Attention as Heat Kernel Diffusion: A Geometric Framework for Interpreting Language Models

Saar Hamam

Independent Researcher

Abstract

I propose a novel interpretation of attention mechanisms in large language models (LLMs) through the lens of differential geometry. Specifically, I generalize attention as a heat kernel defined over a learned Riemannian manifold induced by the model's internal representations, referred to here as the *learned space*. This formulation yields a coordinate-invariant, mathematically principled framework for interpreting semantic flow, information propagation, and inductive biases in language models. We can extract local geometric invariants—including the metric tensor, Christoffel symbols, and the Laplace–Beltrami operator—directly from the learned space using autodifferentiation, without access to model weights. This framework enables a geometric interpretation of attention as heat diffusion, and offers new foundations for interpretability, robustness, and curvature-aware modeling. I further hypothesize that the curvature of the learned space corresponds to natural language structures in a testable way.

1 Introduction and Hypothesis

Language models achieve remarkable performance, yet their inner workings remain opaque. While attention is typically framed as an algebraic similarity operation, I argue that it is better understood as a geometric diffusion process over a manifold induced by the model's learned space.

Hypothesis. *Let $\mathcal{M}$ be the learned space induced by a large language model (LLM), with a differentiable structure inherited from its representation function $f : \mathcal{X} \to \mathcal{M} \subset \mathbb{R}^d$, where $\mathcal{X}$ is the space of discrete or continuous token inputs. Then:*

1. *$\mathcal{M}$ admits a Riemannian manifold structure with a pullback metric $g = Df^\top Df$, from which geometric quantities such as curvature and connection can be extracted.*

2. *The attention mechanism can be interpreted as a heat kernel $K_t(x, y)$ over $(\mathcal{M}, g)$, approximating semantic diffusion under the Laplace–Beltrami operator Δ_g.*

3. *The curvature of $(\mathcal{M}, g)$, extracted from the learned space via the connection coefficients Γ^k_{ij}, correlates with structural properties of natural language such as semantic hierarchies, syntactic trees, and contextual dependencies.*

This hypothesis is empirically testable by extracting local geometric invariants from pretrained language models and comparing them against known linguistic structures.

2 Attention as Heat Kernel Diffusion

2.1 Classical Attention

In standard Transformers, attention uses the scaled dot-product:

$$\mathrm{Attn}(x_i, x_j) = \mathrm{softmax}\left(\frac{Q_i \cdot K_j}{\sqrt{d}}\right),$$

which is equivalent to applying a Gaussian kernel in a Euclidean space.

2.2 Heat Kernel Generalization

I propose replacing this similarity kernel with the heat kernel $K_t(x, y)$ on a Riemannian manifold $\mathcal{M}$, defined via:

$$\left(\frac{\partial}{\partial t} - \Delta_x\right) K_t(x, y) = 0, \quad K_0(x, y) = \delta_y(x),$$

where Δ is the Laplace–Beltrami operator and $t > 0$ is diffusion time. Subsequently, attention weights are defined as:

$$\alpha_{ij} \propto K_t(x_i, x_j).$$

3 Extracting the Geometry of the Learnt Space

Let $f : \mathbb{R}^n \to \mathbb{R}^d$ be the differentiable map implemented by the model from token inputs to representation space.

3.1 Metric Tensor

The metric at x is derived from the Jacobian $J = Df(x)$:

$$g_{ij}(x) = \langle \partial_i f, \partial_j f \rangle = (J^\top J)_{ij}.$$

3.2 Christoffel Symbols

Compute the Levi–Civita connection coefficients:

$$\Gamma^k_{ij} = \tfrac{1}{2} g^{kl} \big(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}\big),$$

using autodifferentiation or finite differences over neighboring tokens.

3.3 Laplace–Beltrami Operator

For scalar functions h, the Laplace–Beltrami operator is given by:

$$\Delta h = \frac{1}{\sqrt{\det g}} \partial_i \Big(\sqrt{\det g}\, g^{ij} \partial_j h\Big).$$

Using this operator, it is possible to compute or approximate the heat kernel K_t necessary for defining geometric attention weights.

4 Interpretability and Testability

- Extracted curvature should correlate with linguistic features (e.g., hierarchical structures, semantic distances). - The hypothesis is empirically testable by comparing curvature distribution in pretrained LLMs with known language hierarchies.

5 Related and Emerging Research

Several works explore geometry in LLMs, but none fully implement this dynamic, learned manifold framework:

- Transformer Dissection (Wang et al., 2019): interprets attention as kernel smoothing—`https://arxiv.org/abs/1908.11775`.
- Implicit Kernel Attention (Xu et al., 2020): proposes learnable attention kernels—`https://arxiv.org/abs/2006.06147`.
- Heat Diffusion Transformers for Mesh (Wong et al., CVPR 2023): applies heat kernels to mesh geometry—`https://openaccess.thecvf.com/content/CVPR2023/papers/Wong_Heat_Diffusion_Based_Multi-Scale_and_Geometric_Structure-Aware_Transformer_for_Mesh_CVPR_2023_paper.pdf`.
- HELM: Hyperbolic LLMs via Mixture-of-Curvature Experts (May 2025): trains fully hyperbolic LLMs at billion-parameter scale—`https://arxiv.org/abs/2505.24722` :contentReference[oaicite:0]index=0.
- Rethinking LLM Training through Information Geometry (Di Sipio, July 2025): explores curvature in optimization landscapes of LLMs via Fisher and quantum metrics—`https://arxiv.org/abs/2506.15830` :contentReference[oaicite:1]index=1.

Unlike prior work, my approach dynamically learns geometry from the model itself and extracts curvature and geometric invariants without predefining the curvature type.

6 Conclusion and Future Directions

I attempt to present a geometric interpretation of attention as heat diffusion over a learned manifold. The framework could perhaps provide:

- A coordinate-invariant approach to semantic interpretability.
- Empirical testability by comparing curvature distributions with linguistic hierarchies.
- A theoretical foundation for geometry-aware model design and regularization.

Future work will involve:

1. Model the entire system as a dynamic system.
2. Empirical extraction of curvature and geometric invariants from pretrained LLMs.
3. Comparative studies of curvature alignment with linguistic structures.

4. Architectural extensions that incorporate curvature-awareness during training.
5. Exploration of spectral properties of the Laplacian and their relation to model behavior.

This line of work opens up pathways toward interpretable and robust AI grounded in intrinsic geometry.

7 Future Work: Attention as a Geometric Dynamical System

A compelling future direction is to extend this framework by interpreting the transformer not merely as a kernel smoother or heat diffusion operator, but as a *discrete geometric dynamical system* evolving over the learned manifold.

In this view, each transformer layer defines a map $\Phi_t : \mathcal{M} \to \mathcal{M}$, and the sequence of hidden states $(x_0, x_1, \ldots, x_L)$ represents a trajectory through the manifold $\mathcal{M}$. This allows for reinterpretation of attention and token updates as discrete steps of a flow induced by a vector field.

Two canonical geometric frameworks can govern such dynamics:

- **Riemannian gradient flow:** If attention arises from minimizing a local energy function $E(x)$, the evolution corresponds to gradient descent in a curved space: $\frac{dx}{dt} = -\nabla_g E(x)$, where ∇_g is the Riemannian gradient.
- **Symplectic (Hamiltonian) flow:** If attention preserves some invariant structure or semantic "volume", the evolution may follow Hamiltonian mechanics. With a symplectic form ω, the flow is given by $\iota_{X_H}\omega = dH$, where $H(x)$ is a scalar Hamiltonian function and X_H is the corresponding vector field.

Modeling attention as a geometric dynamical system would allow analysis of long-term semantic trajectories, attractor structures, conserved invariants, and curvature-driven transitions. It also opens the door to using tools from symplectic integration, geometric ODE solvers, and topological phase analysis in the study of LLMs. This extension offers a mathematically principled foundation for deeper insights into reasoning, compositionality, and model behavior over time.

Disclaimer: This research was AI assisted.