the space:

$$R_M \sim \left(\frac{e^{(N-1)L}}{e^r}\right)^{1/N} = \Theta(1) \quad \text{when } r \lesssim (N-1)L. \tag{26}$$

Consequently, the capacity remains:

$$\alpha_c^{\mathbb{H}} \propto \frac{1}{R_M\sqrt{D_M}} \sim \Theta(1). \tag{27}$$

The ratio $\alpha_c^{\mathbb{H}}/\alpha_c^{\mathbb{R}}$ therefore exhibits an exponential advantage:

$$\frac{\alpha_c^{\mathbb{H}}}{\alpha_c^{\mathbb{R}}} \sim \frac{L\sqrt{D_M}}{e^{r/N}\cdot\Theta(1)} = \Omega(e^{r/N}) \tag{28}$$

in favor of the hyperbolic substrate. □

A.3 H-AKOrN Convergence Analysis

Proof Sketch of Phase Synchronization Theorem. Define Lyapunov function

$$V = \frac{1}{2}\sum_{i,j} A_{ij} d_{\mathbb{L}}^2(\mathbf{h}_i, \mathbf{h}_j). \tag{29}$$

The time derivative satisfies

$$\frac{dV}{dt} = -\eta K \sum_{i,j} A_{ij} \left\|\log_{\mathbf{h}_i}(\mathbf{h}_j)\right\|^2 \leq 0. \tag{30}$$

By LaSalle's invariance principle, the system converges to the largest invariant set where $\dot{V} = 0$, which occurs when $\log_{\mathbf{h}_i}(\mathbf{h}_j) = 0$ for all connected pairs, i.e., phase synchronization. □