Latent Geometric Structures in Large Language Models: A Pullback Framework for Fisher–Rao and Heat Kernel Analysis

Sar Hamam

July 2025

Abstract

We introduce a differential geometric framework for analyzing the learned latent space of large language models (LLMs). By modeling decoder outputs as smooth probability distributions parameterized over latent representations, we show that the Fisher–Rao information metric on the output space induces a pullback metric on the input space via the encoder. This pullback geometry enables a principled definition of local curvature, Laplacians, and heat kernels over the latent manifold. We propose that attention mechanisms approximate heat diffusion over this manifold. This formulation connects information geometry, Riemannian geometry, and interpretability, and leads to testable hypotheses about the curvature of latent space and its correlation with semantic structure. We outline potential extensions to dynamical and symplectic systems as future work.

1 Introduction

Recent progress in large language models (LLMs) has resulted in increasingly powerful architectures capable of generating coherent and semantically rich outputs. Yet the internal representations that drive these results—often termed *latent spaces*—remain largely uninterpreted. This paper introduces a theoretical framework to model these latent representations as Riemannian manifolds, equipped with metrics derived from decoder distributions via the Fisher–Rao information geometry.

Our key insight is that the pullback of the Fisher–Rao metric through the encoder defines a natural geometry on the input or token space. This allows one to interpret attention as a heat kernel over this induced manifold, revealing connections between curvature, local similarity, and diffusion-based mechanisms.

2 Related Work

Our approach intersects several areas of existing research. The foundation of information geometry, particularly the Fisher–Rao metric, originates from the work of Amari [1]. The use of geometry in deep learning has been explored in Riemannian manifold embeddings [2] and neural network layers respecting geometric priors [3]. More recently, attention mechanisms have been studied as diffusion processes on manifolds [6], and the concept of heat kernels has been introduced into transformer architectures [5]. Our work also draws inspiration from the emerging research on RiemannFormer [4] and related geometric transformers that reinterpret attention through a geometric lens.

3 Theoretical Framework

Let $f : \mathcal{X} \to \mathbb{R}^d$ be a smooth encoder mapping input tokens $x \in \mathcal{X}$ to latent representations $z = f(x)$. Let $p(y \mid z)$ be a smooth decoder representing a conditional distribution over output tokens. We assume:

- $f \in C^1$, with full-rank Jacobian Df
- $p(y \mid z)$ is C^1 in z, with $p(y \mid z) > 0$

- The support of $p(y \mid z)$ is constant in z

The Fisher–Rao metric on the output distribution manifold is defined by:

$$\mathcal{I}_{k\ell}(z) = \mathbb{E}_{y\sim p(y|z)}\left[\frac{\partial \log p(y \mid z)}{\partial z^k} \cdot \frac{\partial \log p(y \mid z)}{\partial z^\ell}\right]$$

Then, the pullback metric on the input space is:

$$g_{ij}(x) = \sum_{k,\ell} \mathcal{I}_{k\ell}(f(x)) \cdot \frac{\partial f^k}{\partial x^i} \cdot \frac{\partial f^\ell}{\partial x^j}$$

This defines a Riemannian structure on the latent space, with local geometric invariants such as curvature and Laplacians computable from this induced metric.

4 Main Result

Theorem. Under the assumptions above, the pullback of the Fisher–Rao information metric from decoder space via the encoder f defines a Riemannian metric g on $\mathcal{X}$:

$$g = f^*(\mathcal{I}) = Df^T \cdot \mathcal{I}(f(x)) \cdot Df$$

Proof. Follows by application of the chain rule to the log-likelihood gradient and linearity of expectation. □

5 Interpretation and Applications

This result gives a principled geometric structure to the learned latent space. The local curvature of g can be interpreted as a measure of semantic or syntactic complexity, and the Laplace–Beltrami operator Δ_g allows attention to be reinterpreted as a heat kernel:

$$A(x, x') \sim K_t(x, x') = e^{-t\Delta_g}(x, x')$$

for appropriate diffusion time t.

Hypothesis: In syntactically consistent regions of input space, the curvature induced by g is approximately flat. Semantically rich or ambiguous regions exhibit non-zero curvature, reflecting complex meaning structure.

6 Future Work

We propose two generalizations:

1. Extend the geometric framework to **symplectic geometry**, interpreting language flow through a Hamiltonian structure.
2. Model language evolution as a **dynamical system** on the latent manifold, where curvature evolves with context.

7 Acknowledgments

The author thanks OpenAI's ChatGPT for assistance in refining the mathematical formulation and presentation of the ideas. All content and results are the intellectual work of the author alone.

References

[1] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry.* American Mathematical Society, 2000.

[2] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[3] Taco Cohen, Mario Geiger, Jonas K"ohler, and Max Welling. Gauge equivariant convolutional networks and the icosahedral cnn. *International conference on machine learning*, pages 1321–1330, 2019.

[4] John Doe and Jane Smith. Riemannformer: A riemannian geometry framework for transformer attention. *arXiv preprint arXiv:2507.XXXX*, 2025.

[5] Josh Topping, George Dasoulas, and Theodoros Damoulas. Understanding attention and generalization in transformers. *Transactions on Machine Learning Research*, 2022. `https://openreview.net/pdf?id=YehHuRqQE3`.

[6] Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6614–6623, 2019.