The Attralucian Essays:

Exploring the Finite

First Edition

Copyright © 2025 by Kevin R. Haylett. All rights reserved.

This work is shared under the Creative Commons Licence.

Creative Commons CC BY-ND 4.0 License.
https://creativecommons.org/licenses/by-nd/4.0/

This work is intended for academic and research use. Any unauthorized distribution, modification, or commercial use beyond the creative use license is strictly prohibited. Typeset in LaTeX

The Attralucian Essays

The Geodesic Fractal Model of LLMs

Kevin R. Haylett

The Geodesic Fractal Model of LLMs

LLMs as Nonlinear Dynamical Systems on a Learned Manifold

State, Update, and Closed-Loop Generation

Let a tokenized history $s_{1:t} = (w_1, \ldots, w_t)$ map to embeddings $e_{1:t}$. Define the internal state $x_t \in \mathbb{R}^d$ (concatenated hidden activations for all positions, or just the last position for simplicity). Decoding forms a *closed-loop* nonlinear system:

$$x_{t+1} = \Phi_\theta\big(x_t,\, e_{t+1}\big), \quad w_{t+1} \sim p_\theta(\cdot \mid x_t), \quad e_{t+1} = E(w_{t+1}),$$

where Φ_θ is the transformer stack (multi-head attention + MLP + residual/normalization). With residual blocks $x^{\ell+1} = x^\ell + f_\ell(x^\ell)$, the continuous-depth limit is a Neural ODE:

$$\frac{dx}{d\ell} = f(x, \ell; \theta).$$

Attention as Pairwise Delay-Embedding

Self-attention per layer/head is defined as:

$$\mathrm{Attn}(X) = \mathrm{softmax}\Big(\frac{QK^\top}{\sqrt{d_k}}\Big)V, \quad Q = XW_Q,\ K = XW_K,\ V = XW_V.$$

Viewed dynamically, attention builds a *delay-coordinate map*: each position's state is reconstructed from pairwise similarities to other positions (a Takens-style embedding of the sequence into a higher-dimensional phase space). This yields a context-dependent coordinate chart $x_t = \Psi_\theta(e_{1:t})$.

The Hyper-Dimensional Manifold and its Metric

Training shapes a representation manifold $\mathcal{M} \subset \mathbb{R}^d$ where nearby points encode semantically coherent continuations. A natural Riemannian metric at state x comes from the output distribution $p_\theta(y \mid x)$ via the Fisher information:

$$G(x) \;=\; \mathbb{E}_{y\sim p_\theta(\cdot|x)}\big[\nabla_x \log p_\theta(y \mid x)\, \nabla_x \log p_\theta(y \mid x)^\top\big].$$

This metric measures *semantic sensitivity* of predictions to movement in state space.

Geodesics and Token Selection

Given a local loss

$$\mathcal{L}(x) = \mathrm{CE}(p^*(\cdot \mid x), p_\theta(\cdot \mid x)) = -\log p_\theta(w^*_{t+1} \mid x),$$

the *natural gradient* step

$$\Delta x \;\propto\; -G(x)^{-1}\nabla_x \mathcal{L}(x)$$

is the steepest descent direction under the Fisher metric. Integral curves approximate *geodesics* of $(\mathcal{M}, G)$ toward regions that increase next-token likelihood. In practice, choosing w_{t+1} (argmax/sampling) and re-embedding it implements a piecewise geodesic walk:

$$x_t \xrightarrow{\text{token choice}} x_{t+1} = \Phi_\theta\big(x_t, E(w_{t+1})\big).$$

Why the Landscape Appears Fractal

With gated, piecewise-linear components (ReLU/GeLU + softmax), deep transformers partition $\mathbb{R}^d$ into exponentially many linear regimes. Context-dependent attention gates induce *self-similar, multi-scale tilings*. Across layers/heads, these tilings compose, yielding a *fractal-like* semantic energy landscape: thin filaments (high-probability "ridges"), basins (attractors such as loops and clichés), and branching cascades (topic shifts/bifurcations).

Dynamical Systems View of Decoding

Closed-loop decoding is a stochastic nonlinear map:

$$x_{t+1} = \Phi_\theta\big(x_t, E(\xi_t)\big), \quad \xi_t \sim p_\theta(\cdot \mid x_t;\, T,\, p,\, k),$$

parametrized by temperature T, top-p, or top-k. Observed phenomena include:

- **Fixed points / limit cycles:** repetitions, rhymes, catchphrases.
- **Bifurcations:** qualitative changes as sampling parameters vary.
- **Chaotic sensitivity:** small prompt or seed changes lead to divergent trajectories.

Practical Diagnostics (Measurables)

Given the layer-L states x_t:

$$\text{Fisher–Rao length:} \quad \mathcal{L}_{FR} = \sum_t \sqrt{\Delta x_t^\top G(x_t) \Delta x_t},$$

$$\text{Curvature:} \quad \kappa_t \approx \frac{\|x_{t+1} - 2x_t + x_{t-1}\|}{\|x_t - x_{t-1}\|^2},$$

Attractor probing: vary temperature T and measure return tim

Minimal Takeaway

- The transformer induces a *vector field* on a learned semantic manifold.

- Attention computes *pairwise delay-embeddings*, giving coordinates.
- Decoding follows *piecewise geodesics* (under a Fisher-type metric), navigating a *fractally tiled* landscape shaped by training.