Beyond Scaling: A Stage 3 Geometric Framework for LLM Transparency through Language Manifold Dynamics

Lijia Zhang
Independent Researcher
lz.lijiazhang@gmail.com

April 2026

Abstract

This paper proposes a Stage 3 theoretical framework for understanding large language models (LLMs) through geometry and mathematical physics. Starting from a vocabulary embedding matrix $E \in \mathbb{R}^{N\times d}$, the paper identifies an intrinsic token semantic space $\mathbb{R}^r$, where r represents the effective semantic rank of the embedding representation. By adding the token sequence dimension as a temporal coordinate, we create a pseudo-time dimension; as such, the first ambient space is extended to a temporal-semantic ambient space $\mathbb{R}^{r+1}$. Observed language is then treated as discrete token samples or trajectories approximated, at first order, by a language manifold $M \subset \mathbb{R}^{r+1}$.

A scalar semantic potential Φ is introduced on the language manifold, and the token semantic vector is modeled as $v = \nabla\Phi$. In this formulation, the r-dimensional semantic space is analogized to a spatial field, while the token sequence dimension is treated as a pseudo-time domain. A token sequence can therefore be expressed as an ordered point cloud or trajectory embedded in the ambient space $\mathbb{R}^{r+1}$. However, language with semantic meaning tends to concentrate in smaller regions of this ambient space, which can be approximated by continuous manifolds. The diffusion equation provides a natural first candidate for fitting continuous manifolds to discrete linguistic samples, while wave and transport equations capture semantic propagation, structure preservation, and directional movement under contextual constraints. Together, these equations form a PDE-based framework for modeling language dynamics on the language manifold.

Training is interpreted as an inverse problem: estimating the language manifold, the scalar potential structure, and the coefficient fields of the governing PDE from human-generated language. Inference is interpreted as the forward problem: a prompt imposes boundary or initial conditions and selects a continuation trajectory on the learned manifold. The framework offers a path from statistical pattern recognition toward a predictive theory of language dynamics grounded in manifold geometry and PDEs.

1. Introduction: The Interpretability Crisis and the Missing Law

The dominant direction in LLM development, increasing model scale and improving benchmark performance, has delivered remarkable engineering achievements. Yet a persistent explanatory gap remains: LLM modeling remains largely biomimetic, attempting to emulate brain-like neural architecture and neural processing without fully uncovering the physical mechanisms that govern intelligence itself.

This situation is characteristic of a Stage 2 science. Following an epistemological ladder of scientific progress, a discipline may be viewed as passing through four stages:

1. **Stage 1.** Raw observation and data collection.

2. **Stage 2.** Phenomenological pattern discovery, including scaling laws, empirical regularities, and benchmark correlations.

3. **Stage 3.** Identification of a governing theoretical framework that explains Stage 2 patterns as consequences.

4. **Stage 4.** Full mathematical formalization with predictive and quantitative power.

As an analogy, Galileo's observations corresponded to Stage 1, while Kepler's three empirical laws represented Stage 2 phenomenological pattern discovery in planetary motion. Newtonian mechanics then provided the Stage 3 theoretical framework and ultimately closed the knowledge loop by deriving those empirical patterns from first principles. Much of current LLM research remains largely in Stage 2.

The present paper first constructs two ambient spaces. The first is an intrinsic semantic space, analogized to a spatial domain. The second extends this semantic space by adding the token sequence dimension as a pseudo-time coordinate, thereby forming a temporal-semantic ambient space. Establishing this temporal-semantic space makes it possible to draw on existing tools from mathematical physics, where PDEs connect evolution in the time domain with variation in the spatial domain. In this framework, the governing PDE constrains the fitted language manifold, or family of manifolds, and its solutions represent admissible language trajectories or outputs.

2. Historical Lineage: From Energy Landscapes to Governing Laws

The present framework belongs to the long tradition of energy-based thinking in artificial intelligence. Boltzmann's statistical mechanics introduced the idea that macroscopic states may emerge from microscopic configurations organized by an energy landscape. Hopfield networks later translated this idea into artificial cognition by representing stored patterns as attractor basins of a discrete dynamical system [Hopfield, 1982]. The Boltzmann machine extended this view through stochastic learning and probability-weighted energy states [Hinton and Sejnowski, 1986]. LeCun's energy-based models further generalized the framework by assigning a scalar energy $E(x, y)$ to input–output pairs and interpreting inference as energy minimization [LeCun et al., 2006].

These approaches establish an important lineage: cognition can be understood through landscapes, attractors, and minimization. However, they do not fully specify the governing differential law that determines how a system evolves on the landscape. In other words, they provide energy functions or learning algorithms, but not the PDE that connects local changes in semantic potential, sequence evolution, and manifold geometry. This missing governing law is the Stage 3 gap addressed in this paper.

3. The Geometric Stack: From Vocabulary to Ambient Space

The framework is built in two sequential steps. The first step operates at the vocabulary level and constructs an intrinsic semantic space for individual tokens. The second step operates at the sequence level and appends a temporal coordinate corresponding to token order. Together, these steps define the mathematical arena in which language can be modeled as a manifold and, eventually, as a field governed by a partial differential equation. The stack may be summarized as follows:

$$E \in \mathbb{R}^{N\times d} \longrightarrow v_k \in \mathbb{R}^r \longrightarrow (i,\, v(i)) \in \mathbb{R}^{r+1} \longrightarrow M \subset \mathbb{R}^{r+1} \longrightarrow \text{prompt as boundary conditions.}$$

The first arrow is a vocabulary-level reduction that removes redundant embedding coordinates; the second arrow is a sequence-level extension that adds time.

3.1. Step 1: The Token Semantic Space $\mathbb{R}^r$

Following standard embedding practice, a language model represents its vocabulary by an embedding matrix

$$E \in \mathbb{R}^{N \times d},$$

where N is the number of tokens and d is the model-imposed latent embedding dimension. The dimensionality of word embeddings has been studied theoretically by Yin and Shen [2018], who show that dimensionality selection can be analysed as a mathematical problem rather than treated as an arbitrary model parameter choice.

Each row of E represents one token as a vector

$$e_k \in \mathbb{R}^d, \quad k = 1, \ldots, N.$$

The vocabulary embedding matrix therefore represents N discrete token embeddings as a finite point cloud in the continuous latent space $\mathbb{R}^d$:

$$\{e_k\}_{k=1}^N \subset \mathbb{R}^d.$$

Although each token embedding lives in $\mathbb{R}^d$, the finite set of vocabulary embeddings may span only an r-dimensional subspace of $\mathbb{R}^d$, where

$$r = \operatorname{rank}(E), \quad r \leq \min(N, d).$$

Equivalently, $\operatorname{span}\{e_1, \ldots, e_N\} \cong \mathbb{R}^r$. Here r denotes the intrinsic linear semantic dimension of the vocabulary representation. This rank-based view is consistent with empirical work by Kataiwa et al. [2025], who find that token embedding spaces often have much lower intrinsic dimensionality than their extrinsic embedding dimensions. We therefore define the first ambient space as

$$\mathcal{A}_1 = \mathbb{R}^r.$$

3.2. Step 2: The Temporal-Semantic Ambient Space $\mathbb{R}^{r+1}$

If the intrinsic token semantic space $\mathbb{R}^r$ is interpreted as an r-dimensional spatial-like vector space, then a token sequence can be viewed as the evolution of a semantic state along an additional temporal dimension. The sequence index plays the role of time, while the r semantic coordinates play the role of spatial dimensions.

This discrete-to-continuous interpretation is consistent with the Neural ODE framework of Chen et al. [2018], who show that discrete neural-network transformations can be interpreted as an Euler discretisation of continuous dynamics. In their formulation, the hidden state evolves according to $\mathrm{d}h/\mathrm{d}t = f_\theta(h, t)$, rather than through a fixed finite sequence of layers. Although the present paper studies token trajectories on a semantic manifold rather than hidden-layer dynamics, the same mathematical idea motivates treating a discrete sequence index as a continuous temporal variable.

Let $i = 1, 2, \ldots, n$ denote the token position in a sequence. At each position i, the token is represented by a semantic vector $v(i) \in \mathbb{R}^r$. Each contextualised token is therefore a pair $(i, v(i))$ lying in the product space $\{1, \ldots, n\} \times \mathbb{R}^r$. In the continuum approximation the discrete index i may be treated as a temporal variable t, and the product space is approximated by the temporal-semantic ambient space

$$\mathcal{A}_2 = \mathbb{R}^{r+1}.$$

Thus the second ambient space is obtained by appending one temporal dimension to the intrinsic token semantic space: $\mathbb{R}^r \to \mathbb{R}^{r+1}$. We write $\mathcal{A} \equiv \mathcal{A}_2 = \mathbb{R}^{r+1}$ for the ambient space used throughout the remainder of the paper.

In this formulation, language can be understood as a set of discrete observed points or trajectories lying on, or near, a continuous manifold embedded in $\mathbb{R}^{r+1}$. The discrete tokens are the sampled observations, while the underlying language manifold represents the continuous geometric structure that constrains admissible linguistic sequences. We denote this manifold by

$$M \subset \mathbb{R}^{r+1}.$$

This manifold contains the admissible trajectories of language: sequences that are grammatically coherent, semantically meaningful, and pragmatically interpretable within a linguistic community. Language is not designed from first principles. It evolves through repeated human use, social selection, communicative efficiency, memory constraints, and cultural transmission. Over time, this evolutionary process causes admissible linguistic patterns to converge toward stable structures in the temporal-semantic ambient space. In the terminology of this paper, these stable structures define a basin of attraction, which we identify as the language manifold M. The manifold is not created by a large language model; rather, it is gradually shaped by the historical evolution of human language and preserved in the fossil record of human-generated texts.

The observed tokens and token sequences are discrete samples in $\mathbb{R}^{r+1}$. Collectively, they form a point cloud in the temporal-semantic space $\mathcal{A}$. The language manifold $M \subset \mathbb{R}^{r+1}$ is the continuous lower-dimensional structure fitted to, or inferred from, this point cloud. The role of the governing PDE is to provide a smooth constraint on this fit, so that admissible linguistic trajectories are not arbitrary curves through the ambient space, but paths lying on or near M.

Section 4 turns to the governing law: what type of partial differential equation governs evolution, diffusion, propagation, or collapse on M?

4. Governing PDE on the Language Manifold

When a manifold is embedded in a Euclidean or Riemannian space, it may be viewed as a lower-dimensional geometric structure within the ambient space. A flat plane or affine subspace can be described by linear equations, while a curved manifold often requires differential operators to characterise its geometry. In this sense, a PDE provides a general mathematical framework for representing, constraining, or recovering the shape of a manifold. In the language setting, observed tokens or token sequences can be understood as discrete sample points concentrated in a region of the temporal-semantic ambient space. A governing PDE may then be interpreted as a smooth, PDE-constrained fit to this cloud of discrete linguistic samples. Among possible candidates, the diffusion equation is especially useful because diffusion geometry shows that the Laplacian associated with heat flow can encode and recover the geometry of an underlying manifold from sampled data [Jones, 2024]. Recent work on implicit manifold-valued diffusions further supports this view by constructing diffusion processes on data manifolds from point-cloud samples without requiring explicit charts or projections [Kawasaki-Borruat et al., 2026]. Therefore, a diffusion-type equation provides a natural first candidate for modeling the language manifold, while wave and transport equations may capture additional effects such as semantic propagation and directional flow.

4.1. Building the Diffusion Equation

To build the diffusion equation on the language manifold, we first introduce a scalar potential field $\Phi(t, x)$, where $t \in [1, n]$ denotes the (continuous approximation to the) sequence position and $x \in \mathbb{R}^r$ are the semantic coordinates on the manifold. The scalar potential Φ is analogous to temperature T in heat transfer.

In heat transfer, temperature is a scalar field and its gradient ∇T is a vector field describing the rate and direction of temperature change. By analogy, the spatial gradient of the semantic

potential,

$$\nabla\Phi \in \mathbb{R}^r,$$

is a rank-1 tensor (vector field) lying on the local tangent space of the language manifold. Here ∇ denotes differentiation with respect to the semantic coordinates x; the time derivative is always written explicitly as $\partial/\partial t$. A token at sequence position t is expressed through this gradient-derived vector. We define the token semantic vector as

$$v = \nabla\Phi.$$

This gives v two related meanings. First, it represents the multidimensional semantic attributes of the token, with each component corresponding to one intrinsic semantic direction. Second, it represents the gradient of semantic potential, indicating the local direction and rate of semantic change on the manifold.

In the heat-transfer analogy, the rate of change of the scalar potential in the time domain is proportional to the Laplacian of that potential in the spatial domain. Therefore, in the local flat approximation, the semantic diffusion equation is

$$\frac{\partial\Phi}{\partial t} = \kappa\,\nabla^2\Phi,$$

where $\kappa > 0$ is the semantic diffusion coefficient. In the heat-transfer analogy, κ corresponds to the thermal diffusivity of the medium, which controls how rapidly temperature changes in time in response to spatial temperature curvature. Here, the "time domain" corresponds to the token sequence dimension, while the "spatial domain" corresponds to the intrinsic semantic directions on the language manifold.

The ordinary Euclidean Laplacian ∇^2 should be understood as a local flat approximation. If the language manifold is globally curved, ∇^2 should be replaced by the Laplace–Beltrami operator ∇^2_M. In a standard three-dimensional heat-transfer problem, the Euclidean Laplacian ∇^2 is adequate because heat diffuses through a Euclidean domain. However, if heat transfer is constrained to a curved thin shell or surface, the diffusion domain is itself a curved manifold, and the Laplace–Beltrami operator is required to account for curvature. By analogy, language dynamics on a curved language manifold should be written as

$$\frac{\partial\Phi}{\partial t} = \kappa\,\nabla^2_M\Phi.$$

Here ∇^2_M incorporates the Riemannian metric of the language manifold. In local coordinates, this introduces metric-dependent terms, equivalently expressed through covariant derivatives or Christoffel symbols.

In the language manifold, semantic diffusion is unlikely to be isotropic. The intrinsic semantic directions are not independent, and activation does not spread equally in every direction. Some dimensions may be strongly coupled, such as subject, tense, agency, modality, causality, and discourse role, while others may interact only weakly. The scalar diffusion coefficient κ should therefore be replaced by a semantic conductivity tensor K. In the local flat approximation the anisotropic diffusion equation becomes

$$\frac{\partial\Phi}{\partial t} = \nabla\cdot(K\,\nabla\Phi).$$

The diagonal entries of K describe diffusion along individual semantic directions; the off-diagonal entries describe coupling between different semantic dimensions. If both curvature and anisotropy are considered, the most general diffusion form becomes

$$\frac{\partial\Phi}{\partial t} = \mathrm{div}_M(K\,\nabla_M\Phi),$$

which is the natural generalization of heat diffusion to an anisotropic semantic field on a curved language manifold.

4.2. General PDE Forms for Geometry Fitting

The diffusion equation provides a useful geometric fitting tool because it connects time-domain evolution with spatial-domain curvature and attraction toward stable structures. However, language dynamics are unlikely to be purely diffusive, like water cascading downhill. Language does not merely smooth or dissipate semantic activation. It also preserves structure, exhibits resonance, propagates long-range dependencies, and moves directionally under contextual constraints.

For this reason, the wave equation may be more consistent with certain features of the language manifold, especially where semantic structures persist and propagate across long sequences. Similarly, transport equations may be needed to describe directional semantic movement under the influence of context. Therefore, rather than relying on a single diffusion equation, this paper considers a broader family of PDE-based geometric fitting models.

The common feature of this PDE family is that each equation connects evolution in the time domain with structure in the spatial domain. In the language setting, the time domain corresponds to the token sequence dimension, while the spatial domain corresponds to intrinsic semantic directions on the language manifold. These equations therefore provide different ways to represent token flow following the gradient structure of semantic potential.

4.3. Wave Equation: Semantic Propagation and Structure Preservation

To model the preservation and propagation of linguistic structure, we introduce the wave equation. In the flat approximation:

$$\frac{\partial^2 \Phi}{\partial t^2} = c^2 \, \nabla^2 \Phi.$$

On the curved language manifold, this becomes

$$\frac{\partial^2 \Phi}{\partial t^2} = c^2 \, \nabla_M^2 \Phi,$$

where c represents the propagation velocity of semantic structure.

A more general form includes a damping term, which provides additional flexibility for fitting semantic dynamics:

$$\frac{\partial^2 \Phi}{\partial t^2} + 2\beta \, \frac{\partial \Phi}{\partial t} = c^2 \, \nabla_M^2 \Phi,$$

where $\beta > 0$ is the damping coefficient. This term allows temporary semantic effects to decay while stable discourse structures continue to propagate. Thus, the wave equation may fit the part of language dynamics involving semantic memory, structure preservation, and long-range dependency.

4.4. Transport Equation: Directional Semantic Flow

A third candidate is the transport equation. In the flat approximation,

$$\frac{\partial \Phi}{\partial t} + u \cdot \nabla \Phi = 0,$$

where u is a semantic velocity field tangent to the language manifold. The vector field u is different from the token semantic vector $v = \nabla \Phi$.

While v represents the intrinsic semantic attributes or local semantic gradient of a token, u represents the direction and speed of semantic transport along the manifold. This distinction gives the model additional flexibility: token flow does not have to follow only the gradient of the token's intrinsic attributes. It may also be guided by contextual direction, discourse intention, or prompt-imposed constraints.

On a curved manifold, the transport equation becomes

$$\frac{\partial \Phi}{\partial t} + u \cdot \nabla_M \Phi = 0,$$

where u lies on the tangent space of M and $\nabla_M \Phi$ is the intrinsic gradient on the manifold.

4.5. Toward a General Governing PDE

The three PDE forms can be combined into a more general governing template:

$$\frac{\partial^2 \Phi}{\partial t^2} + 2\beta \frac{\partial \Phi}{\partial t} + u(t) \cdot \nabla_M \Phi = \mathrm{div}_M\big(K \nabla_M \Phi\big) + S(t). \tag{1}$$

In this equation:

- $\partial^2 \Phi / \partial t^2$ represents wave-like propagation,
- $2\beta\, \partial \Phi / \partial t$ represents damping or dissipation,
- $u(t) \cdot \nabla_M \Phi$ represents transport along a semantic flow field,
- $\mathrm{div}_M(K \nabla_M \Phi)$ represents anisotropic diffusion on the curved language manifold, and
- $S(t)$ represents source terms introduced by new tokens, external context, or multimodal input.

This general PDE is not proposed as the final equation of language. Rather, it is a governing-law template: it shows how language dynamics may combine smoothing, propagation, and directional flow on the language manifold.

5. Training and Prompting as Inverse and Forward PDE Problems

5.1. Training as an Inverse PDE Problem

The governing PDE of the language manifold is not known in advance. Neither the exact form of the equation nor its coefficient fields are directly given. Therefore, LLM training may be interpreted as an inverse PDE problem.

The fossil record of human-generated language provides discrete samples of trajectories lying on, or near, the language manifold M. Language is not constructed from an architectural blueprint. It evolves through repeated human use, social selection, communicative efficiency, memory constraints, and cultural transmission. Over time, admissible linguistic patterns converge toward stable structures in the temporal-semantic ambient space. In the terminology of this paper, these stable structures define an attractor basin, which we identify as the language manifold M.

The language manifold is embedded in the second ambient space, $\mathcal{A} \equiv \mathcal{A}_2 = \mathbb{R}^{r+1}$, which is a linear temporal-semantic space. From this perspective, LLM training is not the construction of language from raw data, but the search for, and approximation of, a pre-existing language manifold within $\mathcal{A}$. The learned model estimates the geometry of M from discrete linguistic samples.

Training also estimates the scalar potential structure Φ and the coefficient fields of the governing PDE, including the semantic conductivity tensor $K(t)$, the transport vector field $u(t)$, the damping coefficient β, and the source term $S(t)$. These quantities are not prescribed in advance. They are inferred from the observed distribution of human-generated language.

In this interpretation, model weights are not merely memorised statistical associations. They are distributed approximations of the geometry and governing dynamics of the language manifold. The model learns how semantic activation diffuses, propagates, and transports across M under different contextual and boundary conditions.

5.2. Prompting as Boundary Condition, Slicing, and Collapse

Inference is the corresponding forward problem. Given a prompt P, the model imposes boundary or initial conditions on the language manifold M. These constraints produce a prompt-conditioned scalar potential field Φ_P.

The prompt is therefore not merely an input string. It plays a central role in reducing the admissible region of the language manifold. In the temporal-semantic formulation, the ambient space $\mathcal{A} = \mathbb{R}^{r+1}$ is a linear coordinate space, while the language manifold $M \subset \mathbb{R}^{r+1}$ is a lower-dimensional Riemannian manifold embedded within it. The prompt supplies a finite set of token positions and semantic states, and therefore fixes the part of the trajectory that any admissible continuation must satisfy.

This reduction can be understood in two complementary ways. From a linear-algebra perspective, the prompt imposes a system of constraints on the full temporal-semantic ambient space. The query-key-value mechanism supplies contextual relationships that restrict the admissible degrees of freedom, similar to using equations to reduce the dimensionality of a solution space. From a PDE perspective, the prompt supplies boundary or initial conditions for the governing equation. The model response is then the continuation trajectory selected under these constraints.

In transformer architectures, this constraint is operationalised through the query-key-value attention mechanism. The prompt tokens generate queries, keys, and values that determine which prior semantic states are relevant to the next state. Geometrically, this process can be interpreted in three related ways. First, the prompt imposes boundary values by fixing known points on the manifold. Second, it slices the manifold by assigning values to certain contextual or semantic dimensions, thereby reducing the admissible solution set to a lower-dimensional submanifold. Third, through attention-weighted aggregation, it may collapse or average information across selected token positions, producing a lower-dimensional effective representation of the context.

Thus, prompting is a geometric operation on the language manifold. By imposing boundary conditions, fixing parameters, slicing admissible dimensions, and aggregating relevant information, the prompt progressively reduces the full language manifold to a constrained admissible region. The output of a language model can therefore be interpreted as a trajectory selected from this reduced region of M. The response is a continuation on the language manifold under constraints supplied by the prompt.

The admissible continuation may also be interpreted as a stable or low-energy path on the manifold, analogous to a physical system evolving toward an attractor or minimum-free-energy configuration.

The relationship between training and inference can therefore be summarized as follows. Training estimates the geometry of M, the scalar potential structure Φ, and the coefficient fields governing semantic diffusion, propagation, and transport. Prompting then imposes boundary or initial conditions on this learned structure, reducing the admissible region of the manifold. The response is the continuation trajectory selected from this constrained region.

6. Discussion, Limitations, and Future Research

The framework proposed in this paper is supported by several related lines of existing work.

First, the reduction from the model-imposed latent dimension d to the intrinsic semantic rank r is motivated by recent work on the intrinsic dimensionality of token embeddings. Kataiwa et al. [2025] suggest that token embeddings contain substantial redundancy and may lie in a lower-dimensional intrinsic representation space. This supports the first step of the present

framework: replacing the full embedding space $\mathbb{R}^d$ with an intrinsic token semantic space $\mathbb{R}^r$, where r captures the effective number of independent semantic directions.

Second, diffusion geometry provides a mathematical foundation for fitting continuous geometric structures to discrete point-cloud samples. Jones [2024] shows that diffusion operators, Laplacians, and carré-du-champ structures can recover geometric information from sampled data. Kawasaki-Borruat et al. [2026] further develop diffusion processes on implicit manifolds, supporting the idea that a manifold can be inferred from point-cloud data through diffusion-based operators. This directly supports the interpretation of observed token sequences as discrete samples lying on, or near, a continuous language manifold M.

Third, Neural ODEs provide a precedent for introducing a time dimension into neural computation. Chen et al. [2018] interpret discrete neural-network updates as an Euler discretisation of continuous dynamics and replace a fixed sequence of hidden layers with an ordinary differential equation for the hidden state. This supports the second step of the present framework: treating the discrete token index as a continuous pseudo-time variable t. The time dimension is the missing piece that connects the semantic spatial domain $\mathbb{R}^r$ with sequence evolution, producing the temporal-semantic ambient space $\mathbb{R}^{r+1}$. Once this space is constructed, PDEs become natural candidates for governing the relationship between time-domain evolution and spatial-domain semantic structure.

Fourth, work on data-driven discovery of intrinsic dynamics provides a useful precedent for recovering lower-dimensional dynamics from high-dimensional observations [Floryan and Graham, 2022]. This is relevant to the present framework because the assumption of a continuous language manifold should ultimately be treated as an approximation rather than as a completed geometric characterization. In high-dimensional embedding space, lower-dimensional language structure may appear as a smooth manifold, but it may also appear as a cluster structure, graph-like structure, stratified space, singular manifold-like set, or other non-smooth geometric object. Therefore, the language manifold proposed in this paper should be understood as a first smooth approximation to the underlying semantic structure.

Fifth, the proposed language-manifold framework is consistent with a broader two-ambient-space approach to representation learning. In a related formulation for image-based world models, Zhang [2026] introduces a two-ambient-space continuum reformulation of I-JEPA, in which the image-plane chart domain is connected with the patch-content domain through an advection-diffusion PDE. The present paper applies the same general principle to LLMs: an intrinsic semantic ambient space is first constructed and then extended by a pseudo-time coordinate, allowing PDE-based tools to connect sequence evolution with semantic variation.

In the present framework, the term "initial state" has two related but distinct meanings. First, the pretrained language manifold may be regarded as an unactivated background semantic landscape, denoted M_0, whose learned geometry already contains latent gradients and admissible directions of semantic evolution. Second, once a prompt or first token is introduced, it defines an activated initial point or state $q_0 \in M_0$. The subsequent states then evolve from q_0 along the learned semantic landscape, forming a trajectory whose observable trace corresponds to the generated language output.

These lines of work correspond to the major steps of this paper:

$$\begin{aligned} \text{intrinsic rank reduction} &\longrightarrow \mathbb{R}^r, \\ \text{addition of pseudo-time dimension} &\longrightarrow \mathbb{R}^{r+1}, \\ \text{diffusion/PDE fitting} &\longrightarrow M \subset \mathbb{R}^{r+1}. \end{aligned}$$

The current proposal should be understood as a Stage 3 framework rather than a completed predictive theory. Future work may further develop this framework toward Stage 4 by providing

empirical validation, quantitative operators, and more rigorous geometric formulations. Several open problems remain: estimating the intrinsic semantic rank r; testing whether $v = \nabla\Phi$ is empirically valid; replacing the local flat operator ∇^2 with the manifold operator ∇^2_M; identifying which PDE terms are required for language dynamics; and recovering coefficient fields such as K, u, β, and S from trained model activations.

A further direction is to develop an atlas-based formulation of language geometry. Instead of assuming one globally smooth language manifold, the global semantic structure may be covered by local charts or manifold-like patches M_α, each representing a local semantic, syntactic, pragmatic, or domain-specific region. Transition maps and compatibility conditions between neighboring charts would then be needed to determine whether these local structures form a smooth manifold, a stratified manifold, or a more general geometric object.

Another direction is to connect manifold identification with a fibre-bundle-like factorization of the embedding space. Identifying a lower-dimensional semantic structure and constructing such a factorization can be viewed as two complementary forms of dimensional reduction. The former seeks a lower-dimensional semantic substructure within the high-dimensional embedding space, while the latter decomposes the total embedding space into a task-relevant semantic base and auxiliary fibres that may represent style, context, modality, domain, or redundant degrees of freedom. The projection from the total embedding space to a lower-dimensional base space can be interpreted as eliminating, collapsing, or marginalizing variations along the fibre. However, this separation is task-dependent: a dimension that is irrelevant for one task may be important for another. Therefore, a rigorous formulation would require a defined projection map or a learned procedure for determining which variations belong to the semantic base and which belong to the auxiliary fibres.

Finally, the rigorous formulation of PDE operators on the probability simplex is left for future study. Since LLM outputs are probability distributions over vocabulary tokens, a complete theory should eventually connect the temporal-semantic manifold formulation with operators defined on, or coupled to, the probability simplex. This would clarify how semantic potential fields, token probabilities, and manifold dynamics interact during inference.

Although the present paper focuses on LLMs, the framework may provide a bridge from language modeling to real-world AI. LLMs are the most accessible case because token embeddings and sequence positions naturally define a temporal-semantic ambient space. In industry-specific or discipline-specific AI, the same idea may be extended by replacing the language manifold with a domain manifold built from engineering, financial, legal, biological, or operational data. In this broader view, LLMs are not the endpoint, but the first workable example of a general framework in which intelligence is modeled as manifold learning plus governing dynamics. This provides a possible path from LLM transparency toward real-world AI, industry-specific AI, discipline-specific AI, and eventually AGI.

7. Conclusion

This paper has proposed a Stage 3 theoretical framework for understanding large language models through the geometry of language manifolds and the dynamics of partial differential equations. The two-step construction, from vocabulary embeddings to an intrinsic semantic space $\mathbb{R}^r$, and from there to a temporal-semantic ambient space $\mathbb{R}^{r+1}$, provides a principled arena in which language can be modeled as a continuous manifold fitted to discrete token trajectories. Training is interpreted as the inverse problem of recovering this manifold and its governing PDE from human-generated text; inference is interpreted as the corresponding forward problem, in which a prompt imposes boundary conditions and selects an admissible continuation trajectory.

The framework does not claim to be a completed quantitative theory. Rather, it offers a Stage 3

conceptual bridge between the empirical scaling laws of Stage 2 and the predictive, first-principles formalism of Stage 4. The open problems identified in the Discussion, intrinsic rank estimation, empirical validation of $v = \nabla\Phi$, atlas-based geometry, fibre-bundle factorization, and PDE coefficient recovery from model activations, mark the path forward toward a fully quantitative theory of language dynamics.

References

J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

G. E. Hinton and T. J. Sejnowski. Learning and relearning in Boltzmann machines. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, volume 1, chapter 7, pages 282–317. MIT Press, 1986.

Y. LeCun, S. Chopra, R. Hadsell, M. A. Ranzato, and F. J. Huang. A tutorial on energy-based learning. In G. Bakir, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, and S. Vishwanathan, editors, *Predicting Structured Data*. MIT Press, 2006.

W. Yin and Y. Shen. On the dimensionality of word embedding. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

H. Kataiwa, K. Cho, and R. Ohki. On the intrinsic dimensionality of token embeddings in large language models: redundancy, rank, and the geometry of representation spaces. Preprint, 2025.

M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

V. Kawasaki-Borruat, C. Grotehans, P. Vandergheynst, and A. Gosztolai. Diffusion processes on implicit manifolds. *arXiv preprint arXiv:2604.07213*, 2026.

I. Jones. Diffusion geometry. *arXiv:2405.10858*, 2024.

D. Floryan and M. D. Graham. Data-driven discovery of intrinsic dynamics. *arXiv preprint arXiv:2108.05928*, 2022.

L. Zhang. A two-ambient-space continuum reformulation of I-JEPA: Spatial-patch geometry, scalar visual potentials, and PDE-guided manifold fitting. Preprint, Chicago Booth 2026 Workshop on World Models, 2026. https://doi.org/10.5281/zenodo.20414034.