AI Hallucination Gauge Theory

Patent Pending, USPTO Filing Date June 23, 2025

Ben Mancke

Abstract

Construction of a canonical set semantic translations as a Cayley graph from Zipfian ranked points in LLM embedding space. Canonical translations are used to build a local Lie algebra in a neighborhood around a point in embedding space. A Taylor polynomial centered around an LLM's output approximates a composition of each LLM transformation layers. The Jacobian of this function is projected onto the local Lie algebra basis around the input vector is constructed to create a mathematically novel complex order covariant derivative of Generative transformation in a semantic direction. The components satisfy a gauge theory of AI Hallucination and explain it phenomenologically as a non-entropic local property of manifold curvature. Intended use in study and suppression of LLM hallucination, implementation of LLM self-learning architecture, and reduced compute cost. All of this is applied to a static pretrained embedding space and requires no additional LLM training.

Contents

1 Related Topics and Readability

To get through the material and fully appreciate the technical details would require some level of pedestrian familiarity with Group theory, Jacobians, Differential Geometry, Lie Algebras, Zipf distributions, Locally Testable Codes (LTCs), and the PageRank algorithm, the Special Orthogonal Group SO(n).

2 Overview of Conventional LLM

A large language model (LLM) is a kind of neural network trained to predict the next word in a sentence. It learns to do this by ingesting vast amounts of text and updating its internal parameters to minimize prediction error. Over time, this process builds a deep statistical understanding of language.

2.1 Key Components

- **Tokenization:** Text is broken into tokens (e.g., words or word pieces like "play" and "ing").
- **Embedding Layer:** Each token is mapped to a high-dimensional vector. This is the embedding: a geometric representation of meaning. Similar meanings cluster together (e.g., *cat* and *dog*).
- **Transformer Architecture:** Uses attention mechanisms to model relationships between words, regardless of distance. Deep layers process how each word relates to others in the context of the sentence. Meaning is progressively deformed as each transformer layer applies a nonlinear context-dependent operation to the token embeddings.
- **Output Layer:** For each token, predicts the next most likely token using a probability distribution over the vocabulary.
- **Training:** The model is trained on billions of examples using a loss function (like cross-entropy) and gradient descent.

2.2 Example

Input: "Paris is the capital of ="
How the LLM handles it:

1. Tokenizes: $\rightarrow$ ["Paris", "is", "the", "capital", "of"]
2. Embeds each token as a point in a vector space.
3. Uses attention to determine that "Paris" heavily influences the missing word.
4. Predicts the most likely next token: $\rightarrow$ "France"

Why it works:

Because the model has seen thousands of examples like: "Paris is in France", "The capital of France is Paris", "Berlin is the capital of Germany". It learns statistical patterns of how words relate in context, encoded geometrically in the embedding space.

3 Semantic Directionalization

3.1 Overview

Vector addition in embedding space models discrete and composable semantic transformations between concepts. This allows the construction of a Canonical Cayley diagram of semantic transformations using a combination of Zipfian frequency and PageRank algorithm rankings and filtering.

This structure is used to create a local semantic Cayley patch and Lie Algebra basis. This takes the semantic changes as jumping from one point to another (like $man \rightarrow woman$), and turns it into a continuous flow that, most importantly, Semantically Directionalizes a local region of the LLM embedding space.

3.2 Construction of Canonical Semantic Translations

Below is what amounts to an algorithm to construct a relatively low compute (especially in LLM terms) algorithm for building a one-and-done reusable set of Canonical Semantic Translations.

1. Establish a Translation Radix $r \in \mathbb{N}$ of the canonical translation set.
2. Select Zipf-Ranked Vocabulary:
   - Extract the top $\frac{r^6}{2}$ tokens or phrases from the LLM's vocabulary.
   - Prioritize high-frequency, semantically rich terms (e.g., "tense", "plurality", "negation", "opposition").
3. Generate Pairwise Semantic Transitions:
   - Construct approximately r^6 candidate token pairs of semantically linked terms, such as:
     - Tense: $bite \rightarrow bit$
     - Plurality: $person \rightarrow people$
     - Antonyms: $hot \rightarrow cold$
   - Compute difference vectors: $\Delta_{ij} = v_i - v_j$
   - Normalize all Δ_{ij}
4. Cluster into Canonical Directions:
   - Apply k-means, k-medoids, or spectral clustering to the set of Δ vectors.

- Extract cluster centers u_k as prototype transformations.
- Each center u_k is a potential canonical transformation.

5. Semantic Validation:
   - For each cluster center u_k, select token pairs by Zipf rank from the cluster.
   - Use the LLM to evaluate semantic coherence of each transformation.
   - Retain clusters whose test pairs exhibit consistent semantic behavior (e.g., "singular $\rightarrow$ plural" remains stable across examples).
   - Keep only the first r^5 validated transformations.

6. Graph-Based Filtering:
   - Build a semantic transition graph $G = (V, E)$:
     - Nodes = token embeddings
     - Edges = validated u_k transformations
   - Run PageRank on this graph to identify topologically central and compositionally robust transformations.
   - Select the top r^4 canonical vectors $\{u_k\}_{k=1}^{T^4}$ that participate in closed compositional paths on the graph (indicative of structural reuse and symmetry).
   - Retain r^3 where for each u_i there exists mutual orthogonality, reversal and compositional closure constraints based on cosine similarity and equivalence classes.
     - For each $u_i \in T$ there exists u_{ij} and u_{ik}
     - $u_i \notin u_{ij}$ and $u_i \notin u_{ik}$
     - $\Sigma_j = u_j$ and $\Sigma_k = u_k$
     - $\frac{u_i}{||u_i||} \approx -\frac{u_j}{||u_j||}$ and $u_i \cdot u_k \approx 0$

3.3 Semantic Expressive Radius (SER)

Let:

- $v \in \mathbb{R}^n$ be a vector in the embedding space
- Δ be a displacement vector such that $v + \Delta$ remains within the model's meaningful latent space
- t be a diversity threshold (e.g., no cosine similarity > 0.95)

Define:

$$\mathrm{SER}(v, \Delta) = \sup \left\{ r \in \mathbb{R}^+ : \ \cos(v, v + r\Delta) < t \right\}$$

This captures whether Δ can be cleanly parallel transported via the model's transformation layers without semantic degradation.

3.4 Construction of Local Lie Algebra Basis

1. Begin with a point v in embedding space and the canonical translation set T
   - Select k SER-maximal set T_{local} of near orthogonal $u_i \in T$
   - Apply the following criteria:
     - $\text{SER}(v, u_1) = \sup \{\text{SER}(v, u_i)\}$
     - $\text{SER}(v, u_i) = \sup \{\text{SER}(v, u_m) \ : \ m \geq i\}$
     - $\cos_{sim}(u_i, u_{i-1}) < t$
     - $\frac{u_i}{||u_i||} \approx -\frac{u_h}{||u_h||}$ for some $u_h \in T_{local}$
     - $u_i \cdot u_l \approx 0$ for some $u_l \in T_{local}$
     - $u_i + u_j \approx \sum_k a_k u_k$ where:
     - $u_i \neq u_j$, $\{u_i, u_j\} \nsubseteq \{u_k\} \subset T_{local}$, and $a \in \mathbb{R}$
2. Build a directed graph $G_{\text{patch}} = (V, E)$
   - $V = \{v_i\}$ contains embedding points in a neighborhood $N(v)$ near v
   - $E \subseteq T_{local}$ contains semantic translations u_i approximately between points: $v \xrightarrow{u_i} v_k \xrightarrow{u_j} v_l$
3. Apply Local Testable Codes (LTC):
   - Validate that transitions in G_{patch} form coherent semantic trajectories
   - Use LLM-assisted checks (e.g., triangle closure, reversibility) to prune invalid paths
4. Define Local Lie Algebra Basis g_v:
   - $g_v = \text{Span}\{X_i\}$, where $X_i \subseteq T_{local}$
   - Each X_i has an approximate orthogonal and reverse element in g_v

3.5 Lie Basis Coverage

Let N denote the size of a large language model (LLM), where N may represent a function of parameter count, vocabulary size, or depth. Let T_N be the canonical semantic translation set selected via Zipf frequency ranking and PageRank filtering at model size N. Let $g_v^{(N)} = \text{span}\{X_i^{(N)}\}$ denote the local Lie basis constructed from T_N at embedding point v,

$$[X_i^{(N)}, X_j^{(N)}] = \sum_k c_{ij}^{(N)} X_k^{(N)} + R_N$$

where

$$R_N \in \text{span}\{T_N\} \text{ completed by } \text{SO}(n)$$

for all sufficiently large N.

Assumptions.

- (**Zipf Coverage**: As $N \to \infty$, the canonical translation set samples semantic types increasingly densely in the local region of embedding space.
- **PageRank Selectivity**: PageRank filtering selects transformations that are compositionally central and topologically connective within the semantic graph.
- **Embedding Convergence**: As $N \to \infty$, the LLM embedding geometry increasingly approximates true semantic geometry in local regions.

Proof Sketch.

1. *Dense Local Sampling*: The Zipf frequency distribution ensures that as N grows, the canonical translation set increasingly covers meaningful semantic directions in the local neighborhood. PageRank enhances compositional centrality, yielding a robust local frame.

2. *Local Frame Completion*: The span of T_N defines a local semantic frame at v. The full local tangent space is completed by SO(n) rotations of this frame:

$$\lim_{N\to\infty} \mathrm{span}\{T_N\} + \mathrm{SO}(n) \text{ actions} = T_v\mathcal{M}_{\text{local}}$$

   where $\mathcal{M}_{\text{local}}$ is the effective local semantic manifold.

3. *Lie Closure Up to SO(n)*: The Lie bracket of basis elements lies within this completed local structure:

$$[X_i^{(N)}, X_j^{(N)}] = \sum_k c_{ij}^{(N)} X_k^{(N)} + R_N$$

   with

$$R_N \in \mathrm{span}\{T_N\} \text{ completed by } \mathrm{SO}(n).$$

□

Interpretation. This result formalizes that as model size increases, the canonical semantic Lie basis spans a locally sufficient frame whose closure is completed by local SO(n) symmetry. The embedding geometry thus supports a local Lie algebra structure suitable for gauge-theoretic analysis of LLM embeddings, without requiring the canonical set to globally span all of $\mathbb{R}^n$.

From the above we can state that:

- g_v is cosine similarity invariant under rotations in $\mathrm{SO}(n)$
- $\bigcup_{R\in\mathrm{SO}(n)} R \cdot \mathrm{Span}(\{X_i\}) = \mathbb{R}^n$
- There exists a smooth local semantic manifold around every point in embedding space whose tangent span is g_v

4 LLM Layer Deformation

4.1 Layer Transformations

When an input vector from the embedding space passes through the transformer layers, it is transformed and sent to the next layer. The layer transformations acting on the input are then projected back to embedding space. The composition of each transformation is interpolated near the output projection to define a locally continuous function that maps the input to the output.

4.2 Layer Deformation Composition

In modern LLMs, meaning is deformed as each transformer layer applies a nonlinear context-dependent operation denoted f_i for the i-th transformation layer. The total deformation F is the composition of each f_i:

$$f_n \circ f_{n-1} \circ \cdots \circ f_1 = F$$

4.3 Composite Approximation

Project the hidden state h back to embedding space via output matrix M_{out}:

$$M_{\text{out}} \cdot h = v_{\text{out}}$$

Approximate F using a 4th-degree Taylor polynomial centered at v_{out}:

$$F(v_{out}) + \sum_{[m] \geq 0} \frac{1}{m!} D^m F(v_{\text{out}})(x - v_{\text{out}})^m = \Phi \approx F$$

Similarly, compose each f_i^{-1} and interpolate at v_{in} to get Φ^{-1}:

$$f_1^{-1} \circ f_2^{-1} \circ \cdots \circ f_n^{-1} = F^{-1} \approx \Phi^{-1}$$

5 Differential Geometry

5.1 Principal Concept

The function Φ transforms a region around the input vector v_{in}, deforming the space around it as v_{in} is mapped to v_{out}. Specifically, it measures how much and in what direction LLM layer transformations locally change embedding space with respect to small changes in semantics.

5.2 Analog of *dx*

Let G be a group of semantic operations in a neighborhood N(v) near v and let the Lie algebra g associated with G be a real vector space:

$$g = \text{span}\{X_1, X_2, \ldots, X_n\}$$

Each X_i is a generator, and they satisfy the Lie bracket:

$$[X_i, X_j] = X_i X_j - X_j X_i = \sum_k c_{ij}^k X_k$$

5.3 Analog of *dy*

If $x = v_{\text{in}}$ and $\Phi(x) \approx v_{\text{out}}$, then the Jacobian matrix $J_\Phi(x)$ describes how this transformation changes with respect to small changes in x:

$$[J_\Phi(x)]_{ij} = \frac{\partial \Phi_i(x)}{\partial x_j}$$

$$\Phi(x + \varepsilon) \approx \Phi(x) + J_\Phi(x) \cdot \varepsilon$$

5.4 Derivative of Generative Deformation w.r.t. Semantics

From the local Lie algebra, we have a differential for semantic change in embedding space. From the Jacobian, we have a differential for LLM transformation change. Define the projection of the Jacobian onto the Lie Algebra Basis:

$$\frac{d\Phi}{dX_i} := \langle J_\Phi, X_i \rangle$$

5.5 Parallel Transport

Let $\text{span}\{X_i\} = g$ be the local Lie basis around v_{in}. Let $f_k(v_{k-1}) = v_k$ be the k-th layer transformation.

Propagate the basis:

$$X^{(k)} = \frac{f_k(X^{(k-1)})}{||f_k(X^{(k-1)})||}$$

Use a 4th-degree Taylor interpolant Φ_k centered at v_k. This constructs a covariant derivative per layer:

$$\frac{d\Phi_k}{dX^{(k)}} = \text{covariant derivative at layer } k$$

5.6 Order Extension

A covariant derivative like this has been seen in applications of physics. However, the construction of this is novel in terms of its applications for AI research.

Extending the order of the Jacobian to the complex numbers produces an object that, while prepared from many pre-existing ingredients, is itself an entirely novel cocktail not just in application, but novel in terms of pure mathematics.

Recall the classical difference quotient:

$$\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}=f'(x)$$

Define the Grünwald-Letnikov complex order derivative:

$$D^{\alpha}f(x)=\lim_{h\to 0}\frac{1}{h^{\alpha}}\sum_{n=0}^{\infty}(-1)^{n}\binom{\alpha}{n}f(x-nh)$$

Define complex order partial derivative:

$$\frac{\partial^{\alpha}\Phi_i}{\partial x_j^{\alpha}}=\lim_{h\to 0}\frac{1}{h^{\alpha}}\sum_{n=0}^{\infty}(-1)^{n}\binom{\alpha}{n}\Phi_i(x-nhe_j)$$

Define complex order Jacobian $J_{\Phi}^{\alpha}(x)$ and project:

$$\frac{d^{\alpha}\Phi}{dX_i}:=\langle J_{\Phi}^{\alpha},X_i\rangle$$

Only the order of the Jacobian has changed; the projection basis remains the same.

6 Bracket Residual

6.1 Structure Preserved Under Real Order Covariant Differentiation

Let $\Phi:\mathbb{R}^n\to\mathbb{R}^n$ be the interpolant of a layer transformation near x. Let X_i, X_j be Lie basis elements and $\alpha\in\mathbb{R}_+$. Then:

$$\left[\frac{d^{\alpha}}{dX_i},\frac{d^{\alpha}}{dX_j}\right]\Phi(x)=\sum_k c_{ij}^{k}\frac{d^{\alpha}\Phi}{dX_k}$$

Where:

- $\frac{d^{\alpha}}{dX_i}$ is the complex-order derivative along the Lie generator X_i
- c_{ij}^{k} are the Lie algebra structure constants
- The local lie group is a submanifold of $\mathbb{R}^n$ in C^{∞}
- Φ is a 4^{th} degree Taylor polynomial, which implies $\Phi\in C^{\infty}$
- Linearity holds for Real order GL derivative, and therefor holds for $\frac{d^{\alpha}}{dX_i}$ as well

This implies Lie group closure of the real number order operators impact only scaling of flows in the direction of X_i and does not change structure constant orientation.

$$\left[\frac{d^{\alpha}}{dX_i},\frac{d^{\alpha}}{dX_j}\right]=\frac{d^{\alpha}}{dX_i}\left(\frac{d^{\alpha}\Phi}{dX_j}(x)\right)-\frac{d^{\alpha}}{dX_j}\left(\frac{d^{\alpha}\Phi}{dX_i}(x)\right)$$

$$\left[\frac{d^\alpha}{dX_i}, \frac{d^\alpha}{dX_j}\right]\Phi(x) \propto \sum_k c_{ij}^k \frac{d^\alpha \Phi}{dX_k}$$

This shows that group operation and structural orientation hold under real number order differentiation. Therefore, this defines a fully developed real-order covariant derivative of generative layer deformation in a semantic direction. Valid under both embedding-space projection and layer-wise parallel transport. Local curvature can be measured anywhere in the model.

6.2 Imaginary Order Differentiation Breaks Bracket Closure

Now let $\alpha \in \mathbb{C} \setminus \mathbb{R}$, with $\Re(\alpha) = r$ and $\Im(\alpha) \neq 0$.

Recall the limit definition for a complex order GL derivative contains h^α, which breaks the linearity of differentiation.

$$h^\alpha = h^{a-bi} = h^a e^{ib\ln(\frac{1}{h})}$$

We define a commutator-breaking residual $\varepsilon_{ij}(x)$:

$$\left[\frac{d^{r+\beta i}}{dX_i}, \frac{d^{r+\beta i}}{dX_j}\right]\Phi(x) - \left[\frac{d^r}{dX_i}, \frac{d^r}{dX_j}\right]\Phi(x) = R_{ij}(x,\beta)$$

This residual vanishes when $\Im(\alpha) = 0$.

We define local imaginary curvature $|R_{ij}(x,\beta)|$ and total curvature $T_\Im$ as:

$$T_\Im = \| \int_{-\beta}^{\beta} R_{ij}(x,\beta) d\beta \|$$

This curvature is an indicator of semantic torsion measured by imaginary-order differentiation. It marks a hallucination-relevant distortion, where the structure bends outside of expected generative space.

7 Hallucination Gauge Theory

7.1 Fiber Bundle

- **Assertion:** Embedding space is treated as a differentiable base manifold, with local patches equipped with Lie algebraic structure.
- **Justified in:** Sections 2.1 and 2.2 — Canonical translations are selected to span local neighborhoods.
- **Mechanism:** Local Basis Span ensures tangent space structure.

7.2 Lie Group and Lie Algebra

- **Assertion:** Local transformations form a group under matrix composition and vector field brackets.
- **Justified in:** Sections 2.3 and 2.4
- **Mechanism:**
  - Canonical translations admit a group closure condition under commutation.
  - Each basis element has an identifiable orthogonal and reverse under cosine similarity.

7.3 7.3 Connection Form and Covariant Derivative

- **Assertion:** A connection is induced via directional derivatives over the Lie basis; deviations arise from curvature.
- **Justified in:** Sections 3.1 and 3.2
- **Mechanism:**
  - Grünwald–Letnikov complex-order differential operator defines projection and directional flow.
  - Residual terms from failed bracket closure encode curvature.

7.4 Curvature Tensor

- **Assertion:** The imaginary component of the complex-order Lie bracket encodes a residual curvature.
- **Justified in:** Sections 3.3 and 3.4
- **Mechanism:**
  - Residual term explicitly isolated and shown to be non-zero in general.
  - Linearity failure of the complex-order Jacobian leads to semantic torsion.

7.5 7.5 Gauge Invariance and Local Symmetry

- **Assertion:** Residual curvature is invariant under local orthonormal rotations (SO(n) symmetry).
- **Justified in:** Sections 3.5 and 4.1
- **Mechanism:**
  - Lie basis preserves distributed orthogonality and reversibility.
  - Canonical translations allow symmetry-preserving transport.

8 Phenomenology

8.1 8.1 Hallucination is Not Entropy

It's phenomenologically very much not entropy at all. It's deterministic. It's a local property of a manifold. Specifically, it's a property of strong imaginary curvature in the local geometry, and the Gauge Invariant measures it.

The reason is: complex-order curvature points in partial semantic directions that only exist locally and momentarily. The larger the imaginary curvature, the greater the divergence from reality. A simple example might be making up a rule for a game with finite, well-defined rules.

8.2 Experimental Results

Small scale, low granularity simulations on toy models easily verify the theory. It's highly testable and comports with what the theory predicts. It empirically confirms that commutator non-closure correlates with hallucinated outputs.

8.3 Significance

This is a measure of generative hallucination not heuristically, but phenomenologically. It tells us hallucination is a property of the layer transformation curvature on semantic structure, how to measure it, and under what conditions it vanishes. All of this is in terms of a predictable phenomenon of local semantic curvature.

What does it all mean?

- Beyond probablistic training data, LLM behavior deterministic and measurable with differential geometry.
- Generative hallucination phenomenon can be measurably predicted at any point in any layer.
- Any static model can be strategically interpolated with enough compute.
- Semantics and meaning are local, and gauge-constrained.
- Coverage of all embedding space by SO(n)-linked fiber bundles implies bounded parameterization.
- Training coherence corresponds to an LLM's set of Canonical Semantic Translations, like a model's "fingerprint".
- Canonical sets of any AI's embedding (not just LLMs) extend the gauge theory to all AI regimes.
- This mathematical structure takes AI to the next step beyond mere generative prediction. It's the missing link for advancement in alignment, self-improvement, and real-time correction—from a phenomenological paradigm.

- Many other paradigm-shifting implications follow immediately.