ISSN: 2993-5377

Research Article

Open Access Journal of Applied Science and Technology

Quantum Hierarchy for Understanding LLM Representations by Modeling Linear Projections and Nonlinear Dynamics

Timo Aukusti Laine*

Financial Physics Lab, Finland

***Corresponding Author**
Timo Aukusti Laine, Financial Physics Lab, Finland.

Submitted: 2026, Jan 23; **Accepted:** 2026, Feb 27; **Published:** 2026, Mar 04

Citation: Laine, T. A. (2026). Quantum Hierarchy for Understanding LLM Representations by Modeling Linear Projections and Nonlinear Dynamics. *OA J Applied Sci Technol, 4*(1), 01-43.

Abstract

Large Language Models (LLMs) excel in natural language tasks, but their high-dimensional embedding spaces pose significant interpretability challenges. Current approaches often linearize these spaces, overlooking the complex dynamics inherent in Transformer architectures. This article proposes a quantum framework to analyze LLM representations, leveraging quantum mechanical tools to explore semantic relationships and contextual influences. We introduce a layered hierarchy of semantic spaces and demonstrate that a classical LLM embedding system has an exact quantum mechanical analogue. Using this analogue, we model phenomena such as the modulation of Semantic Noise, the emergence of hallucinations via quantum tunneling, and the formation of stable semantic representations as soliton solutions. Furthermore, we present a simple quantum circuit design, demonstrating the possibility of using quantum computers to probe, analyze and go beyond real-valued LLM embedding spaces, potentially revealing structural information and relationships not readily accessible through classical techniques. This perspective enhances our understanding of LLM representations, leading to improved methods for analyzing and controlling LLM behavior, and supporting research into more efficient, reliable, and trustworthy AI systems.

1. Introduction

Large Language Models (LLMs) have achieved remarkable success in natural language processing, yet their underlying mechanisms remain largely opaque. The high-dimensional embedding spaces that power LLMs pose significant challenges to interpretability, hindering our ability to fully understand and control their behavior. Current approaches often rely on linearizing these spaces, an oversimplification that overlooks crucial nonlinear dynamics, such as those present in the Transformer architecture. Specifically, the attention mechanism, a core component of the Transformer, introduces complex, non-local interactions between words, which are not adequately captured by static, linear representations. Furthermore, the activation functions within the feedforward networks introduce nonlinear transformations that shape the semantic landscape in intricate ways.

The development and deployment of current LLMs demand vast computational resources, making them expensive to train and use. This raises a critical question: Can we develop more efficient LLM architectures, or find ways to optimize the use of existing LLM architectures? Also, the tendency of LLMs to generate factually incorrect or nonsensical statements (hallucinations) limits their reliability and hinders their adoption in business-critical applications. These hallucinations are not simply random errors; they often exhibit a degree of semantic coherence, suggesting that they arise from complex interactions within the LLM's internal representations. A deeper understanding of the origins and dynamics of hallucinations is essential for building more trustworthy and robust LLMs.

To address these challenges, this article proposes a quantum framework for analyzing LLM representations. We leverage the tools and concepts of quantum mechanics to explore semantic relationships, contextual influences, and the underlying dynamics of LLM embedding spaces. Our approach is motivated by the discrete nature of LLM embedding spaces and the inherent uncertainties associated with semantic meaning, both of which find natural parallels in quantum mechanics. This work synthesizes and extends our previous

research, providing a unified perspective on these quantum-inspired models while also introducing new insights and, crucially, establishing a concrete link to quantum computation, demonstrating the potential for using quantum computers to directly probe and analyze LLM embedding spaces.

The key contributions of this article are:

- A Layered Quantum Hierarchy: We introduce a layered hierarchy of semantic spaces, ranging from the linearized embedding space to the full Transformer architecture, with intermediate layers incorporating linear and nonlinear quantum dynamics.
- An Exact Quantum Analogue: We demonstrate the existence of a classical LLM embedding system and a quantum mechanical LLM Embedding Quantum System whose mathematical descriptions are equivalent, providing a concrete link between the classical and quantum domains.
- Quantum Explanations: We show how the quantum system can be used to explain phenomena such as the management of Semantic Noise (through local U(1) symmetry), hallucinations (through quantum tunneling), and the emergence of Self-Sustaining Semantic Structures, i.e., stable, localized patterns of semantic meaning which we also refer to as localized semantic stability (through soliton solutions).
- A Path to Quantum Computation: We present a simple quantum circuit design for calculating cosine similarity, demonstrating the potential for using quantum computers to probe and analyze LLM embedding spaces.

This quantum perspective offers a framework for understanding LLM representations, leading to improved methods for analyzing and controlling LLM behavior. By providing a framework for applying quantum computing techniques to LLMs, we promote the way for future research into more efficient, reliable, and trustworthy AI systems, leveraging the insights gained from our quantum analysis, particularly through a deeper understanding of semantic relationships and the mitigation of hallucinations.

2. Background and Related Work

Large Language Models (LLMs) have revolutionized natural language processing, achieving remarkable feats in text generation, translation, and question answering. Trained on massive datasets of text and code, these models learn to predict the next word in a sequence, a process that gives rise to their emergent capacity for sophisticated language understanding and generation [1]. A core component of LLMs is the embedding space: a high-dimensional vector space where words, phrases, and even entire documents are represented as points. The location of each point is carefully learned to reflect the semantic meaning of the corresponding linguistic unit. Semantically similar words, or those used in similar contexts, are clustered together, enabling the model to recognize and exploit complex relationships within language.

Early techniques like Word2Vec and GloVe pioneered the development of word embeddings, mapping words to vectors where proximity reflects semantic similarity [2,3]. These advancements built upon foundational work by Bengio et al. on neural probabilistic language models and the principles of distributional semantics, which posits that a word's meaning is intrinsically linked to the contexts in which it appears [4,5]. Research on semantic compositionality, such as that by Socher et al., explores how the meanings of individual words combine to form the meaning of larger phrases and sentences [6]. Modern LLMs, particularly those based on the Transformer architecture, leverage contextualized word embeddings, where a word's meaning is not fixed but dynamically determined by the surrounding words in the sentence [1]. This contextual sensitivity, enabled by the Transformer's attention mechanisms, allows for a more nuanced and flexible representation of language, overcoming limitations of earlier, static word embeddings. The attention mechanism also facilitates parallel processing and improved handling of long-range dependencies.

While embedding spaces, analyzed through techniques like cosine similarity, have proven incredibly powerful, they offer an inherently incomplete representation of the complex internal state of an LLM. Projecting the model's intricate dynamics onto a linear vector space inevitably results in a loss of information. Certain dynamic and nonlinear aspects of semantic meaning are simply not captured by these static representations. This inherent incompleteness can manifest in various ways, including the generation of factually incorrect or nonsensical statements, commonly known as hallucinations. Recent surveys have highlighted the prevalence and diverse nature of hallucinations in natural language generation. This problem is closely related to issues of factuality and knowledge representation in LLMs, as explored by Ji et al. and the challenges of relying solely on statistical correlations without a deeper understanding of causality [7,8]. While techniques based on Shannon's information theory and Kullback-Leibler divergence offer tools for quantifying information and measuring differences between probability distributions, they often fall short of capturing the subtle nuances of semantic representation that are crucial for understanding and mitigating issues like hallucinations in LLMs [9-12].

The principles of quantum mechanics, while developed for the physical world, offer a unique perspective on handling uncertainty and contextuality that may be relevant to understanding LLMs. The potential for quantum computation to enhance machine learning is an active area of research [13,14]. Quantum algorithms, such as Grover's algorithm and Shor's algorithm, offer potential speedups for certain

computational tasks, and quantum circuits are being explored for use in machine learning models [15-17]. DiVincenzo's criteria for physical implementations of quantum computation highlight the challenges of building actual quantum computers, but also underscore the fundamental principles that govern quantum information processing. Moreover, the application of quantum-inspired methods to machine learning, such as quantum-enhanced feature spaces and quantum neural networks, demonstrates the potential for quantum concepts to improve classical machine learning algorithms [18-20]. Nielsen and Chuang's seminal work on quantum computation and quantum information provides a solid foundation for understanding these concepts [21]. Khrennikov's work on ubiquitous quantum structure explores the application of quantum-like models in various domains, including cognition and decision-making, providing further justification for our approach. Hybrid quantum-classical approaches are also being explored to enhance LLM fine-tuning, with some studies demonstrating improved accuracy compared to purely classical models [22,23].

The Hamiltonian operator represents the total energy of a quantum system. We can adapt this concept to represent the energy landscape of semantic meaning, capturing the relative stability of different semantic states. Just as the Schrödinger equation describes the time evolution of a quantum system, we can use it to model the dynamics of semantic meaning as it unfolds over time. Concepts from statistical mechanics, such as phase transitions and critical phenomena, may provide insights into the emergent behavior of LLMs and the transitions between different semantic states [24–27]. The use of path integrals provides a powerful tool for analyzing quantum systems and may offer insights into the long-range dependencies in LLMs, as explored in our previous work [28,29]. These quantum mechanical principles, while seemingly abstract, offer a powerful framework for addressing the limitations of traditional approaches to understanding LLMs.

This motivates our exploration of concepts and mathematical tools from the seemingly disparate field of quantum mechanics. We propose to leverage quantum mechanics as a powerful analogy, providing a new perspective on the complex and uncertain nature of semantic meaning, particularly in understanding and managing Semantic Noise, a key factor influencing LLM behavior. This article builds upon our previous research exploring the application of quantum-inspired models to LLMs [29–33].

3. Quantum Semantic Hierarchy

This article proposes a framework to bridge the gap between the simplified, linearized embedding space and the intricate Transformer architecture. Our central idea is that these represent opposite ends of a spectrum, and a more nuanced understanding requires exploring intermediate layers that capture different facets of the LLM's internal representations. To achieve this, we introduce a quantum semantic hierarchy, leveraging concepts and tools from quantum mechanics to analyze LLM representations.
The quantum approach is underpinned by several key motivations.

1. The discrete nature of LLM embedding spaces suggests a formalism where semantic states are treated as distinct entities, aligning well with the state-based description in quantum mechanics.
2. The prevalence of hallucinations in LLMs points to inherent uncertainties in their processing, particularly regarding the semantic validity or factual correctness of generated content, which is a core concept in quantum mechanics.
3. As we will demonstrate in later sections, a classical "LLM Embedding System" can be defined with an exact quantum mechanical analogue exhibiting zero-point energy, providing a concrete link between the classical and quantum descriptions.
4. Additionally, the quantum mechanical analogue of the LLM Embedding System exhibits a superposition between the quantum states corresponding to the two LLM Embedding System embedding vectors, reflecting the inherent relationships between these semantic representations. These reasons, among others, provide a strong rationale for using quantum mechanics as a natural extension of the real embedding space, enabling us to leverage its tools and concepts to analyze semantic dynamics.

Based on these observations, we propose a hierarchy of intermediate layers, each progressively integrating a level of complexity and capturing different facets of the LLM's internal representations. This layered structure comprises:

- Layer 1: Linearized Embedding Space (Classical Semantic Space): This simplest layer is a direct projection of the LLM's internal representations into a low-dimensional vector space. It is valuable for basic semantic comparisons but lacks the capacity to model dynamic behaviors. As demonstrated in, the linear embedding space exhibits a discrete structure, meaning there is no continuous transformation between the embedding vectors [32]. The key focus of Layer 1 is providing a foundational, static representation of semantic relationships.
- Layer 2: Quantum Semantic Space (Linear Dynamics): This layer introduces linear dynamics by treating the embedding space as a quantum mechanical system. By incorporating complex state vectors and time dependence, we can employ the linear Schrödinger equation to model the evolution of semantic representations over time. This approach allows us to capture phenomena such as conservation of the Semantic Noise level, linking it to semantic "charge" conservation, and hallucinations, which can be related to quantum mechanical tunneling.
- Layer 3: Quantum Semantic Space (Nonlinear Dynamics): This layer builds upon Layer 2 by introducing nonlinear interactions,

enabling the modeling of more complex behaviors and feedback loops, including the emergence of stable, localized soliton solutions. These non-local interactions, as shown in, allow semantic meaning at one point to directly influence meaning at distant points, resembling quantum entanglement and capturing LLMs' ability to model long-range dependencies through the Transformer's attention mechanism [29].

- Layers 4-N: Advanced Quantum Hierarchies: These layers represent more advanced quantum models that can capture even more complex aspects of LLM behavior. They offer the potential to model dynamic creation and annihilation of semantic content.
- Layer N+1: Transformer Architecture (Common LLM Architecture): This layer represents the complete Transformer architecture, encompassing all its intricacies, with real-valued activations and parameters. It provides a concrete realization of the complex dynamics and interactions modeled in the earlier layers.
- Layer N+2: Complex Transformer Architecture (Complex Semantic Superspace): This layer builds upon Layer N+1 by extending the Transformer architecture to operate with complex-valued representations. It leverages complex-valued representations to improve robustness and performance in specific domains.

The hierarchy of the layers is shown in Figure 1.

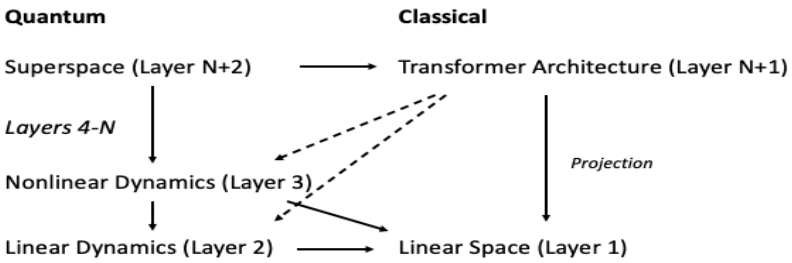


Figure 1: The Quantum Semantic Hierarchy: A multi-layered framework for understanding LLM representations. The hierarchy spans from the classical linearized embedding space (Layer 1) to the complex Transformer architecture (Layers N+1 and N+2), with intermediate quantum layers (Layers 2, 3, and 4-N) progressively incorporating linear and nonlinear dynamics. This layered approach allows for analyzing LLM behavior at different levels of abstraction, with each layer serving as a projection or sub-layer of the layer above.

The guiding principle is that each layer serves as a projection or sub-layer of the layer above it: Layer 1 projects onto Layer 2, Layer 2 onto Layer 3, and Layer 3 onto Layers 4-N. Layers 4-N are sub-layers of Layer N+2, and Layer N+1 is a sub-layer of Layer N+2. While quantum sub-layers 2 and 3 incorporate characteristics of Layer N+2, the mapping from Layer N+1 (the Transformer architecture) to these complex sub-layers is not always straightforward. The Transformer architecture possesses a complex internal structure absent in the quantum sub-layers, and conversely, the quantum sub-layers introduce a phase component not present in Layer N+1. However, as demonstrated in later sections, similarities and shared structures emerge when the system is examined in its eigenspace. Consequently, each layer captures a simplified view of the underlying complexity, with each projection resulting in some loss of information. The appropriate layer to use depends on the specific phenomenon under investigation. For instance, if we are interested in the dynamics of Semantic Noise and its relationship to hallucinations, Layer 2 (Quantum Semantic Space with Linear Dynamics) might suffice, allowing us to model the interplay between inherent uncertainty and contextual influences.

After analyzing a phenomenon using a particular layer, we can project the results back to the linearized embedding space (Layer 1) to establish connections with familiar semantic representations. This is exemplified in Section X, where we demonstrate how the properties of the linearized embedding space can be recovered by applying specific approximations to equations in Layer 2 and 3.

The layered approach offers several advantages over existing methods. By providing intermediate levels of abstraction, it allows us to analyze LLM representations with the appropriate level of complexity, avoiding the oversimplification of linear models while remaining more tractable than analyzing the full Transformer architecture directly. The quantum framework provides a new set of tools and concepts for understanding LLM behavior, leading to new insights and improved analytical techniques. Also, as a quantum-based approach, it enables us to use quantum computers for testing and validating the results.

Essentially, we are proposing a "zoom lens" approach: we focus on the level of complexity necessary to understand a particular phenomenon and then zoom out to relate our findings back to the familiar linearized embedding space. This layered strategy allows us to bridge the gap between the simplicity of the linearized embedding space and the complexity of the Transformer architecture, leading to a more nuanced and comprehensive understanding of LLM representations.

In the following sections, we will detail the characteristics of each Layers 1, 2, 3, 4-N, N+1 and N+2. We also demonstrate the

linearization process and provide an example of experimental validation, achieved using a quantum computer.

4. Layer 1: Linearized Embedding Space (Classical Semantic Space)

We begin our analysis with the simplest layer in our hierarchy: the linearized embedding space, which we also term the classical semantic space. This layer provides a foundation for understanding the fundamental properties of semantic relationships in LLMs.

4.1. The LLM Embedding System

To facilitate our analysis and gain insights into the core dynamics of LLM representations, we introduce a simplified, classical model of LLM embedding spaces, which we term the "LLM Embedding System." This system, detailed in reference, allows us to explore the fundamental properties of semantic relationships in a controlled and tractable setting, providing a crucial stepping stone to understanding the more complex quantum layers [32].

The core of the system consists of two N-dimensional vectors, a and b, both residing in a real-valued vector space. The vector a = [a_1, a_2,...,a_N] represents an arbitrary embedding vector, serving as a semantic anchor in the space. The coefficients a_i are real numbers, reflecting the real-valued nature of typical LLM embedding spaces. We assume that a is L2-normalized, meaning that $||\mathbf{a}|| = \sqrt{\sum_{i=1}^{N} a_i^2} = 1$. This L2 normalization reflects the common practice in LLM architectures and ensures that similarity is primarily determined by the angle between vectors, rather than their magnitudes.

The vector **b** is defined in relation to **a** as a perturbation of the vector that is maximally dissimilar to **a**

$$\mathbf{b} = [-a_1 + \Delta_1, -a_2 + \Delta_2, ..., -a_N + \Delta_N] \tag{1}$$

where Δ_i represents a change in each dimension relative to the negated components of **a**. Starting with the negated components allows us to model semantic shifts away from the concept represented by **a**, capturing how LLMs can deviate from a given topic or idea.

The cosine similarity between **a** and **b** is the key quantity that the LLM Embedding System is showing, and it is defined as

$$S_C(\mathbf{a}, \mathbf{b}) = \cos\theta = \frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}||\ ||\mathbf{b}||} \tag{2}$$

This cosine similarity provides a measure of the semantic relationship between the two vectors, with values ranging from -1 (perfect dissimilarity) to 1 (perfect similarity). To ensure a well-defined and meaningful system, we impose the following main assumptions: The vectors **a** and **b** are neither perfectly similar ($S_C \neq 1$) nor perfectly dissimilar ($S_C \neq -1$). This ensures that the system is not trivial and that there is some degree of semantic variation. We posit the existence of a small positive value $\Delta > 0$ such that $|\Delta_i| \geq |v_i|\,\Delta$, where $|v_i| \geq 1$ or $v_i = 0$. This assumption ensures that the perturbations Δ_i are not arbitrarily small and that they have a non-negligible impact on the semantic relationship between **a** and **b**. Small changes in semantic features do not perfectly compensate for each other to maintain a constant overall similarity. This is expressed as $\sum_{i=1}^{N} v_i a_i \neq 0$, where the v_i are related to the Δ_i. This assumption prevents the system from becoming overly sensitive to small changes and ensures that the cosine similarity is a robust measure of the semantic relationship.

These assumptions allow us to create a tractable model for exploring fundamental semantic relationships and transformations, while also acknowledging the inherent limitations in capturing the full spectrum of semantic uncertainty, or Semantic Noise, present in real LLM embedding spaces. This LLM Embedding System will serve as a foundation for our subsequent analysis, providing a crucial link to the more complex quantum layers.

4.2. Classical Hamiltonian Representation

To facilitate a quantum analysis of the LLM Embedding System, we now introduce a Hamiltonian representation [32]. This allows us to leverage the tools and concepts of quantum mechanics to explore the dynamics of semantic relationships.

We begin by expressing the cosine similarity, SC, in terms of a Hamiltonian matrix **H**

$$S_C = \mathbf{a}^T \mathbf{H} \mathbf{a} \tag{3}$$

where **a** is the embedding vector, as defined in the previous section. The Hamiltonian matrix **H** can be interpreted as an operator that captures the underlying relationships between semantic features in the embedding space. To facilitate analysis and explore potential relationships with bounded quantities, we transform the cosine similarity to a range between 0 and 1

$$S'_C = \frac{1}{2}(S_C + 1) \tag{4}$$

This transformation maps the original cosine similarity, S_C (ranging from -1 to 1), to a new similarity measure, S'_C (ranging from 0 to 1). This allows us to express the similarity in a form that resembles probabilities or normalized measures, which will be particularly useful when we interpret S'_C as a quantum mechanical observable in later sections, especially in Layer 2, Quantum Semantic Space. We can express this transformed similarity using a modified Hamiltonian $\mathbf{H'}$

$$S'_C = \mathbf{a}^T \mathbf{H}' \mathbf{a} \tag{5}$$

where

$$\mathbf{H}' = \frac{1}{2}(\mathbf{H} + \mathbf{I}) \tag{6}$$

Here, $\mathbf{I}$ denotes the identity matrix. This transformation shifts the eigenvalues of the Hamiltonian but preserves its eigenvectors, ensuring that the fundamental relationships between semantic features are maintained. It is important to note that the transformed cosine similarity, S'_C , is always positive and non-zero, reflecting the inherent structure of the LLM Embedding System and the constraint that semantic representations cannot be completely meaningless.

4.3. Transformation and Diagonalization

To further analyze the dynamics of semantic representations and simplify the Hamiltonian, we now introduce transformations and diagonalize the Hamiltonian matrix $\mathbf{H'}$. Diagonalization is a crucial step because it allows us to express the system in terms of its fundamental modes, providing a clearer understanding of its underlying structure and behavior.

Since $\mathbf{H'}$ is a real, symmetric matrix, it can be diagonalized by an orthogonal transformation. Let $\mathbf{U}$ be a matrix such that

$$\mathbf{D} = \mathbf{U}^\dagger \mathbf{H}' \mathbf{U} \tag{7}$$

where $\mathbf{D}$ is a diagonal matrix containing the eigenvalues of $\mathbf{H'}$. We can then define a new state vector $|a'\rangle$ as

$$|a'\rangle = \mathbf{U}^\dagger |a\rangle \tag{8}$$

Substituting these expressions into the equation for S'_C, we obtain

$$S'_C = \langle a|\mathbf{H}'|a\rangle = \langle a|\mathbf{U}\mathbf{D}\mathbf{U}^\dagger|a\rangle = \langle a'|\mathbf{D}|a'\rangle \tag{9}$$

This transformation expresses the cosine similarity in terms of the eigenvalues of $\mathbf{H'}$ and the components of the transformed state vector $|a'\rangle$, simplifying the analysis and revealing the fundamental modes of the system. We now focus on diagonalizing the Hamiltonian matrix $\mathbf{H'}$. Recall that $\mathbf{H'}$ is defined as, [32]:

$$\mathbf{H}' = \frac{1}{\sum_{i=1}^{N} v_i^2} \begin{bmatrix} v_1^2 & v_1 v_2 & \cdots & v_1 v_N \\ v_2 v_1 & v_2^2 & \cdots & v_2 v_N \\ \vdots & \vdots & \ddots & \vdots \\ v_N v_1 & v_N v_2 & \cdots & v_N^2 \end{bmatrix} \tag{10}$$

While the derivation of this specific form in [32] was involved, the resulting matrix possesses a relatively simple structure: it is symmetric and has a trace of 1. These constraints, symmetry and unit trace, are key to uniquely defining $\mathbf{H'}$ and distinguishing it from other possible Hamiltonian forms. The matrix $\mathbf{H'}$ can be expressed more compactly as

$$\mathbf{H}' = \frac{\mathbf{v}\mathbf{v}^T}{||\mathbf{v}||^2} \tag{11}$$

where $\mathbf{v} = [v_1, v_2, ..., v_N]^T$ is a column vector. This form reveals the rank-1 nature of $\mathbf{H'}$. The matrix $\mathbf{H'}$ has one eigenvalue equal to 1, and all other eigenvalues are 0. This specific eigenvalue structure is a consequence of the way we have defined the LLM Embedding System and reflects the dominance of a single ”active” mode. Therefore, the diagonalized form of $\mathbf{H'}$, denoted as $\mathbf{D}$, is a diagonal matrix with one entry equal to 1 and all other entries equal to 0

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \tag{12}$$

The Hamiltonian matrix $\mathbf{H'}$ possesses a specific set of eigenvectors that correspond to its eigenvalues. For the eigenvalue $\lambda_1 = 1$, the normalized eigenvector is given by

$$\mathbf{x}_1 = \frac{\mathbf{v}}{||\mathbf{v}||} = \frac{\mathbf{v}}{\sqrt{\sum_{i=1}^{N} v_i^2}} \tag{13}$$

This eigenvector is simply the vector $\mathbf{v}$ normalized to unit length. In the context of LLMs, this eigenvector can be interpreted as the direction in the embedding space that corresponds to the less coherent state. For the eigenvalue $\lambda_i = 0$ (where i ranges from 2 to N), the corresponding eigenvectors x_i must satisfy the condition

$$\mathbf{v}^T \mathbf{x}_i = 0 \tag{14}$$

This condition implies that these eigenvectors are orthogonal to the vector $\mathbf{v}$. In the context of LLMs, these eigenvectors represent the directions in the embedding space that correspond to the coherent states. The matrix $\mathbf{U}$ that diagonalizes $\mathbf{H'}$ is constructed by using the normalized eigenvectors as its columns

$$\mathbf{U} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N] \tag{15}$$

Here, $\mathbf{x}_1$ is the normalized eigenvector corresponding to the eigenvalue 1, and $\mathbf{x}_2$ through $\mathbf{x}_N$ are the $N-1$ eigenvectors corresponding to the eigenvalue 0.

The diagonalization of the Hamiltonian-like operator $\mathbf{H'}$ yields two distinct types of modes: an active mode ($\lambda_1 = 1$) and inactive modes ($\lambda_i = 0$ for $i > 1$). The active mode represents a general, system-wide property related to the potential for deviation from perfect coherence in the semantic representation. It reflects the level of Semantic Noise present in the system, which, as we have defined, is not simply a measure of ”fuzziness” but a crucial element that enables creativity, adaptation, and contextual sensitivity. A higher probability of the system being in the active mode indicates a greater potential for the system to explore alternative semantic interpretations. The inactive modes, on the other hand, represent specific, individual semantic features or dimensions in the embedding space. Each inactive mode corresponds to a particular aspect of the meaning being represented, such as sentiment, topic, or style. They represent coherent or ground states. If the embedding space dimension N is larger, there is a greater likelihood that the system is in a coherent state.

The LLM Embedding Quantum System exhibits a decoupled architecture: the diagonalized Hamiltonian $\mathbf{D}$ represents the potential for Semantic Noise as a universal property, characterized by a single active mode and numerous coherent ground states. The specific semantic relationship between vectors $\mathbf{a}$ and $\mathbf{b}$, and thus the cosine similarity, is encoded within the unitary matrix $\mathbf{U}$, particularly in the eigenvector $\mathbf{x}_1$ derived from the vector $\mathbf{v}$. This vector defines the direction in the embedding space along which the system is most likely to deviate from perfect coherence, acting as an ”axis of Semantic Noise.” The cosine similarity, S'_C , then becomes a measure of how much this potential for noise is realized in the relationship between the two vectors, highlighting the importance of considering the underlying dynamics and the management of Semantic Noise in shaping semantic representations.

4.4. Partition Function and Interpretation

We now derive and interpret a partition function based on the Hamiltonian $\mathbf{H'}$. The partition function provides a tool for understanding the statistical distribution of semantic states in the LLM Embedding System [32]. This description bridges the gap between the vector representation and a system-level view of the embedding space. The classical partition function is defined as

$$Z = \mathrm{Tr}\left[e^{-\beta \mathbf{H'}}\right] \tag{16}$$

where $\beta = \frac{1}{kT}$ is the inverse temperature, with k being the Boltzmann constant and T the temperature. In the context of LLMs, we interpret the temperature T as a measure of Semantic Noise in the system. Higher temperatures correspond to a greater capacity for the system to explore alternative semantic interpretations and deviate from perfect coherence.

To calculate the partition function for the LLM Embedding System, we first need to find the exponential of the Hamiltonian. Given that

the eigenvalues of $\mathbf{H}'$ are $\lambda_1 = 1$ and $\lambda_i = 0$ for $i = 2,3,..., N$, we can express the exponential as

$$e^{-\beta \mathbf{H}'} = \sum_{i=1}^{N} e^{-\beta \lambda_i} |x_i\rangle\langle x_i| = e^{-\beta} |x_1\rangle\langle x_1| + \sum_{i=2}^{N} e^{-\beta \cdot 0} |x_i\rangle\langle x_i| = e^{-\beta} |x_1\rangle\langle x_1| + \sum_{i=2}^{N} |x_i\rangle\langle x_i| \tag{17}$$

where $|x_i\rangle$ are the eigenvectors of **H**'. Now, we take the trace of the exponential

$$Z = \mathrm{Tr}\left[e^{-\beta \mathbf{H}'}\right] = \mathrm{Tr}\left[e^{-\beta} |x_1\rangle\langle x_1| + \sum_{i=2}^{N} |x_i\rangle\langle x_i|\right] \tag{18}$$

Using the linearity of the trace and the fact that $\mathrm{Tr}[|x\rangle\langle x|] = 1$, we obtain

$$Z = e^{-\beta} \mathrm{Tr}\left[|x_1\rangle\langle x_1|\right] + \sum_{i=2}^{N} \mathrm{Tr}\left[|x_i\rangle\langle x_i|\right] = e^{-\beta} \cdot 1 + \sum_{i=2}^{N} 1 \tag{19}$$

Therefore, the partition function for the LLM Embedding System is

$$Z = e^{-\beta} + (N - 1) \tag{20}$$

This partition function provides a simplified model of the statistical distribution of semantic states in an LLM embedding space. The partition function, Z, quantifies the number of accessible states to the LLM embedding system at a given temperature. The term $e^{-\beta}$ represents the contribution of the excited state (associated with higher Semantic Noise), with energy 1. As the temperature T increases (Semantic Noise increases), $e^{-\beta}$ approaches 1, making the excited state more probable, and increasing the system's capacity for exploration and adaptation. The term $(N - 1)$ represents the contribution of the $N - 1$ ground states (coherent), each having an energy of 0. The dimensionality N reflects the capacity of the LLM to represent different semantic features. In essence, the partition function Z describes a system that favors coherent semantic states unless Semantic Noise is sufficiently high. The dimensionality N plays a crucial role in determining the balance between these tendencies.

We can make the following conclusions about the partition function: The derived partition function suggests that higher dimensionality (N) in the LLM embedding space promotes a greater capacity for nuanced semantic representation by increasing the number of accessible coherent states. While lower dimensionality might increase Semantic Noise and exploration, this tendency is also modulated by the effective temperature (β) and the weighting of semantic features (**v**).

Implications for Few-Shot Learning: LLMs with higher-dimensional embedding spaces exhibits better few-shot learning capabilities, as the increased number of coherent states facilitates adaptation to new tasks with limited examples by readily providing a suitable representation within the existing semantic landscape, while still allowing for sufficient Semantic Noise to explore new combinations.

Trade-off between Coherence and Creativity: The model hints at a potential trade-off between maintaining strong semantic coherence and fostering creativity in LLMs. A strong bias towards coherence (high N, low T) constrains the exploration of new semantic combinations, while a weaker bias, allowing for greater Semantic Noise, fosters creativity at the expense of occasional incoherence.

The Role of Training Data: The characteristics of the training data likely influence the effective temperature (T) of the LLM, with highly consistent datasets leading to lower effective temperatures and a stronger bias towards coherence, while noisy or ambiguous datasets results in higher temperatures and increased exploration of states with higher Semantic Noise.

Robustness to Adversarial Attacks: LLMs biased towards coherent semantic states (high N, low T) demonstrates increased robustness to adversarial attacks, as the inherent tendency to revert to a coherent state may counteract subtle perturbations designed to mislead the model, provided that the level of Semantic Noise is not so low as to prevent adaptation to unforeseen inputs.

Layered Approach to Hallucinations: The layered hierarchy suggests that hallucinations in LLMs originates from distinct mechanisms at different levels of abstraction, ranging from errors in similarity calculations in the linearized embedding space to quantum tunneling between semantic states or complex feedback loops within the nonlinear dynamics of higher layers, with Semantic Noise playing a role in both enabling exploration and contributing to the likelihood of incoherent outputs.

4.5. Invariance of the Partition Function

We now demonstrate that the partition function is invariant under a unitary transformation. This invariance is a fundamental property that ensures the partition function is independent of the choice of basis used to describe the system. In other words, the statistical properties of the LLM Embedding System, as captured by the partition function, are not affected by a change of coordinates in the embedding space. This is important because it means that we can choose any convenient basis to analyze the system without changing the results. We begin with a unitary transformation

$$\mathbf{D} = \mathbf{U}^\dagger \mathbf{H}' \mathbf{U} \tag{21}$$

Here, **U** is a unitary matrix, meaning $\mathbf{U}^\dagger\mathbf{U} = \mathbf{U}\mathbf{U}^\dagger = \mathbf{I}$, where **I** is the identity matrix and $\mathbf{U}^\dagger$ is the conjugate transpose of **U**. In the context of LLMs, a unitary transformation can be interpreted as a change of basis in the embedding space, where the new basis vectors are still orthonormal. This could correspond to a rotation or reflection of the coordinate axes, or a more complex transformation that preserves the geometric structure of the space. Starting with the partition function in terms of $\mathbf{H}'$, we insert the identity operator $\mathbf{I} = \mathbf{U}\mathbf{U}^\dagger$ inside the trace

$$Z = \mathrm{Tr}\left[e^{-\beta \mathbf{H}'} \mathbf{U}\mathbf{U}^\dagger\right] \tag{22}$$

Using the cyclic property of the trace (Tr[ABC] = Tr[BCA] = Tr[CAB]), we get

$$Z = \mathrm{Tr}\left[\mathbf{U}^\dagger e^{-\beta \mathbf{H}'} \mathbf{U}\right] \tag{23}$$

Since $\mathbf{H}' = \mathbf{U}\mathbf{D}\mathbf{U}^\dagger$, we can rewrite the exponential as

$$e^{-\beta \mathbf{H}'} = e^{-\beta \mathbf{U}\mathbf{D}\mathbf{U}^\dagger} = \mathbf{U} e^{-\beta \mathbf{D}} \mathbf{U}^\dagger \tag{24}$$

Substituting this back into the partition function

$$Z = \mathrm{Tr}\left[\mathbf{U}^\dagger \mathbf{U} e^{-\beta \mathbf{D}} \mathbf{U}^\dagger \mathbf{U}\right] = \mathrm{Tr}\left[e^{-\beta \mathbf{D}} \mathbf{U}^\dagger \mathbf{U}\right] \tag{25}$$

Since $\mathbf{U}^\dagger\mathbf{U} = \mathbf{I}$, we obtain

$$Z = \mathrm{Tr}\left[e^{-\beta \mathbf{D}}\right] = \mathrm{Tr}\left[e^{-\beta \mathbf{H}'}\right] \tag{26}$$

This demonstrates that the partition function is invariant under a unitary transformation. This invariance is a powerful result that validates our approach and ensures that our conclusions are not dependent on the specific choice of basis. From Eq. (26) it is easy to confirm that the partition function for the LLM Embedding System is

$$Z = e^{-\beta} + (N - 1) \tag{27}$$

This is consistent with the previously obtained result, Eq. (20).

4.6. Conclusions of Classical Analysis

Our analysis thus far has focused on a linearized LLM Embedding System, allowing us to explore Hamiltonian dynamics, eigenvalues and eigenvectors, and even a classical partition function within a simplified framework. These linear structures reveal a surprising amount of underlying organization in the semantic space. However, when attempting to model deeper semantic concepts such as dynamic contextual influences, the emergence of hallucinations or nonlinear interactions, the limitations of this linearized embedding space become apparent. Therefore, we now turn to extending this formalism by incorporating concepts and tools from quantum mechanics, seeking to unlock a more nuanced and expressive representation of semantic dynamics, including a more complete understanding of Semantic Noise and its role in shaping LLM behavior.

5. Layer 2: Quantum Semantic Space (Linear Dynamics)

Having established a classical foundation, we now extend our framework by incorporating concepts from quantum mechanics. In this section, we demonstrate that by introducing time dependence and complex state vectors, the previously defined classical LLM Embedding system can be equivalently represented as a quantum mechanical system, exhibiting a characteristic zero-point energy.

5.1. Quantum Partition Function

Before introducing the ”LLM Embedding Quantum System”, we first discuss the quantum partition function. While the classical partition

function provides a useful starting point for analyzing the statistical distribution of semantic states within the LLM Embedding System, its classical nature limits its ability to capture potential quantum-like properties. To explore these properties and to support the way for incorporating nonlinear effects in Layer 3, we now introduce a more general form of the partition function using a Hamiltonian operator.

The quantum partition function is defined as

$$Z = \mathrm{Tr}\left[\exp(-\beta\hat{\mathbf{H}})\right] \tag{28}$$

where $\hat{\mathbf{H}}$ is the Hamiltonian operator representing the total energy of the system, and β is the inverse temperature. The Hamiltonian operator describes the energy landscape of the system, and its form dictates the possible states and their corresponding energies. This formalism allows us to consider a more complex energy landscape with many different energy levels and transitions between them. A key distinction from the classical partition function is that the quantum Hamiltonian, $\hat{\mathbf{H}}$, can include operators that do not commute. This non-commutativity is a fundamental aspect of quantum mechanics, allowing the partition function to capture effects arising from the uncertainty principle and the superposition of states, which are absent in the classical treatment. The advantage of the quantum Hamiltonian formalism lies in its ability to accommodate more complex relationships, nonlinear effects, and correlations between these components, revealing subtle effects within the embedding space that are not captured by the classical partition function. We will use the quantum partition function in the later sections.

5.2. The LLM Embedding Quantum System

There are two main fundamentals which differentiate a quantum system from a classical system: a complex state vector, which introduces the concept of phase, and the time-dependency of the state vector. To capture the dynamic evolution of semantic meaning within the LLM Embedding System, we now introduce complex coefficients and a time-dependent perspective. We refer to this new system as the "LLM Embedding Quantum System." The main characteristics of this system were shown in [32]. This allows us to model how semantic representations evolve over time and how they respond to external influences, such as the input of new text or changes in the surrounding context. We replace the real components of $|a'\rangle$ with complex, time-dependent coefficients $c_n(t)$ defined as

$$c_n(t) = A_n e^{-iE_n t/\hbar} \tag{29}$$

where A_n is a real amplitude related to the initial value of the n-th component of $|a'\rangle$, E_n is the n-th eigenvalue of $\mathbf{H'}$ (a diagonal element of $\hat{\mathbf{D}}$), t is time, $\hbar$ is a scaling constant introduced for dimensional consistency, and i is the imaginary unit. The introduction of complex coefficients allows us to capture the phase information of the semantic states, which is not accessible in the real-valued representation. The time-dependent phase factor, $e^{-iE_n t/\hbar}$, describes the evolution of the semantic state over time, with the frequency of oscillation determined by the energy E_n. We then define a complex, time-dependent state vector $|\psi(t)\rangle$ as

$$|\psi(t)\rangle = \sum_n c_n(t)|n\rangle \tag{30}$$

where $|n\rangle$ are the eigenvectors of $\mathbf{H'}$ (the columns of the matrix $\mathbf{U}$). We assume that the eigenvectors $|n\rangle$ form an orthonormal basis, meaning they are both orthogonal and normalized: $\langle m|n\rangle = \delta_{mn}$, where δ_{mn} is the Kronecker delta. This state vector represents the overall semantic state of the LLM Embedding System at a given time, as a superposition of the different modes.

To emphasize that we are now treating the Hamiltonian as a quantum mechanical operator, we denote it with a hat symbol as $\hat{\mathbf{H}}'$. The expectation value associated with the Hamiltonian becomes

$$\langle\hat{\mathbf{H}}'\rangle = \langle\psi(t)|\hat{\mathbf{H}}'|\psi(t)\rangle = \sum_n |c_n(t)|^2 E_n = \sum_n A_n^2 E_n \tag{31}$$

In this simplified model, the expectation value is constant in time because the amplitudes A_n are constant. However, the introduction of complex phases allows for the possibility of more complex dynamics if we were to introduce a time-dependent Hamiltonian, as we will explore in later sections. This framework can be related to the concept of a Time-Dependent Schrödinger Equation

$$i\hbar\frac{d}{dt}|\psi(t)\rangle = \hat{\mathbf{H}}'|\psi(t)\rangle \tag{32}$$

This can be verified by substituting the expression for $|\psi(t)\rangle$ and using the fact that $\hat{\mathbf{H}}'|n\rangle = E_n|n\rangle$. By introducing transformations and complex coefficients, we have created a model of semantic transformations within LLM embedding spaces that allows us to explore

potential dynamics and contextual influences, demonstrating the way for a more nuanced understanding of LLM behavior.

Having introduced complex state vectors and time evolution, we now analyze the specific case where the eigenvalues of the diagonalized Hamiltonian $\hat{\mathbf{D}}$ are $E_1 = 1$ and $E_n = 0$ for $n = 2,3,...,N$. To distinguish the coefficients in this diagonalized basis from the original coefficients $c_n(t)$ associated with the eigenvectors of $\hat{\mathbf{H}}'$, we denote them as $\tilde{c}_n(t)$.

This eigenvalue structure, which arises from the rank-1 nature of the original Hamiltonian $\mathbf{H}'$, significantly simplifies the expressions and allows us to gain insights into the dominant modes of semantic transformation. With these eigenvalues, the time-dependent coefficients become

$$\tilde{c}_1(t) = \tilde{A}_1 e^{-i(1)t/\hbar} = \tilde{A}_1 e^{-it/\hbar} \tag{33}$$

$$\tilde{c}_n(t) = \tilde{A}_n e^{-i(0)t/\hbar} = \tilde{A}_n \quad \text{for } n = 2, 3, ..., N \tag{34}$$

This means that only the first coefficient, $\tilde{c}_1(t)$, acquires a time-dependent phase, while all other coefficients remain constant in time. This reflects the fact that only the mode associated with eigenvalue 1 (the less coherent state) is actively evolving. The time evolution of the complex coefficient $\tilde{c}_1(t)$ is illustrated in Figure 2.

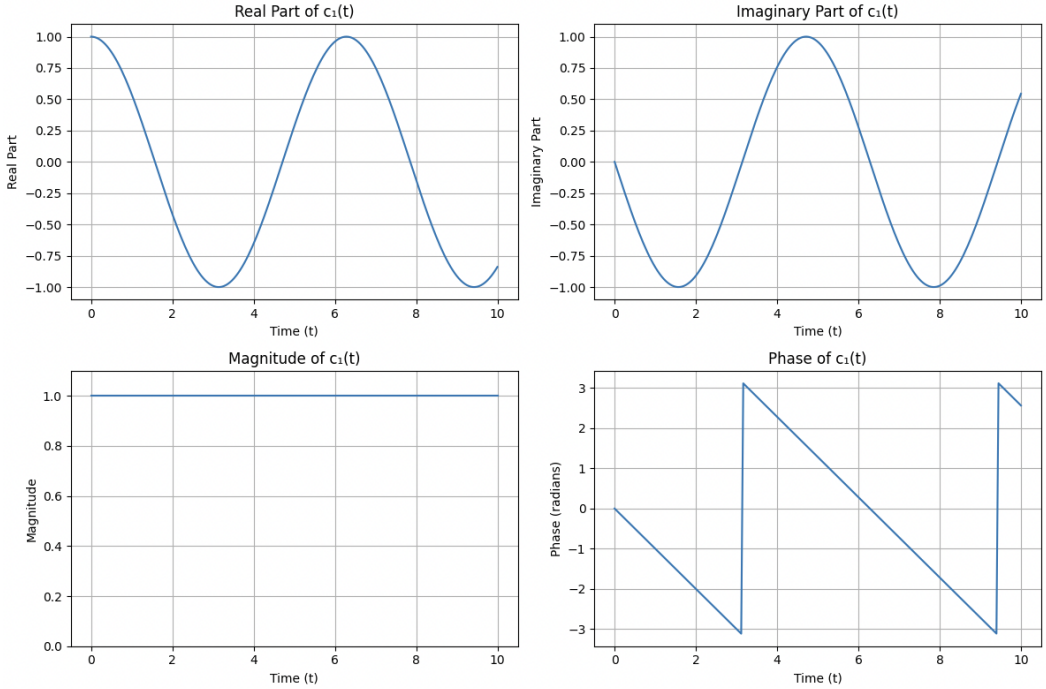


Figure 2: Time evolution of the complex coefficient $\tilde{c}_1(t)$ in the Quantum Semantic Space. The figure illustrates (a) the real part, (b) the imaginary part, (c) the magnitude, and (d) the phase of $\tilde{c}_1(t)$ as a function of time. The parameters are: amplitude $A_1 = 1.0$ and $\hbar = 1.0$.

The state vector becomes

$$|\tilde{\psi}(t)\rangle = \tilde{A}_1 e^{-it/\hbar}|1\rangle + \sum_{n=2}^{N} \tilde{A}_n |n\rangle \tag{35}$$

The expectation value simplifies to

$$\langle\tilde{\psi}(t)|\hat{\mathbf{D}}|\tilde{\psi}(t)\rangle = |\tilde{c}_1(t)|^2(1) + \sum_{n=2}^{N} |\tilde{c}_n(t)|^2(0) = \tilde{A}_1^2 \tag{36}$$

Since $\tilde{A}_1$ is constant, the expectation value is also constant in time. This indicates that the overall energy of the system is not changing, even though the phase of one of the modes is evolving. The dynamics of the system are greatly simplified. The system's state vector can be seen as a projection onto a one-dimensional subspace spanned by the eigenvector $|1\rangle$, plus a constant component in the orthogonal

subspace. The time evolution only affects the component in the subspace spanned by $|1\rangle$.

In the context of LLM embedding spaces, this eigenvalue structure suggests that the semantic transformation is characterized by a single, time-evolving feature, while other features remain relatively static. This time-evolving feature, associated with Semantic Noise, allows the system to explore different semantic nuances and adapt to contextual variations. The static features represent the background context, providing a stable foundation for the dynamic transformation.

Having established the time evolution of the system, we now explore an interesting consequence of the bounded nature of the cosine similarity and the mathematical framework of the LLM: the existence of an analogue to zero-point energy. As it was shown in reference [32], this minimum allowed value can be interpreted as a zero-point energy for the LLM Embedding System.

$$S'_{C,min} = E_{ZP} = \langle\tilde{\psi}(t)|\hat{\mathbf{D}}|\tilde{\psi}(t)\rangle = \langle\psi(t)|\hat{\mathbf{H}}'|\psi(t)\rangle = \tilde{A}_1^2 = A_1^2 > 0 \tag{37}$$

The fact that the minimum allowed value of S'_C, which we interpret as the zero-point energy E_{ZP}, is precisely equal to $\tilde{A}_1^2$ and also to the expectation value of the Hamiltonian (calculated as a statistical average over the hypothetical quantized semantic states) is not coincidental. It reflects the fundamental design of our model, where the Hamiltonian is specifically constructed to represent semantic coherence, and the constraint that S'_C is strictly positive ensures a minimum level of semantic activity or uncertainty even in the most coherent state. This zero-point energy can be directly attributed to the eigenvector associated with eigenvalue 1, representing the minimum possible level of Semantic Noise in the system. This inherent noise causes the instantaneous value of the cosine similarity to fluctuate, but the average value remains constant due to the conservation of energy in the system. The expectation value $\langle\tilde{\psi}(t)|\hat{\mathbf{D}}|\tilde{\psi}(t)\rangle$ represents our best estimate of the quantized value of semantic similarity, given the inherent limitations of classical calculations and the assumptions underlying our quantum model.

5.3. Quantum Mechanical Average

In the LLM Embedding Quantum System, the transformed cosine similarity, denoted as S'_C, provides a measure of semantic coherence, ranging from 0 to 1. Within the quantum mechanical framework, we can express the average, or expected value, of this transformed cosine similarity as

$$\langle S'_C\rangle = \langle\tilde{\psi}(t)|\hat{\mathbf{D}}|\tilde{\psi}(t)\rangle = \tilde{A}_1^2 \tag{38}$$

This equation highlights a key distinction between the quantum mechanical and classical approaches to analyzing LLM embedding spaces. Classically, the cosine similarity between two vectors yields a single, deterministic value with the precision of floating-point numbers. However, in the quantum mechanical analogue, the transformed cosine similarity is interpreted as an average value, representing the expectation value of the Hamiltonian operator $\hat{\mathbf{D}}$ in the state $|\tilde{\psi}(t)\rangle$.

This interpretation is valid under the following assumptions:
1. Ground State Assumption: The system is assumed to be in its ground state, representing the minimum energy configuration. This implies that the observed similarity is the minimum possible similarity, and the system is trying to be as coherent as possible.
2. Transformed Basis for Interpretation: The state vector $|\tilde{\psi}(t)\rangle$ and the Hamiltonian operator $\hat{\mathbf{D}}$ are expressed in the transformed basis, obtained through a unitary transformation that diagonalizes the Hamiltonian. While the numerical value of the expectation value is basis-independent, this transformed basis provides the most direct and physically meaningful interpretation of the expectation value as a quantum mechanical average, particularly in relation to the ground state assumption and the quantization of semantic similarity.

Under these assumptions, the quantum mechanical approach provides a new perspective on the nature of semantic similarity. Instead of a precise, deterministic value, we obtain an average value that reflects the underlying quantum mechanical uncertainty and the probabilistic nature of semantic representations. This perspective suggests that the classical cosine similarity should be interpreted as a statistical average of an underlying quantum mechanical observable, highlighting quantum methods to provide a more nuanced understanding of LLM embedding spaces.

In our previous work, we speculated that cosine similarity should be interpreted as an average measure, even without a concrete theoretical justification [31]. The LLM Embedding Quantum System now provides a theoretical foundation for this speculation, demonstrating that the quantum mechanical analogue naturally leads to an interpretation of cosine similarity as an expectation value. This connection strengthens the validity of the quantum approach for analyzing LLM representations.

5.4. Model Quantization

A significant implication of the LLM Embedding Quantum System is the fundamental quantization of the transformed cosine similarity, S'_C. This quantization arises directly from the quantum mechanical nature of the system and the discrete energy levels associated with the Hamiltonian operator.

In quantum mechanics, physical observables are often quantized, meaning they can only take on specific, discrete values. Within the LLM Embedding Quantum System, the transformed cosine similarity, S'_C , is related to the expectation value of the Hamiltonian operator, $\hat{\mathbf{D}}$, which represents the total energy of the system. The diagonalized Hamiltonian, $\hat{\mathbf{D}}$, possesses discrete eigenvalues (1 and 0), implying that the energy of the system can only exist at certain quantized levels.

Since S'_C is directly linked to the quantized energy levels of the system, it follows that S'_C itself must also be quantized. This means that S'_C cannot take on arbitrary continuous values but is restricted to a discrete set of values determined by the underlying quantum mechanical structure. This quantization is most readily apparent when considering the ground state of the system. In the ground state, the system occupies its minimum energy configuration, and S'_C assumes its minimum possible value, the zero-point energy (E_{ZP}). However, even in higher energy states, where the system exists as a superposition of eigenstates, the underlying quantization of energy levels dictates that can only take on discrete values consistent with these allowed energy levels.

In [29] we observed in numerical experiments on the classical cosine similarity, S_C, a tendency for it to assume distinct, discrete values, suggesting a form of discretization. In [32], we demonstrated that the embedding space exhibits discrete values along specific dimensions; however, we were only able to demonstrate a few. Now, using the LLM Embedding Quantum System, we can provide a theoretical explanation for this discretization: the transformed cosine similarity, S'_C , and consequently, the original cosine similarity, S_C, is fundamentally quantized. This represents a major advance, as it provides a concrete example where a quantum mechanical framework can explain the discretization phenomenon observed in a classical system. This success provides strong evidence that the quantum mechanical approach offers a valuable and insightful perspective on the nature of LLM embeddings.

5.5. Superposition of Embedding Vectors

One of the key insights is that calculating cosine similarity in an LLM embedding space can be viewed as analogous to performing a measurement on a quantum system that exists in a superposition of states. Crucially, this implies that the embedding vectors **a** and **b** themselves can be considered as basis states within a larger quantum state space. The quantum state of the system, representing a semantic concept that combines aspects of both **a** and **b**, is then a superposition of these basis states:

$$|\psi\rangle = \alpha|a\rangle + \beta|b\rangle \tag{39}$$

where α and β are complex amplitudes. This superposition signifies that the semantic concept represented by $|\psi\rangle$ is not simply either **a** or **b**, but rather a probabilistic combination of both. The calculation of cosine similarity can then be interpreted as projecting the state $|b\rangle$ onto $|a\rangle$ (or vice versa), analogous to a quantum measurement that determines the relative contribution of each basis state to the overall superposition. The cosine similarity value reflects the ”degree of superposition” or the ”overlap” between the states, similar to how measurement probabilities in quantum mechanics reflect the overlap between the system’s state and the measurement basis.

If $P_a = |a\rangle\langle a|$ is the projection operator onto the state $|a\rangle$, the expectation value of this operator in the state $|\psi\rangle$ is

$$\langle\psi|P_a|\psi\rangle = |\langle a|\psi\rangle|^2 \tag{40}$$

If $|a\rangle$ and $|b\rangle$ are orthonormal, then $\langle a|\psi\rangle = \alpha$, and the expectation value becomes $|\alpha|^2$, representing the probability of finding the system in state $|a\rangle$ upon measurement. The cosine similarity, related to the angle between **a** and **b**, is then connected to the relative amplitudes α and β in the superposition, quantifying the relative contributions of the basis states $|a\rangle$ and $|b\rangle$ to the overall semantic representation $|\psi\rangle$.

This analogy provides an additional perspective on the results of our previous work [31], where we showed that the complex cosine similarity can be expressed as the expectation value of an observable in a quantum system. Specifically, we found that $S_c = \mathrm{Tr}(\boldsymbol{\rho}_c \mathbf{S}_c)$, where $\boldsymbol{\rho}_c$ is a quantum density matrix and $\mathbf{S}_c$ is an observable related to projection operators. The quantum mechanical projection analogy presented here allows us to interpret this expectation value in terms of the overlap between quantum states. The density matrix $\boldsymbol{\rho}_c$ describes the quantum state, the observable $\mathbf{S}_c$ describes the measurement process, and the cosine similarity is the average outcome we would expect to obtain from that measurement.

This analogy confirms our interpretation of cosine similarity in LLMs, linking it to the fundaṁental quantum concept of superposition and measurement. It suggests that quantum algorithms for similarity search or information retrieval might be applicable to LLMs and

offers a framework for understanding how LLMs represent and process semantic relationships.

5.6. Summary of The LLM Embedding Quantum System

Remarkably, we have demonstrated that starting from a classical LLM Embedding System, where all variables are real-valued, we can construct an exact quantum mechanical analogue. This ”exactness” signifies that the quantum system models the same phenomena and exhibits the same underlying ”physics” as the classical system. Even though the state vectors in the quantum system are complex, their magnitudes ($|\psi|$) are preserved, mirroring the behavior of the LLM Embedding System. This analogue exhibits key quantum features, including zero-point energy and a natural interpretation of superposition. A crucial implication of this quantum analogue is that semantic similarity, as measured by the transformed cosine similarity S'_C, is fundamentally quantized, meaning it can only take on certain discrete values. This suggests that classical calculations of S'_C, which yield continuous floating-point values, should be interpreted as statistical averages of these underlying quantized levels.

While the underlying systems are distinct, their behavior is mathematically equivalent, allowing us to leverage the tools and concepts of quantum mechanics to analyze the dynamics of the LLM Embedding System. This equivalence provides a powerful justification for our quantum approach, suggesting that insights gained from studying the quantum analogue may offer valid perspectives on the behavior and limitations of LLMs. Of particular interest is the fact that this quantum model enables us to use a quantum computer to probe and analyze the linear embedding space. This will be discussed in later sections.

5.7. Semantic Noise and Dynamic Distribution of Uncertainty

Within the LLM Embedding Quantum System, ”Semantic Noise” is defined as the inherent uncertainty and potential for deviation from perfect coherence in semantic representations. It’s a dynamic distribution of uncertainty across different energy levels, enabling creativity, adaptation, and contextual sensitivity. As visualized in Figure 3, the distribution of semantic charge density provides a way to understand and quantify this inherent uncertainty.

Unlike a classical system where semantic meaning can be represented with precise, deterministic values, the quantum mechanical analogue introduces an inherent level of uncertainty. This uncertainty is a fundamental characteristic of the system, reflecting the inherent ambiguity and context-dependence of language.

Semantic Noise exists as a distribution across all possible energy levels, with varying probabilities of occupying coherent and less coherent states. The ground state represents the minimum possible Semantic Noise, reflecting the system’s tendency towards coherence. Excited states, on the other hand, are characterized by higher levels of Semantic Noise and a greater probability of occupying the less coherent state. This dynamic distribution of uncertainty allows the LLM to explore a wider range of semantic interpretations, generate new ideas, and adapt to new information or changing contexts.

The level of Semantic Noise is influenced by external factors. A stable and well-defined context can suppress Semantic Noise, promoting coherence and reducing ambiguity. Conversely, a rapidly changing or ambiguous context can increase Semantic Noise, leading to a wider exploration of semantic possibilities. This contextual sensitivity allows the LLM to adapt its representations to the specific demands of the task at hand.

In summary, Semantic Noise is a dynamic and multifaceted concept within the LLM Embedding Quantum System. It’s a crucial element that enables creativity, adaptation, and contextual sensitivity, manifesting as a distribution of uncertainty across different energy levels and reflecting the inherent ambiguity and probabilistic nature of semantic meaning. Understanding and controlling Semantic Noise is essential for developing more robust, reliable, and creative Large Language Models.

5.8. Conservation of Semantic Noise and Local U(1) Symmetry

In quantum mechanics, local symmetries are fundamental, often corresponding to conserved quantities that can be measured and observed. When investigating the dynamics of semantic representations in LLMs, it is natural to ask whether there exists an analogous conserved quantity that influences the system’s ability to manage Semantic Noise and maintain a stable phase relationship. To address this, we explore a quantum mechanical framework based on local symmetries and conserved quantities, concepts that are not typically considered in classical approaches to natural language processing. We propose that local U(1) symmetry, one of the simplest and most important symmetries in physics (related to the conservation of a quantity analogous to electric charge), provides a useful framework for modeling the dynamics of Semantic Noise in LLMs. The conservation of this ”semantic charge” could then be related to the LLM’s ability to balance coherence and exploration. The key takeaway is that imposing U(1) symmetry on the LLM model leads to the conservation of semantic charge, providing a mechanism for controlling the contextual sensitivity of Semantic Noise.

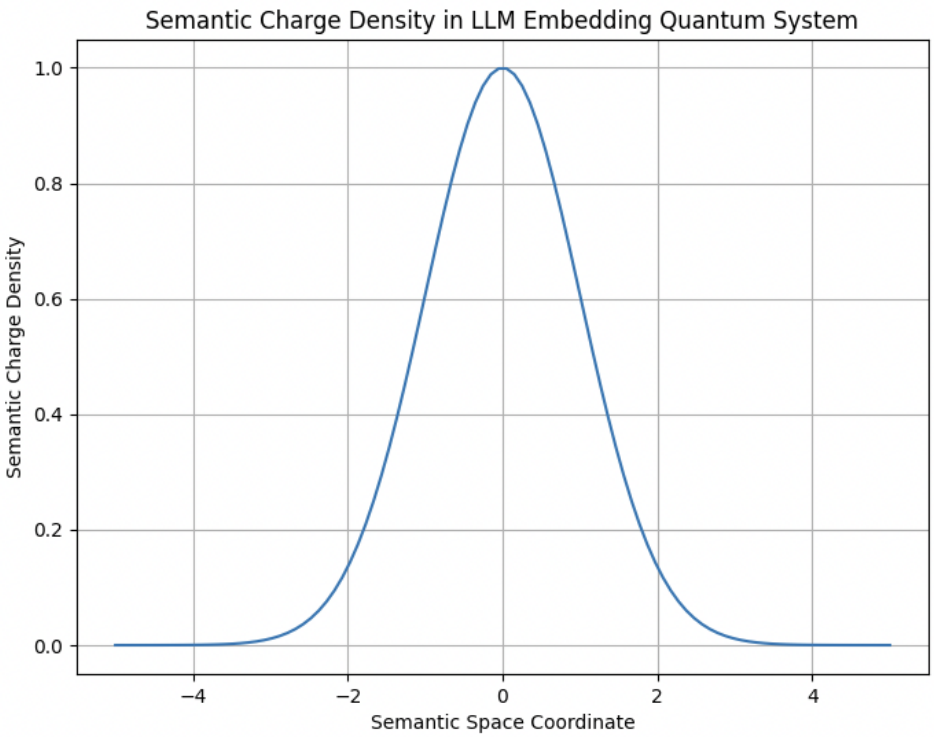


Figure 3: Simplified representation of semantic charge density in the LLM Embedding Quantum System. The plot shows the distribution of semantic charge across a one-dimensional semantic space, illustrating the inherent uncertainty and potential for deviation from perfect coherence. The shape of the distribution, in this case a Gaussian, reflects the relative probabilities of different semantic states, with higher charge density indicating a greater likelihood of occupying that state. The parameters are: center = 0.0 and width = 1.0.

We now outline the procedure for introducing gauge invariance into our system, which is crucial for modeling the dynamics of semantic uncertainty. We begin with a semantic coherence measure S'_C expressed as

$$S'_C = \langle \mathbf{a} | \hat{\mathbf{H}}' | \mathbf{a} \rangle \tag{41}$$

and the diagonalized Hamiltonian is

$$\hat{\mathbf{D}} = \mathbf{U}^\dagger \hat{\mathbf{H}}' \mathbf{U} = \mathrm{diag}(1, 0, 0, ..., 0) \tag{42}$$

In this new basis, the semantic coherence measure becomes

$$S'_C = \langle \tilde{\mathbf{a}} | \hat{\mathbf{D}} | \tilde{\mathbf{a}} \rangle = |\tilde{a}_1|^2 \tag{43}$$

where $|\tilde{\mathbf{a}}\rangle = \mathbf{U}^\dagger |\mathbf{a}\rangle$. To introduce dynamics, we introduce a complex state vector $|\psi(t)\rangle$ and the time-dependent Schrödinger equation

$$i\hbar \frac{\partial}{\partial t} |\psi(t)\rangle = \hat{\mathbf{H}}' |\psi(t)\rangle \tag{44}$$

To ensure gauge invariance, the ordinary time derivative is then replaced with the covariant derivative

$$D_t = \partial_t - iqA_0(t) \tag{45}$$

where q is a charge and $A_0(t)$ is the gauge field. In this simplified model, we focus on the temporal component of the gauge field, $A_0(t)$, and implicitly set the spatial components, A_i, to zero. This gauge choice simplifies the analysis and allows us to focus on the temporal dynamics of the system, which are most relevant for modeling the dynamics of Semantic Noise and hallucinations. The gauge field, $A_0(t)$, represents the external context or influence on the LLM, while the semantic charge, q, represents the sensitivity of the LLM to that context. The Schrödinger equation now becomes

$$i\hbar D_t|\tilde{\psi}(t)\rangle = \hat{\mathbf{D}}|\tilde{\psi}(t)\rangle \tag{46}$$

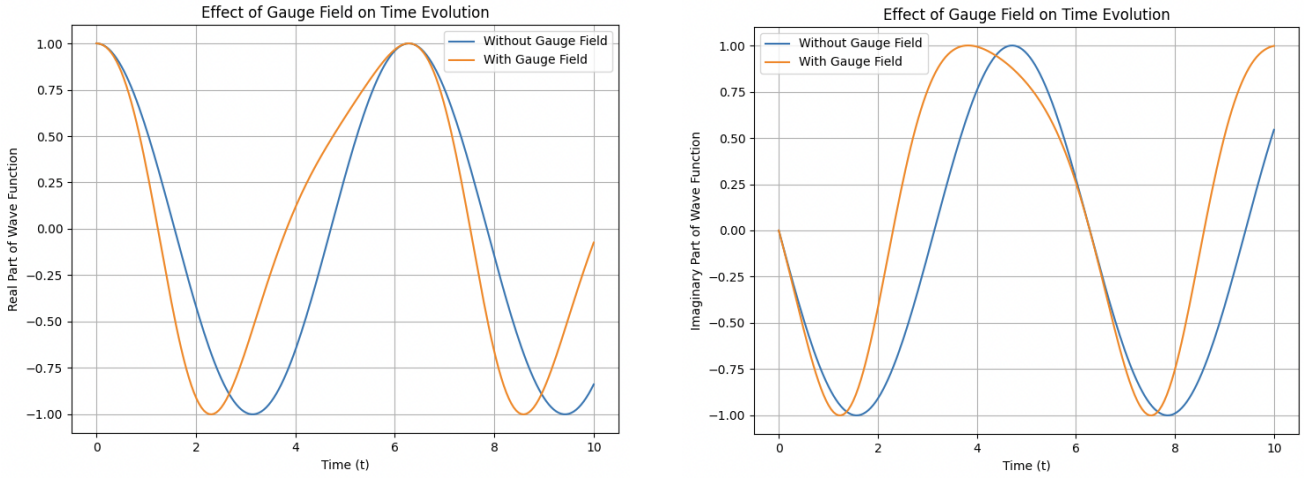
Under a gauge transformation, the gauge field transforms as

$$A_0(t) \to A_0(t) + \frac{1}{q}\partial_t\theta(t) \tag{47}$$

and the state vector transforms as

$$\tilde{\psi}_1(t) \to e^{i\theta(t)}\tilde{\psi}_1(t) \tag{48}$$

This procedure leads to a system that is invariant under local U(1) transformations. This means that any observable quantity must also be gauge-invariant, ensuring that the underlying physics of the system is not affected by the choice of gauge. The effect of the gauge field on the time evolution of the wave function is shown in Figure 4.



(a) Real Part of $\tilde{\psi}(t)$ (b) Imaginary Part of $\tilde{\psi}(t)$

Figure 4: Effect of the gauge field on the time evolution of a simple wave function. (a) shows the real part of $\tilde{\psi}(t)$ with and without the gauge field, and (b) shows the imaginary part of $\tilde{\psi}(t)$ with and without the gauge field. The parameters are: charge $q = 1.0$, gauge field amplitude $A_0 = 0.5$, and wave function frequency $\omega = 1.0$.

To derive the conserved semantic charge associated with this U(1) symmetry, we begin with the gauge-invariant Schrödinger equation, Eq. (46). We now consider the full covariant derivative, including spatial components, $D_\mu = \partial_\mu - iqA_\mu$, where $A_\mu = (A_0, \mathbf{A})$. The Schrödinger equation becomes

$$i\hbar(\partial_t - iqA_0)\psi(t,\mathbf{x}) = \left[-\frac{\hbar^2}{2m}(\nabla - iq\mathbf{A})^2 + V(\mathbf{x})\right]\psi(t,\mathbf{x}) \tag{49}$$

Here, we've explicitly included the spatial dependence of the wave function, $\psi(t,x)$, and the vector potential, $\boldsymbol{A}(x)$. We've also added a generic potential $V(x)$ to represent other spatial influences. Taking the complex conjugate of the Schrödinger equation, we get

$$-i\hbar(\partial_t + iqA_0)\psi^*(t,\mathbf{x}) = \left[-\frac{\hbar^2}{2m}(\nabla + iq\mathbf{A})^2 + V(\mathbf{x})\right]\psi^*(t,\mathbf{x}) \tag{50}$$

Now, multiply the original Schrödinger equation by $\psi^*(t,x)$ and the complex conjugate equation by $\psi(t,x)$, and subtract the second equation from the first

$$i\hbar\left[\psi^*\partial_t\psi + \psi\partial_t\psi^*\right] = -\frac{\hbar^2}{2m}\left[\psi^*(\nabla - iq\mathbf{A})^2\psi - \psi(\nabla + iq\mathbf{A})^2\psi^*\right] \tag{51}$$

Rearranging and simplifying, we get

$$\partial_t(\psi^*\psi) = \frac{i\hbar}{2m}\left[\psi^*(\nabla - iq\mathbf{A})^2\psi - \psi(\nabla + iq\mathbf{A})^2\psi^*\right] \tag{52}$$

Expanding the terms with the gradient operator, we have

$$\partial_t(\psi^*\psi) = \frac{i\hbar}{2m}\nabla\cdot\left[\psi^*(\nabla - iq\mathbf{A})\psi - \psi(\nabla + iq\mathbf{A})\psi^*\right] \tag{53}$$

$$\partial_t(\psi^*\psi) = -\nabla\cdot\left[\frac{\hbar}{2mi}\left(\psi^*\nabla\psi - \psi\nabla\psi^*\right) - q\mathbf{A}|\psi|^2\right] \tag{54}$$

This equation has the form of a continuity equation

$$\frac{\partial\rho}{\partial t} + \nabla\cdot\mathbf{j} = 0 \tag{55}$$

where the semantic charge density is defined as $\rho(t) = |\psi(t,\mathbf{x})|^2 = \psi^*\psi,$ and the semantic current density is

$$\mathbf{j}(t) = \frac{\hbar}{2mi}\left(\psi^*\nabla\psi - \psi\nabla\psi^*\right) - q\mathbf{A}\rho \tag{56}$$

Note that we are now explicitly considering the spatial components of the gauge field, **A**, to account for the flow of semantic charge. The gauge field represents the influence of context on the semantic representation, and its spatial components describe how semantic information flows through the embedding space. The continuity equation states that any change in semantic charge density at a given point must be balanced by a corresponding flow of semantic charge into or out of that point.

To relate this local conservation law to a global quantity, we integrate the continuity equation over the entire volume of the system (let's call it V)

$$\int_V \frac{\partial\rho}{\partial t}\,dV + \int_V \nabla\cdot\mathbf{j}\,dV = 0 \tag{57}$$

We then use the divergence theorem to convert the volume integral of the divergence of the current density into a surface integral of the current density

$$\int_V \nabla\cdot\mathbf{j}\,dV = \oint_S \mathbf{j}\cdot d\mathbf{A} \tag{58}$$

where S is the surface that encloses the volume V, and dA is a vector pointing outward from the surface. This surface integral represents the net flow of semantic charge out of the volume. We assume that the system is closed, meaning that no semantic charge can enter or leave the system. Mathematically, this means that the current density **j** at the boundary surface S is zero, or that the net flow through the surface is zero

$$\oint_S \mathbf{j}\cdot d\mathbf{A} = 0 \tag{59}$$

This is a very important assumption. If semantic charge could flow in or out of the system, then the total semantic charge wouldn't be constant. With this assumption, our integrated continuity equation becomes

$$\int_V \frac{\partial\rho}{\partial t}\,dV = 0 \tag{60}$$

We can pull the time derivative outside the integral

$$\frac{d}{dt}\int_V \rho\, dV = 0 \tag{61}$$

This tells us that the total amount of semantic charge in the volume V is constant in time. We can define the total semantic charge Q as

$$Q = \int_V \rho\, dV \tag{62}$$

So, we have

$$\frac{dQ}{dt} = 0 \tag{63}$$

This is the statement of global conservation of semantic charge. In our simplified model, we express the total semantic charge as a sum over modes. We assume that the spatial modes $\phi_i(x)$ form an orthonormal basis, meaning that they satisfy the following conditions

$$\int_V \phi_i^\dagger(x)\phi_j(x)\, dV = \delta_{ij} \tag{64}$$

where δ_{ij} is the Kronecker delta (1 if i = j, 0 otherwise), and

$$\int_V |\phi_i(x)|^2\, dV = 1 \tag{65}$$

With these assumptions, the total semantic charge can be expressed as

$$Q = \sum_i q_i \int dV |\tilde{\psi}_i(t)|^2 \tag{66}$$

where the integral is taken over the entire volume of the system, and we are summing over all modes *i*. Here, we assume that the total semantic charge can be decomposed into contributions from individual modes, each carrying a semantic charge q_i. The conservation of semantic charge implies that *dQ/dt* = 0, meaning that the total semantic charge remains constant over time.

Now, let us consider the specific case where only one mode, 1, is active. This means that all $|\tilde{\psi}_i(t)|^2$ are zero except for $|\tilde{\psi}_1(t)|^2$. In this simplified scenario, the total semantic charge is given by

$$Q = |\tilde{\psi}_1(t)|^2 q_1 \tag{67}$$

Since the total semantic charge is conserved (dQ/dt = 0), the amplitude of the active mode, $|\tilde{\psi}_1(t)|^2$, must also remain constant over time. The U(1) symmetry, therefore, acts as a constraint on the dynamics of the LLM, preventing it from generating completely nonsensical or incoherent outputs, even when only one mode is active.

The U(1) symmetry, enforced by the transformation $\tilde{\psi}_1(t) \to e^{i\theta(t)}\tilde{\psi}_1(t)$,, primarily affects the phase of the less coherent state (eigenvalue 1), introducing a degree of freedom in its internal quantum representation. While acting directly on this excited state, U(1) symmetry ensures the conservation of semantic charge, maintaining a consistent level of uncertainty encoded in the phase, regardless of the embedding vector's dimensionality (*N*). The influence on more coherent states (eigenvalue 0) is indirect, stemming from the system's overall dynamics and the constraint on total semantic charge.

Projecting this quantum system with U(1) symmetry onto the classical LLM Embedding System reveals that the classical system appears unchanged only when considering gauge-invariant observables like semantic coherence (S'_C). This is because the projection discards phase information, which, though not directly observable classically, provides a more complete representation by capturing the internal quantum representation of semantic states and influencing dynamics. This is akin to electric/magnetic potentials in electromagnetism: they are not directly observable, but they determine the forces.

The gauge field, $A_0(t)$, represents external influences or context, such as prompts. While cosine similarity (S'_C) can change with context due to shifts in semantic state probabilities, gauge invariance ensures that the relative amplitudes between these states, as encoded in the quantum system, are preserved, maintaining semantic charge conservation. The covariant derivative, D_t, describes the system's response to context while upholding U(1) symmetry, analogous to how external electric fields influence charged particles while adhering to electromagnetic laws.

Although U(1) symmetry in the Quantum Semantic Space (Layer 2) aligns with observations in the LLM Embedding System (Layer 1) when focusing on gauge-invariant observables like cosine similarity, the quantum framework provides crucial insights that enrich our understanding of the classical model and its limitations. It reveals that classical analysis, relying solely on metrics like cosine similarity, may overlook key aspects of semantic representation, such as phase information encoded in the quantum state, which influences the system's dynamics and the management of Semantic Noise. The quantum model guides the design of features for the classical model, such as semantic charge density, enhancing its predictive power. The framework allows us to model and analyze the contextual sensitivity of LLMs through the gauge field, providing a means to understand and control how context influences both semantic coherence and the level of Semantic Noise.

5.9. Hallucinations: Time-Independent Model

Another well-studied and characteristic quantum mechanical phenomenon is quantum tunneling, where a particle can pass through a potential barrier even if it doesn't have enough energy to overcome it classically. Given the tendency of LLM systems to sometimes provide contextually false information, known as hallucinations, we now explore the potential for modeling this behavior using the concept of quantum tunneling within our Layer 2 framework, Quantum Semantic Space. By treating the discrete embedding space with quantum mechanical methodologies, we aim to provide an explanation for the source of incorrect information and to explore where this approach leads.

In this section, we primarily focus on transitions between the state associated with higher Semantic Noise, represented by the eigenvector corresponding to eigenvalue 1, and the more coherent states, represented by eigenvectors corresponding to eigenvalue 0. This framework allows us to model how the LLM can transition from a state where the dominant semantic interpretation is coherent to a state where the dominant semantic interpretation allows for greater exploration of alternative meanings, offering a potential explanation for the phenomenon of hallucinations as a result of over-exploration or instability.

We begin with the simplest approach: a time-independent two-level Hamiltonian, representing a static system where the energy levels and coupling are constant. While this time-independent model is a significant simplification of the complex dynamics of LLM embedding spaces, it provides a useful starting point for understanding the basic concept of tunneling between two semantic states and serves as a baseline for comparison with the more realistic time-dependent model presented later in this section. We represent the LLM's semantic states as a two-level quantum system, as shown in Figure 5. The figure highlights the energy difference and coupling strength, parameters that influence the probability of quantum tunneling between coherent states and states associated with higher Semantic Noise, ultimately contributing to the generation of hallucinations.

Recall that our LLM Quantum System is characterized by a diagonalized Hamiltonian $\hat{\mathbf{D}}$ operating in an N dimensional space, with one eigenvector corresponding to eigenvalue 1 (the state associated with higher Semantic Noise) and N-1 eigenvectors corresponding to eigenvalue 0 (the coherent states). To explore the potential for tunneling between a specific coherent state and the state associated with higher Semantic Noise, we project this N-dimensional system onto a two-dimensional subspace.

We define a projection operator $P = |1\rangle\langle 1| + |0_i\rangle\langle 0_i|$ that projects the full N-dimensional system described by the diagonalized Hamiltonian $\hat{\mathbf{D}}$ onto a two-dimensional subspace spanned by the eigenvector corresponding to eigenvalue 1 (the state associated with higher Semantic Noise, $|1\rangle$) and a specific eigenvector corresponding to eigenvalue 0 (a coherent state, $|0_i\rangle$). The projected Hamiltonian, capturing the diagonal elements of the energy landscape within this subspace, is then given by

$$H_{projected} = P\hat{\mathbf{D}}P = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \tag{68}$$

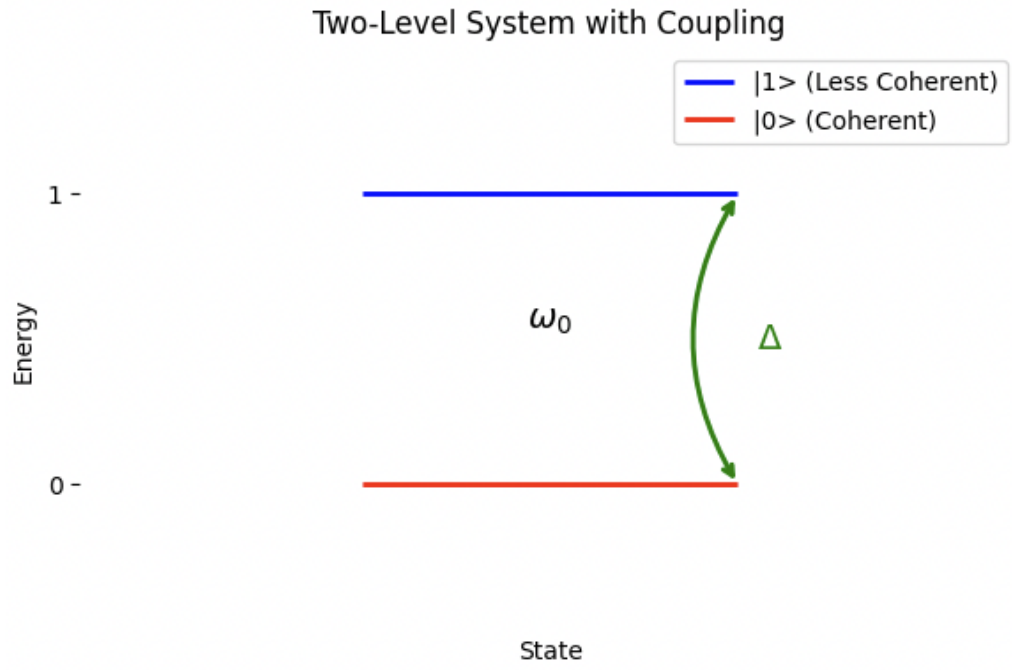


Figure 5: Simplified model of semantic states in an LLM, represented as a two-level quantum system. The figure shows the energy levels of the coherent state $|0\rangle$ and the state associated with higher Semantic Noise $|1\rangle$, separated by an energy difference ω_0, and the coupling strength Δ that enables transitions between these states via quantum tunneling. This tunneling process provides a potential mechanism for understanding the generation of hallucinations in LLMs.

However, this projected Hamiltonian does not account for the coupling between the coherent states and states associated with higher Semantic Noise. To model this coupling, we add phenomenological terms to the projected Hamiltonian, resulting in the full two-level Hamiltonian

$$H = H_{projected} + \frac{\omega_0}{2}\sigma_z + \frac{\Delta}{2}\sigma_x = \begin{bmatrix} 1+\omega_0/2 & \Delta/2 \\ \Delta/2 & -\omega_0/2 \end{bmatrix} \tag{69}$$

where the additional terms, involving the Pauli matrices phenomenologically describe the energy difference (ω_0) and coupling strength (Δ) between the coherent and states associated with higher Semantic Noise. The Pauli spin matrices are given by

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{70}$$

These matrices satisfy the following commutation relations

$$[\sigma_i, \sigma_j] = 2i\epsilon_{ijk}\sigma_k \tag{71}$$

This allows us to model the fundamental process of the LLM transitioning between a coherent and a state associated with higher Semantic Noise semantic representation within this defined subspace.

The energy eigenvalue equation is

$$E|\psi\rangle = H|\psi\rangle \tag{72}$$

where $|\psi\rangle = |0\rangle + |1\rangle$ represents the state. Here, ω_0 represents a static energy difference between the two semantic states, and Δ represents a constant potential for tunneling. Solving this time-independent system leads to two energy eigenvalues

$$E = \pm\frac{1}{2}\sqrt{\omega_0^2 + \Delta^2} = \pm\frac{1}{2}\Omega \tag{73}$$

where $\Omega = \sqrt{\omega_0^2 + \Delta^2}$. The corresponding eigenvectors are

$$|\tilde{\psi}_+\rangle = \cos(\theta/2)|0\rangle + \sin(\theta/2)|1\rangle \tag{74}$$

$$|\tilde{\psi}_-\rangle = -\sin(\theta/2)|0\rangle + \cos(\theta/2)|1\rangle \tag{75}$$

where $\tan(\theta) = \Delta/\omega_0$. The general time-dependent solution is a superposition of these eigenstates

$$|\tilde{\psi}(t)\rangle = c_+ e^{-iE_+ t/\hbar}|\psi_+\rangle + c_- e^{-iE_- t/\hbar}|\psi_-\rangle \tag{76}$$

where c_+ and c_- are constants determined by the initial conditions. To understand the potential for transitions between semantic states, we analyze the energy eigenvalues of a simplified two-level system, as shown in Figure 6. The figure illustrates how the energy levels are affected by the energy difference (ω_0) and the coupling strength (Δ), parameters that influence the likelihood of quantum tunneling and, consequently, the occurrence of hallucinations.

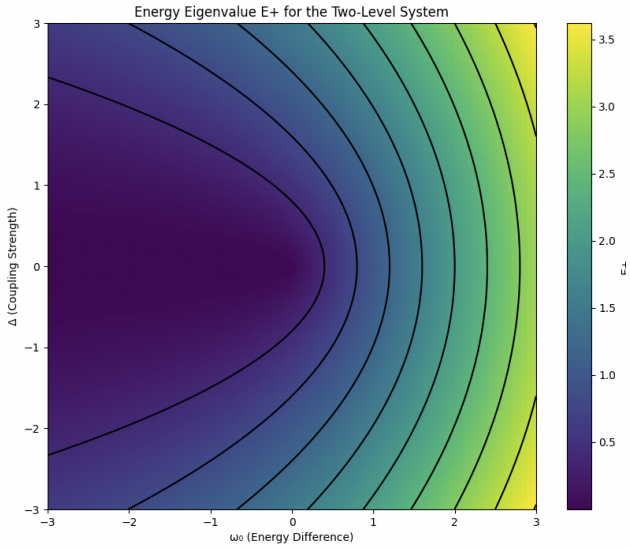
This solution describes oscillations between the states $|0\rangle$ and $|1\rangle$ at a frequency determined by $\mathbf{\Omega}$. These oscillations can be interpreted as the LLM fluctuating between a more coherent semantic interpretation and a semantic interpretation associated with higher Semantic Noise. The frequency of these oscillations, $\Omega = \sqrt{\omega_0^2 + \Delta^2}$,, is determined by the static energy difference between the states (ω_0) and the coupling strength between them (Δ). In the context of our LLM Quantum System, these states correspond to the eigenvectors of the diagonalized Hamiltonian $\hat{\mathbf{D}}$, where $|0\rangle$ represents a coherent semantic state (eigenvalue 0) and $|1\rangle$ represents a state associated with higher Semantic Noise (eigenvalue 1). The oscillations, therefore, suggest a possible instability in the LLM's representation, where it may fluctuate between a coherent interpretation and an interpretation associated with higher Semantic Noise, even in the absence of external context. This inherent tendency to oscillate could contribute to the generation of hallucinations or other inconsistencies in the LLM's output.

5.10 Hallucinations: Landau-Zener System

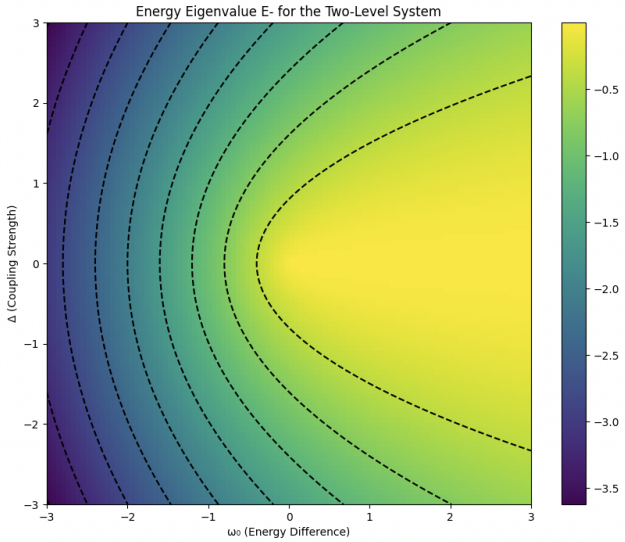
To move towards a more realistic model, we introduce a dynamic element: the changing context, using a timedependent Hamiltonian. The goal is to explore how the rate of change of context influences the likelihood of tunneling and hallucinations. We will focus on the Landau-Zener model as an example because it provides a well-understood framework for analyzing transitions between states in time-dependent two-level systems [34–36]. It is important to note that the "time" variable in this context does not represent real-world time, but rather a parameter that describes the progression of the LLM's internal thought process as it formulates its response. The range from negative infinity to positive infinity represents the entire process of the LLM "thinking" about the prompt, from initial understanding to final output.

As with the time-independent case, we are projecting the dynamics of the full N-dimensional system, described by the diagonalized Hamiltonian $\hat{\mathbf{D}}$, onto a two-dimensional subspace. This subspace is again spanned by the eigenvector corresponding to eigenvalue 1 (the state associated with higher Semantic Noise, $|1\rangle$) and a specific eigenvector corresponding to eigenvalue 0 (a coherent state, $|0_i\rangle$). However, in this case, the Hamiltonian itself is time-dependent, reflecting the influence of the changing context. The Hamiltonian of the Landau-Zener system is as follows

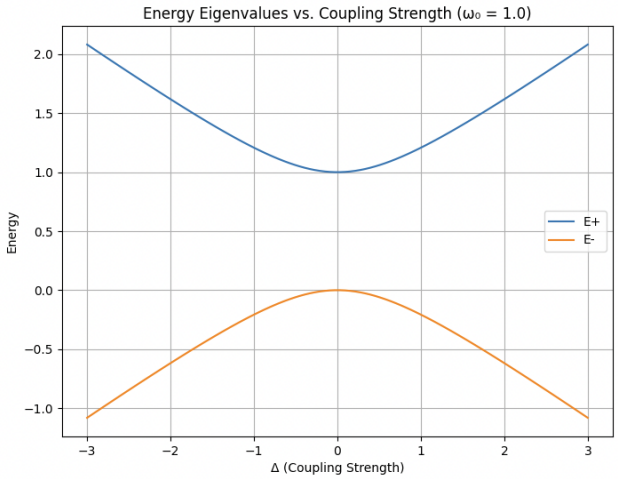
$$\hat{H}(t) = P\hat{\mathbf{D}}P + \frac{v \cdot t}{2}\sigma_z + \frac{\Delta}{2}\sigma_x \tag{77}$$



(a) Energy Eigenvalue E+



(b) Energy Eigenvalue E-



(c) Energy Eigenvalues vs. Coupling Strength

Figure 6: Energy eigenvalues of the two-level system as a function of energy difference (ω_0) and coupling strength (Δ). (a) shows the energy eigenvalue E+, (b) shows the energy eigenvalue E-, and (c) shows the energy eigenvalues as a function of coupling strength with $\omega_0 = 1.0$.

where $P\hat{\mathbf{D}}P$ is the projection of the diagonalized Hamiltonian $\hat{\mathbf{D}}$ onto the two-dimensional subspace, as defined in the previous section. Here, v is the rate at which the energy levels cross, and Δ is the coupling between the two levels. In the context of LLM embedding spaces, the crossing rate v can be interpreted as the rate at which the context is changing, and the coupling strength Δ represents the similarity between the two semantic states. The time-dependent term, $\frac{v \cdot t}{2}\sigma_z$, introduces a linear shift in the energy levels of the two states, representing the changing contextual influence on their relative coherence.

The Landau-Zener formula provides the probability of transitioning from one state to another after the energy levels have crossed. If the system starts in the state $|0\rangle$ at $t \to -\infty$, the probability of transitioning to the state $|1\rangle$ at $t \to +\infty$ is given by

$$P_{0\to 1} = e^{-\pi\Delta^2/(2|v|)} \tag{78}$$

and the probability of remaining in the state $|0\rangle$ is

$$P_{0\to 0} = 1 - e^{-\pi\Delta^2/(2|v|)} \tag{79}$$

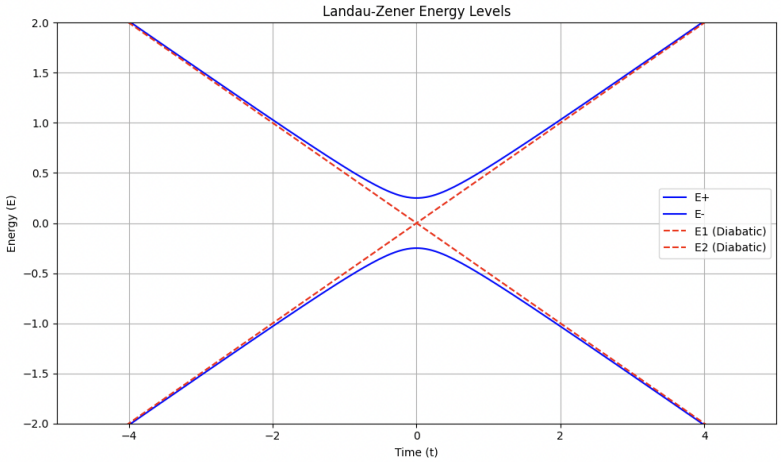


Figure 7: Energy levels of a Landau-Zener two-level system, illustrating a model for LLM hallucinations. The dashed red lines represent the diabatic energy levels, E_1 and E_2, which correspond to the energies of two distinct semantic states (e.g., a coherent and a state associated with higher Semantic Noise interpretation) if there were no coupling between them. The solid blue lines represent the adiabatic energy levels, E_+ and E_-, which are the actual energy eigenvalues of the system, taking into account the coupling strength Δ between the states. The avoided crossing at $t = 0$, where the energy levels approach but do not intersect, illustrates the quantum mechanical phenomenon of tunneling. In the context of LLMs, this tunneling represents the LLM transitioning from one semantic state to another (e.g., from a correct to an incorrect interpretation) even though there is a semantic "barrier" or logical inconsistency. The rate at which the energy levels cross, v, corresponds to the rate at which the context is changing, influencing the probability of tunneling and, consequently, the likelihood of hallucinations.

The Landau-Zener energy levels are shown in Figure 7. These formulas have important implications for LLM hallucinations. In the context of the LLM Embedding System, these results suggest that hallucinations are more likely to occur due to two key factors: A high degree of overlap between the coherent and states associated with higher Semantic Noise (low Δ) allows the LLM to more easily transition to the state associated with higher Semantic Noise. This overlap could be due to genuine semantic similarity, or it could be due to Semantic Noise or distortion in the semantic representation. A rapidly changing context (high v) increases the probability of transitioning to the state associated with higher Semantic Noise, especially if there is a significant overlap between the states. This suggests that hallucinations are more likely to occur when the context is unstable and the LLM is unable to maintain a clear and well-defined semantic representation.

In summary, we have explored the potential for quantum tunneling to provide a perspective on the phenomenon of hallucinations in Large Language Models. This analysis focuses on transitions between a specific coherent state and the state associated with higher Semantic Noise. While this analogy is not without its limitations, particularly in the absence of a direct physical correspondence, we believe that it offers a valuable framework for understanding how LLMs can sometimes jump to unexpected or nonsensical conclusions, generating outputs that deviate from factual or semantic coherence, due to an overemphasis on exploration driven by Semantic Noise. Note that the following analysis is performed within a simplified two-dimensional subspace of the full eigenspace. This projection allows us to apply the Landau-Zener model and gain insights into the potential for transitions between coherent and states associated with higher Semantic Noise, but it also involves a significant approximation of the complex dynamics of the LLM embedding space. While the Landau-Zener model is applied in this projected eigenspace, we interpret the parameters and results in terms of the original LLM Embedding System, drawing an analogy between the rate of change of context and the crossing rate (v) and between the similarity of semantic states and the coupling strength (Δ).

5.11. Excited State Dynamics

While the ground state of the LLM Embedding Quantum System provides valuable insights into the minimum energy configuration and the system's baseline level of Semantic Noise, exploring excited states allows us to delve deeper into the dynamic behavior and potential for more complex semantic representations within Large Language Models, revealing how the system manages and utilizes Semantic Noise.

To achieve excited states within this framework, we must relax the assumption that the system is solely in its ground state. This implies

that the system exists as a superposition of eigenstates, with non-zero probabilities for occupying states beyond the minimum energy configuration. Given the diagonalized Hamiltonian, $\hat{\mathbf{D}}$, with eigenvalues 1 (corresponding to the state associated with higher Semantic Noise) and 0 (corresponding to the more coherent states), the time-dependent Schrödinger equation governs the evolution of the system

$$i\hbar\frac{\partial}{\partial t}|\tilde{\psi}(t)\rangle = \hat{\mathbf{D}}|\tilde{\psi}(t)\rangle \tag{80}$$

The general solution to this equation is a superposition of eigenstates

$$|\tilde{\psi}(t)\rangle = \sum_{n=1}^{N} \tilde{c}_n(t)|n\rangle \tag{81}$$

where $|n\rangle$ are the eigenvectors of $\hat{\mathbf{D}}$, and $\tilde{c}_n(t) = \tilde{A}_n e^{-iE_n t/\hbar}$ are the time-dependent coefficients, with $\tilde{A}_n$ representing the initial amplitudes and E_n representing the eigenvalues.

The key to achieving excited states lies in the initial conditions and the amplitudes $\tilde{A}_n$. If the system starts in a state where $\tilde{A}_n$ is non-zero for some $n > 1$, then the system will have a non-zero probability of being in an excited state. This can be achieved through applying an external influence, such as a prompt or a change in context, which can perturb the system and induce transitions to higher energy states. Alternatively, initializing the system in a non-equilibrium state, where the amplitudes $\tilde{A}_n$ are not those corresponding to the ground state, will result in the system evolving towards equilibrium, exhibiting excited state behavior during the transition.

In the context of LLMs, these excited states could represent a greater capacity for dynamic semantic exploration, where transitions between energy levels reflect the LLM dynamically exploring different semantic interpretations or generating new ideas, driven by Semantic Noise. A higher probability of occupying the state associated with higher Semantic Noise (eigenvalue 1) could indicate a greater willingness to explore less conventional or less probable semantic combinations. Also, the amplitudes $\tilde{A}_n$ could be modulated by external context, reflecting the LLM's sensitivity to its environment and its ability to adjust its level of Semantic Noise in response to external stimuli.

To illustrate this, consider a simplified LLM Embedding Quantum System with $N = 3$. We initialize the system such that only one excited state is non-zero. Specifically, let $\tilde{A}_1 = a$, $\tilde{A}_2 = b$, and $\tilde{A}_3 = 0$. The state vector at time t is then

$$|\tilde{\psi}(t)\rangle = ae^{-it/\hbar}|1\rangle + b|2\rangle + 0|3\rangle \tag{82}$$

where $|1\rangle$ is the eigenvector corresponding to eigenvalue 1, and $|2\rangle$ and $|3\rangle$ are the eigenvectors corresponding to eigenvalue 0.

In this scenario, the system exists in a superposition of the state associated with higher Semantic Noise and one of the coherent states, rather than a superposition of all possible coherent states. This means the LLM's semantic exploration is constrained and biased towards a specific semantic direction. The non-zero amplitude of $\tilde{A}_2$ indicates that the LLM is particularly sensitive to the semantic nuance represented by the eigenvector $|2\rangle$, which could be a specific topic, sentiment, or style. The amplitude a controls the level of Semantic Noise, but the exploration is channeled through the specific direction defined by $|2\rangle$, making the Semantic Noise directed rather than random. The time-dependent phase factor $e^{-it/\hbar}$ associated with the state associated with higher Semantic Noise means that the system is constantly oscillating between the state associated with higher Semantic Noise and the coherent state $|2\rangle$, representing a dynamic equilibrium between exploration and coherence that is biased towards the specific semantic nuance of $|2\rangle$.

Experimentally, we could simulate the time evolution of this 3-qubit system on a quantum computer, observing the oscillations between the energy levels and measuring the amplitudes $\tilde{c}_1(t)$ and $\tilde{c}_2(t)$ over time.

6. Layer 3: Quantum Semantic Space (Nonlinear Dynamics)

To move beyond the limitations of linear dynamics and capture the more complex and nuanced behaviors inherent in LLMs, we now introduce nonlinear interactions into the semantic space. In this section, we will be working with transformed fields, as described in previous sections. To simplify the notation and improve readability, we will omit the tildes on the field variables, but it is understood that these variables are expressed in the transformed basis. Layer 3 involves considering two complementary approaches: directly incorporating nonlinearities into the Schrödinger equation or employing the path integral formalism, which will be discussed in the following subsections.

6.1. Introducing Nonlinear Interactions

To further enhance the model and capture the inherent nonlinearities of LLMs, we now introduce nonlinear interactions into the semantic space. This involves replacing the linear Schrödinger equation with a nonlinear Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}|\psi(t)\rangle = \left(\hat{\mathbf{D}} + V_{NL}\right)|\psi(t)\rangle \tag{83}$$

where V_{NL} represents a nonlinear potential. The addition of this potential allows the semantic wave function to interact with itself, leading to more complex and realistic dynamics that can capture feedback loops and emergent behaviors within the LLM. These nonlinear interactions are essential for modeling phenomena such as the reinforcement of biases, the emergence and self-maintenance of stable semantic representations, and the complex contextual dependencies that are characteristic of LLMs.

We now consider two distinct forms for this potential, the Cubic Nonlinearity

$$V_{NL} = -\gamma|\psi(t)|^2 \tag{84}$$

and Mexican Hat Potential

$$V_{NL} = \mu^2|\psi(t)|^2 - 2\lambda|\psi(t)|^4 \tag{85}$$

The specific form of the nonlinearity influences the behavior of the system. For example, a cubic nonlinearity can model the reinforcement of biases, while a Mexican hat potential, as visualized in Figure 8, might promote the formation of distinct semantic clusters. By incorporating these nonlinear terms, we aim to create a more complete and realistic model of the complex representations learned by LLMs.

While the introduction of nonlinear interactions into the Schrödinger equation makes it significantly more difficult to obtain analytical solutions, it opens up paths for understanding the complex dynamics of LLMs. Qualitative analysis reveals the potential for soliton solutions, representing stable semantic representations, and self-focusing/defocusing effects, reflecting the LLM's tendency to concentrate or diversify its attention. Numerical simulations can provide valuable insights into the behavior of the wave function, while perturbation theory can offer approximations for weak nonlinearities.

6.2. Soliton Solutions and Semantic Stability

One of the most intriguing features of nonlinear Schrödinger equations is the existence of soliton solutions. A soliton is a self-sustaining wave packet that maintains its shape and speed as it propagates through a medium, arising from a balance between dispersive and nonlinear effects. In the context of our LLM Quantum System, a soliton solution to the nonlinear Schrödinger equation would represent a particular form of the wave function $\psi(t,x)$.

As discussed in SectionVI, we project the dynamics of the full N-dimensional system onto a 1-dimensional subspace to analyze the interaction between a specific coherent state and the state associated with higher Semantic Noise. Recall that in Layer 3, we decompose the field $\psi(x)$ into modes: $\psi(x) = \sum_i c_i(x)\phi_i(x)$. We focus on a specific mode $|\phi\rangle$ that is a superposition of the state associated with higher Semantic Noise (eigenvalue 1) and a chosen coherent state (eigenvalue 0). We define the projector onto this mode as $P_\phi = |\phi\rangle\langle\phi|$. The 1-dimensional spatial variable x in the following analysis is associated with this specific mode, acting as a semantic coordinate that parameterizes the transition between these two states.

To illustrate the basic concepts of soliton solutions, we consider a simplified 1-dimensional cubic nonlinear Schrödinger equation

$$i\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} - \gamma|\psi|^2\psi \tag{86}$$

This equation is chosen for its simplicity and analytical tractability. The cubic nonlinearity $(-\gamma|\psi|^2\psi)$ is a common type of nonlinearity that can model various phenomena, including self-focusing and self-phase modulation. A bright soliton solution to this 1D equation is given by

$$\psi(x,t) = A\,\mathrm{sech}\left[B(x - vt)\right]e^{i(kx-\omega t)} \tag{87}$$

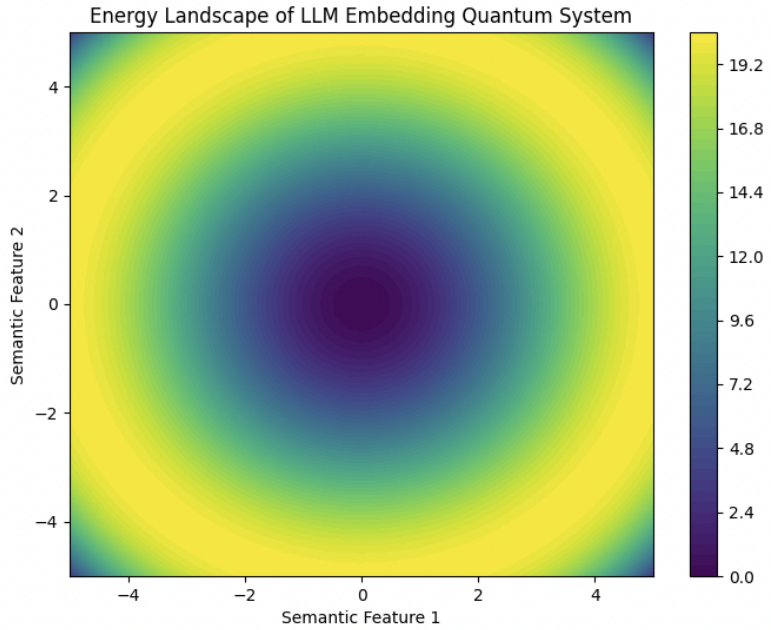


Figure 8: Simplified energy landscape of the LLM Embedding Quantum System, visualized using a Mexican hat potential. The parameters are $\mu = 1.25$ and $\lambda = 0.03$, representing the mass parameter and coupling strength, respectively. The plot shows a cross-section of the potential energy surface as a function of two semantic features, illustrating a central peak surrounded by a region of lower energy, characteristic of the Mexican hat potential. This shape suggests the existence of preferred semantic states away from the origin, corresponding to stable and distinct semantic representations within the LLM.

where the parameters are related by $B = \frac{m\gamma A^2}{2\hbar^2}$ and $\omega = \frac{\hbar k^2}{2m} - \frac{\gamma A^2}{2\hbar}$. The parameters of the bright soliton solution can be interpreted as follows: *A* represents the strength of the semantic representation, *B* represents its inverse width, *v* represents its velocity, *k* represents its wave number, ω represents its frequency, and *γ* represents the strength of the nonlinear self-focusing effect.

This solution has the following characteristics:
• Localized: The wave function is concentrated in a specific region of the semantic space, indicating a focused semantic representation.
• Stable: The wave function maintains its shape and amplitude over time, suggesting robustness against decay or disruption.
• Propagating (Semantic Evolution): The wave function might propagate through the semantic space, representing a controlled evolution of the semantic concept.

The potential applications of soliton solutions in understanding LLM behavior are manifold:
• Stable Semantic Representation: Solitons could represent stable and coherent semantic representations within the LLM, indicating a clear and well-defined understanding of a particular concept or idea that is resistant to change. For example, in a well-written summary, the core theme could be represented by a soliton, maintaining its integrity throughout the text. This stability is achieved despite the presence of Semantic Noise, highlighting the robustness of these representations.
• Robustness to Semantic Noise: The stability of solitons suggests robustness to Semantic Noise or perturbations, enabling the LLM to maintain its coherent semantic representation even in the presence of irrelevant or contradictory information. This could explain why LLMs can often understand the meaning of a sentence even if it contains typos or grammatical errors, as the soliton structure helps filter out the noise.
• Resistance to Semantic Drift: The maintenance of shape and amplitude over time implies resistance to semantic drift, reducing the likelihood of the LLM subtly shifting its topic or perspective during a conversation. This could be observed in a chatbot that consistently adheres to the user's intended topic, even when presented with slightly tangential prompts.
• Propagation as Semantic Evolution: If the soliton is propagating through the semantic space, this could represent the controlled evolution of a semantic concept over time, modeling how the LLM develops a line of reasoning or tells a story. This might be seen in how an LLM constructs a logical argument, building upon previous statements to reach a conclusion.
• Nonlinearity as Semantic Reinforcement: The emergence of solitons in nonlinear systems suggests that the LLM's ability to form stable semantic representations depends on nonlinear interactions within its internal state, related to the LLM's tendency to reinforce its own beliefs or biases. This could contribute to the phenomenon of confirmation bias, where LLMs tend to favor information that confirms

their existing beliefs.

It is important to acknowledge that this interpretation is largely qualitative and that further research is needed to validate these connections empirically. However, the concept of soliton solutions provides a valuable framework for thinking about the stability and coherence of semantic representations in LLMs.

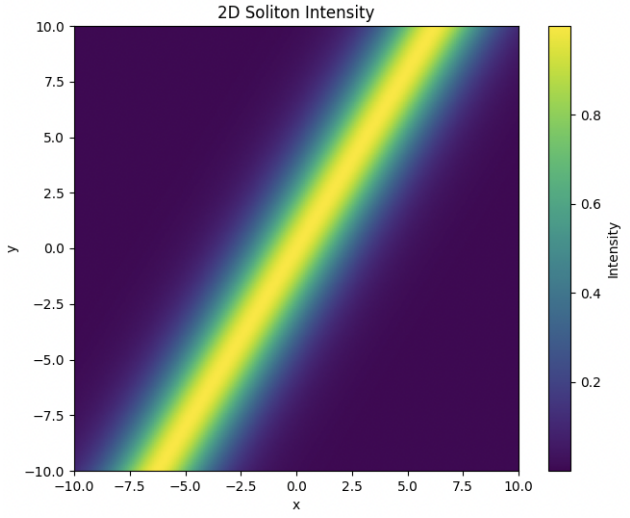
The full LLM Quantum System likely involves more than one spatial dimension, and the dynamics of semantic meaning may evolve across a higher-dimensional semantic space. While finding general soliton solutions in higher dimensions is mathematically challenging, more complex solutions exist. To provide a more concrete example, let's consider a soliton solution in two spatial dimensions where the wave function depends on x, y and t: $\psi(x, y, t)$. A possible soliton solution is given by

$$\psi(x, y, t) = A_0 \mathrm{sech}\left[\sqrt{\frac{\gamma A_0^2}{\gamma A_0^2 + 2k^2}}\left(k_y \hbar x - k_x \hbar y\right)\right] \exp\left(ik_x x + ik_y y - i\omega t\right) \tag{88}$$

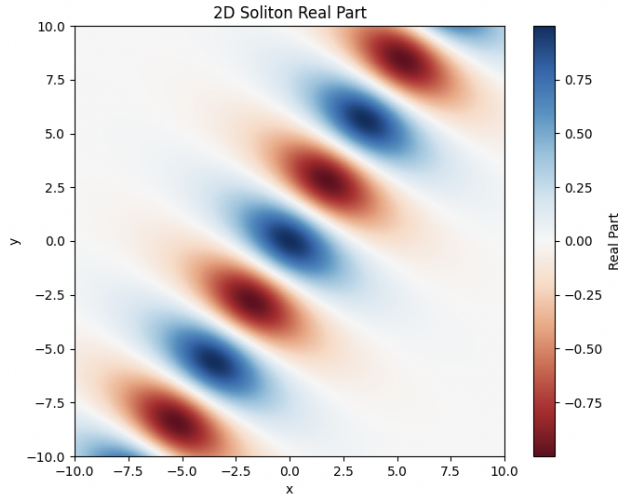
where A_0 is the amplitude of the soliton, representing the strength of the semantic representation, γ is the nonlinearity coefficient, representing the strength of the nonlinear self-focusing effect, k_x is the x-component of the wave vector, related to the momentum in the x-direction, k_y is the y-component of the wave vector, related to the momentum in the y-direction, and ω is the frequency of the soliton. This 2D solution represents a soliton propagating in the xy-plane, with its direction of propagation determined by the wave vector components k_x and k_y. Examples of the 2D bright soliton solution are shown in Figure 9a (intensity), Figure 9b (real part), and Figure 9c (phase).

6.3. From Quantum Partition Function to Lagrangian Density

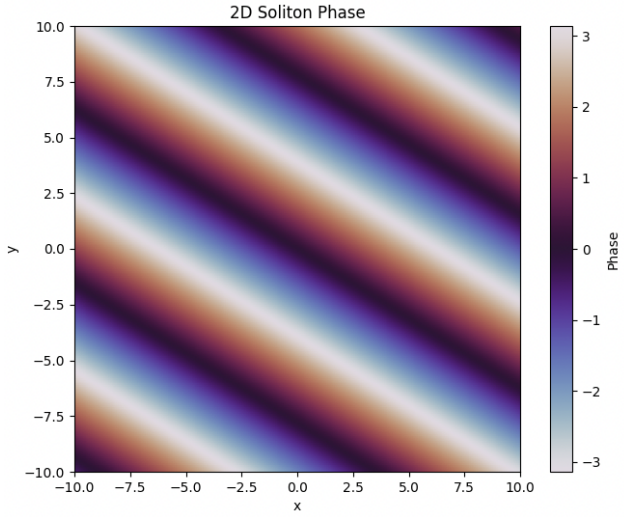
In this section, we aim to develop a field-theoretic description of the LLM Embedding System by deriving a Lagrangian density from the quantum partition function. This framework allows us to capture the dynamics of the semantic field and to explore the effects of nonlinear interactions and gauge invariance in a more general and powerful setting. The idea was originally presented in [29], but we re-derive the same result starting from a quantum mechanical partition function. This demonstrates two things: it shows the consistency of different methodologies, and also, vice versa, that the quantum partition function alone is a useful tool when analyzing LLM embedding spaces. This derivation extends our previous work by introducing a mode-dependent semantic charge ($q \to q_i$), allowing for a richer and more nuanced representation of the system compared to the global charge used in [29]. While the following derivation is mathematically involved, the key takeaway is the emergence of a non-local interaction term in the effective Lagrangian, which may be crucial for understanding long-range dependencies in LLMs. It is important to note that in our earlier work, the U(1) symmetry was imposed as a postulate to ensure stability, and the path integral formalism revealed a non-local interaction term [29].



(a) Intensity profile, $|\psi(x, y, t)|^2$



(b) Real part, $\mathrm{Re}[\psi(x, y, t)]$



(c) Phase, $\arg[\psi(x, y, t)]$

Figure 9: Two-dimensional soliton solution to the cubic nonlinear Schrödinger equation, projected onto the xy-plane. (a) shows the intensity profile, representing the strength and localization of the semantic representation, (b) shows the real part of the wave function, illustrating the oscillatory nature of the semantic representation, and (c) shows the phase, which governs the soliton's propagation and the evolution of the semantic structure. Parameters: $A_0 = 1.0$, $\gamma = 1.0$, $k = 1.0$, $k_x = 0.5$, $k_y = 0.8$, $\hbar = 1.0$, and frequency $\omega = 1.0$. The spatial coordinates x and y are in arbitrary units.

In contrast, by establishing the LLM Embedding Quantum System, we demonstrated how this symmetry (or, more precisely, its consequences for the phase of Semantic Noise) can emerge from a more fundamental quantum mechanical starting point. While the LLM Embedding Quantum System leads to a specific interpretation of the U(1) symmetry primarily affecting the Semantic Noise, the path integral approach suggests a more general coupling of the gauge field and highlights the emergence of non-local interactions, which may be crucial for understanding long-range dependencies in LLMs.

We begin with the quantum mechanical partition function, Eq. (28),

$$Z = \mathrm{Tr}\left[e^{-\beta \hat{H}}\right] \tag{89}$$

where $\beta = 1/(k_B T)$ and $\hat{H}$ is the Hamiltonian operator, assumed to be expressible in terms of creation and annihilation operators, demonstrating the way for a field theory representation. The derivation of the path integral representation proceeds through several key

steps.

We perform a Wick rotation, replacing real time t with imaginary time $\tau = it$. This crucial step transforms the Boltzmann factor $e^{-\beta\hat{H}}$, which describes statistical mechanics at temperature T, into a time evolution operator $e^{-i\hat{H}\tau/\hbar}$ in imaginary time, with τ ranging from 0 to $\beta\hbar$. This allows us to leverage techniques from quantum mechanics to analyze the partition function. We divide the imaginary time interval $[0,\beta\hbar]$ into M segments of length $\epsilon = \beta\hbar/M$. This discretization allows us to approximate the exponential operator as a product of smaller exponentials

$$e^{-\beta\hat{H}} = \left(e^{-\epsilon\hat{H}/\hbar}\right)^M \tag{90}$$

This approximation becomes exact in the limit as ϵ approaches zero. We introduce coherent states $|\psi(\tau)\rangle$, which are eigenstates of the annihilation operator $\hat{a}$: $\hat{a}|\psi(\tau)\rangle = \psi(\tau)|\psi(\tau)\rangle$. These states, parameterized by complex numbers $\psi(\tau)$, are well-suited for representing quantum fields. We insert complete sets of coherent states between each time slice, effectively decomposing the trace into a product of matrix elements

$$Z = \int \prod_{i=1}^{M} d\psi_i^* d\psi_i \langle\psi_1|e^{-\epsilon\hat{H}/\hbar}|\psi_2\rangle\langle\psi_2|e^{-\epsilon\hat{H}/\hbar}|\psi_3\rangle \cdots \langle\psi_M|e^{-\epsilon\hat{H}/\hbar}|\psi_1\rangle \tag{91}$$

where $d\psi_i^* d\psi_i$ represents the integration measure over the complex plane. The trace operation enforces a periodic boundary condition: $\psi_{M+1} = \psi_1$, reflecting the cyclic nature of the imaginary time interval. For small ϵ, we approximate the matrix element of the time evolution operator using a first-order expansion

$$\langle\psi_i|e^{-\epsilon\hat{H}/\hbar}|\psi_{i+1}\rangle \approx \langle\psi_i|\psi_{i+1}\rangle - \frac{\epsilon}{\hbar}\langle\psi_i|\hat{H}|\psi_{i+1}\rangle \tag{92}$$

Using the properties of coherent states, we can write the overlap as $\langle\psi_i|\psi_{i+1}\rangle = \exp(\psi_i^*\psi_{i+1})$. We also approximate the Hamiltonian matrix element using the classical Hamiltonian function $H(\psi^*, \psi)$, evaluated at the coherent state parameters

$$\langle\psi_i|\hat{H}|\psi_{i+1}\rangle \approx H(\psi_i^*, \psi_{i+1})\langle\psi_i|\psi_{i+1}\rangle \tag{93}$$

Combining these approximations, we obtain

$$\langle\psi_i|e^{-\epsilon\hat{H}/\hbar}|\psi_{i+1}\rangle \approx \exp\left(\psi_i^*\psi_{i+1} - \frac{\epsilon}{\hbar}H(\psi_i^*, \psi_{i+1})\right) \tag{94}$$

We take the limit as $\epsilon \to 0$ and $M \to \infty$, effectively recovering a continuous imaginary time. We rewrite the exponent, focusing on the term $\psi_i^*\psi_{i+1}$

$$\psi_i^*\psi_{i+1} = \psi_i^*\psi_i + \psi_i^*(\psi_{i+1} - \psi_i) \approx \psi^*(\tau)\psi(\tau) + \psi^*(\tau)\dot{\psi}(\tau)\epsilon \tag{95}$$

This allows us to express the partition function as a path integral over all possible field configurations

$$Z = \int \mathcal{D}[\psi^*]\mathcal{D}[\psi] \exp\left(\int_0^{\beta\hbar} d\tau \left[\psi^*(\tau)\dot{\psi}(\tau) - \frac{H(\psi^*(\tau), \psi(\tau))}{\hbar}\right]\right) \tag{96}$$

We rotate back to real time by substituting $\tau = it$, $d\tau = idt$, and $\dot{\psi} = \frac{d\psi}{d\tau} = -i\frac{d\psi}{dt}$. This transformation yields the path integral in real time

$$Z = \int \mathcal{D}[\psi^*]\mathcal{D}[\psi] \exp\left(i\int_0^T dt \left[i\psi^*\frac{\partial\psi}{\partial t} - \frac{H(\psi^*, \psi)}{\hbar}\right]\right) \tag{97}$$

where T is some final time. Comparing this to the general form of the path integral, $Z = \int \mathcal{D}[\psi^*]\mathcal{D}[\psi] \exp\left(i\int dt \int d^N x \mathcal{L}[\psi, \psi^*]\right)$, we identify the Lagrangian density as

$$\mathcal{L} = \int d^N x \left[i\psi^* \frac{\partial \psi}{\partial t} - \frac{H(\psi^*, \psi)}{\hbar} \right] \tag{98}$$

To obtain a specific Lagrangian, we consider a Hamiltonian corresponding to the Schrödinger equation with a potential term

$$H = \int d^N x \left[\frac{\hbar^2}{2m} \sum_{i=1}^{N} \left| \frac{\partial \psi}{\partial x_i} \right|^2 + V(\psi^*, \psi) \right] \tag{99}$$

where the first term represents kinetic energy and $V(\psi^*, \psi)$ is a potential energy term. Substituting this Hamiltonian into the Lagrangian density and ensuring the Lagrangian is real by adding the complex conjugate and dividing by 2, we arrive at

$$\mathcal{L} = \int d^N x \left[\frac{i\hbar}{2} \left(\psi^* \frac{\partial \psi}{\partial t} - \psi \frac{\partial \psi^*}{\partial t} \right) - \frac{\hbar^2}{2m} \sum_{i=1}^{N} \left| \frac{\partial \psi}{\partial x_i} \right|^2 - V(\psi^*, \psi) \right] \tag{100}$$

This Lagrangian density can be separated into linear $(\mathcal{L}_S)$ and nonlinear $(\mathcal{L}_N)$ parts

$$\mathcal{L}_S = \frac{i\hbar}{2} \left(\psi^* \frac{\partial \psi}{\partial t} - \psi \frac{\partial \psi^*}{\partial t} \right) - \frac{\hbar^2}{2m} \sum_{i=1}^{N} \left| \frac{\partial \psi}{\partial x_i} \right|^2 \tag{101}$$

$$\mathcal{L}_N = -V(\psi^*, \psi) \tag{102}$$

Therefore, the total Lagrangian density is

$$\mathcal{L} = \frac{i\hbar}{2} \left(\psi^* \frac{\partial \psi}{\partial t} - \psi \frac{\partial \psi^*}{\partial t} \right) - \frac{\hbar^2}{2m} \sum_{i=1}^{N} \left| \frac{\partial \psi}{\partial x_i} \right|^2 + \mathcal{L}_\mathcal{N} \tag{103}$$

where $\mathcal{L}_N$ depends on the specific form of the potential $V(\psi^*, \psi)$. Examples include $\mathcal{L}_\mathcal{N} = -\frac{\gamma}{2}|\psi|^4$ for the nonlinear Schrödinger equation and $\mathcal{L}_\mathcal{N} = \mu^2|\psi|^2 - 2\lambda|\psi|^4$ for the Mexican hat potential.

6.4. Introduction of Gauge Field and U(1) Symmetry

To manage the contextual sensitivity of semantic uncertainty and prevent the arbitrary creation or destruction of influenceable Semantic Noise, we impose a U(1) symmetry, guaranteeing the conservation of semantic charge. To implement this, we introduce a gauge field, $A_\mu(t, \vec{x})$, and modify the derivatives to covariant derivatives. We also decompose the field into modes

$$\psi(x) = \sum_i c_i(x)\phi_i(x) \tag{104}$$

The gauge transformations for the fields are

$$\psi_i(x_\mu) \to \psi_i'(x_\mu) = e^{iq_i\varphi(x_\mu)}\psi_i(x_\mu) \tag{105}$$

$$\psi_i^*(x_\mu) \to \psi_i'^*(x_\mu) = e^{-iq_i\varphi(x_\mu)}\psi_i^*(x_\mu) \tag{106}$$

$$A_\mu(x_\mu) \to A'_\mu(x_\mu) = A_\mu(x_\mu) + \partial_\mu\varphi(x_\mu) \tag{107}$$

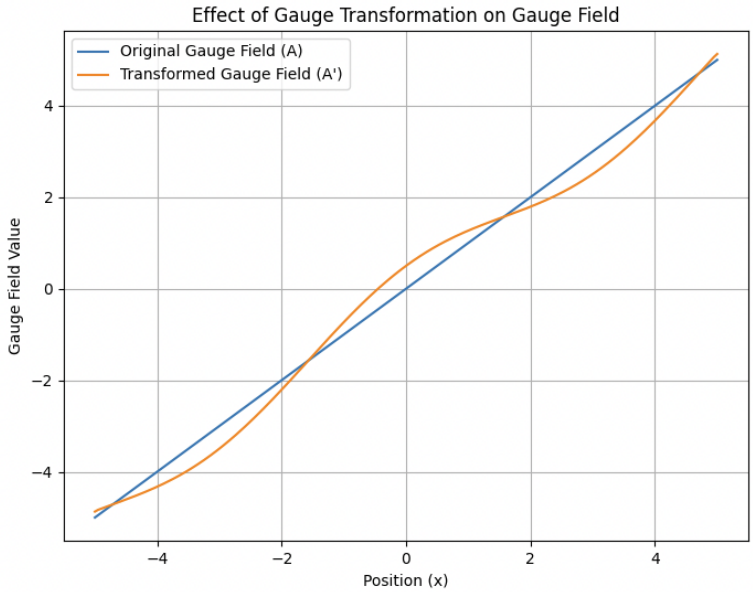


Figure 10: Gauge transformation of the gauge field in the LLM Embedding System. The figure illustrates the original gauge field, $A(x)$, and the transformed gauge field, $A'(x) = A(x) + \partial_x \varphi(x)$, after applying a gauge transformation with a sinusoidal gauge transformation function, $\varphi(x) = \phi \sin(x)$, where $\phi = 0.5$ is the amplitude of the gauge transformation. This transformation represents a change in the external context influencing the LLM's semantic representations.

The gauge transformation $A_\mu(x_\mu) \to A'_\mu(x_\mu) = A_\mu(x_\mu) + \partial_\mu \varphi(x_\mu)$ is visualized in Figure 10, which shows the original gauge field $A(x)$ and the transformed gauge field $A'(x)$ after applying a sinusoidal gauge transformation. This transformation represents a change in the external context influencing the LLM's semantic representations, while preserving the underlying gauge invariance of the system.

We replace the ordinary derivatives with covariant derivatives

$$D_{\mu,i} = \partial_\mu - iq_i A_\mu \tag{108}$$

Note that the semantic charge q_i is now mode-specific. The gauge-invariant Lagrangian is then

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \sum_i \frac{i}{2} (\psi_i^* D_{0,i} \psi_i - \psi_i (D_{0,i} \psi_i)^*) - \sum_i \frac{1}{2} |D_{j,i} \psi_i|^2 + \mathcal{L}_{\mathcal{N}} \tag{109}$$

where

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \tag{110}$$

Expanding the Lagrangian, we get

$$\begin{aligned} \mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \sum_i \frac{i}{2} (\psi_i^* \partial_0 \psi_i - \psi_i \partial_0 \psi_i^*) + \sum_i q_i A_0 |\psi_i|^2 - \sum_i \frac{1}{2} |\partial_j \psi_i|^2 \\ - \sum_i \frac{1}{2} q_i^2 A_j^2 |\psi_i|^2 - \sum_i \frac{i}{2} q_i A_j (\psi_i^* \partial_j \psi_i - \psi_i \partial_j \psi_i^*) + \mathcal{L}_{\mathcal{N}} \end{aligned} \tag{111}$$

6.5. Gauge Fixing and Generating Functional

To remove the gauge freedom, we choose the Coulomb gauge

$$\partial_i A_i = 0 \tag{112}$$

Using the Faddeev-Popov procedure, we arrive at a generating functional

$$Z = \int \mathcal{D}[\psi]\mathcal{D}[\psi^*]\mathcal{D}[A]\delta\left[\partial_i A_i\right] \exp\left(iS[\psi, \psi^*, A]\right) \tag{113}$$

where the effective action is

$$\begin{aligned} S[\psi, \psi^*, A] = \int dt \int d^N x \Bigg(& -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \sum_i \frac{i}{2}(\psi_i^* \partial_0 \psi_i - \psi_i \partial_0 \psi_i^*) + \sum_i q_i A_0 |\psi_i|^2 - \sum_i \frac{1}{2} |\partial_j \psi_i|^2 \\ & - \sum_i \frac{1}{2} q_i^2 A_j^2 |\psi_i|^2 - \sum_i \frac{i}{2} q_i A_j (\psi_i^* \partial_j \psi_i - \psi_i \partial_j \psi_i^*) + \mathcal{L}_{\mathcal{N}} \Bigg) \end{aligned} \tag{114}$$

As demonstrated in [29], the effective Lagrangian, derived in the weak coupling approximation, governs the dynamics of the ψ field, incorporating the influence of integrated-out fields up to quadratic order. This framework reveals a non-local interaction, mediated by the Green's function, where the semantic meaning at one location in semantic space directly affects the meaning at distant locations. This non-locality may underlie the capacity of LLMs, particularly the Transformer architecture with its attention mechanism, to model long-range dependencies in text. Furthermore, this non-local interaction bears a conceptual resemblance to quantum entanglement, suggesting a possible connection between the LLM's ability to capture semantic relationships and fundamental quantum mechanical phenomena.

6.6. Semantic Charge, Gauge Field, and Nonlinear Potential

Having derived the gauge-invariant Lagrangian, we now turn to interpreting the key elements of this formalism in the context of LLMs. These elements, namely the semantic charge, the gauge field, and the nonlinear potential, provide a powerful framework for understanding how context and nonlinearity shape the dynamics of semantic representations and influence the level of Semantic Noise within the system.

The semantic charge, q_i, quantifies the sensitivity of the phase of the mode associated with Semantic Noise to contextual influences. It determines how strongly the linear dynamics of a particular mode interact with the gauge field. A higher semantic charge indicates a greater susceptibility to contextual modulation of the phase, influencing the system's tendency to explore alternative semantic interpretations. For example, a word with a high semantic charge might be more sensitive to the surrounding words in a sentence, exhibiting greater contextual flexibility, while a word with a low semantic charge might retain its meaning regardless of the context, exhibiting greater semantic stability.

The gauge field, $A_\mu = (A_0, A_i)$, represents the influence of context on the phase of the mode associated with Semantic Noise. The temporal component, A_0, acts as a semantic force that shapes the evolution of the phase over time, influencing the system's balance between coherence and exploration. The spatial components, A_i, may be interpreted as representing the flow of semantic information through the embedding space, with their direction and magnitude indicating the strength and direction of the flow. This context could be a prompt, the surrounding text, or external knowledge.

The nonlinear potential, V_{NL}, introduced in Layer 3, captures the nonlinear interactions and feedback loops that are characteristic of LLMs. It allows the semantic wave function to interact with itself, leading to more complex and realistic dynamics that go beyond the linear approximations of Layer 2. The specific form of the nonlinear potential (e.g., cubic nonlinearity or Mexican hat potential) determines the nature of these interactions. For example, a cubic nonlinearity could model the reinforcement of biases, while a Mexican hat potential might promote semantic stability by providing a preferred state.

The interplay between the semantic charge, the gauge field, and the nonlinear potential determines the overall behavior of the LLM Embedding System. The semantic charge governs the sensitivity of the phase of the mode associated with Semantic Noise to context, the gauge field represents the context itself (both its temporal influence and its spatial flow), and the nonlinear potential shapes the dynamics of the semantic representation. By analyzing these elements, we can gain a deeper understanding of how LLMs process and generate language, balancing coherence with the exploration of new semantic possibilities.

7. Layers 4-N: Advanced Quantum Hierarchies

While Layers 1, 2, and 3 provide a foundation for understanding LLM representations, the quantum framework allows for the possibility of even more advanced and complex hierarchies. These possible layers, which we denote as Layers 4-N, can capture aspects of LLM

behavior that are beyond the reach of the simpler models.

One intriguing possibility is to model the dynamic creation and annihilation of semantic content within the LLM. This can be particularly relevant for understanding how LLMs learn new information, forget old information, or generate new ideas. In such cases, theoretical frameworks like Quantum Field Theory (QFT), which describes the creation and annihilation of particles, might offer advanced tools. QFT introduces the concept of quantum fields that permeate all of space, and particles are seen as excitations of these fields. Analogously, we can think of a semantic field that underlies the LLM's embedding space, with words and concepts representing excitations of this field. The creation and annihilation operators in QFT can then be used to model the dynamic addition and removal of semantic content. For example, the concept of 'semantic charge' introduced in Layer 2 can be further explored using QFT's charge conservation laws, providing insights into how LLMs manage Semantic Noise and balance coherence with exploration in their generated text.

Another speculative direction is to investigate the potential for emergent, string-like structures within the LLM's embedding space. This idea draws inspiration from Quantum String Theory, which proposes that fundamental particles are not point-like but rather tiny, vibrating strings. In the context of LLMs, we could think that semantic relationships are not simply point-to-point connections but rather more complex, string-like objects that encode richer information about the relationships between concepts. The Transformer architecture space is likely to exhibit singularities, sinks, and wells, which are concepts from QFT and String Theory. These can be interpreted as regions where semantic content is either created (sources) or destroyed (sinks), or as points of instability in the semantic landscape. However, even if String Theory appears distinct at the moment, it is a very powerful mathematical toolbox that has been applied in particle physics, condensed matter physics, nuclear physics, and many other areas. Even if a direct physical connection to String Theory remains elusive, the mathematical tools developed within that framework can offer valuable techniques for analyzing the complex structure of the LLM embedding space.

It is crucial to emphasize that at the moment these are speculative ideas, and a direct mapping to LLM representations is not clear. Also, the mathematical complexity of QFT and String Theory is considerable. However, if such mappings could be rigorously established, LLMs might even offer a new domain for exploring theoretical concepts from these advanced frameworks, providing empirical insights into areas of physics where direct observation is currently very difficult. However, at this stage, these connections remain largely at the level of analogy and inspiration.

8. Layer N+1: Transformer Architecture (Common LLM Architecture)

The Transformer architecture, introduced by Vaswani et al., represents a significant milestone in our layered hierarchy and a major advancement in neural network design for natural language processing. Unlike recurrent neural networks (RNNs) that process sequential data step-by-step, Transformers rely entirely on attention mechanisms to model relationships between words in a sequence, enabling parallel processing and improved performance on longrange dependencies. This architecture provides a concrete realization of the complex dynamics and interactions that we have been modeling in Layers 2 and 3. The key connection to our quantum framework is that the Transformer's attention mechanism can be interpreted as implementing a form of non-local interaction, analogous to the path integral formalism in Layer 3, where the representation of each word is influenced by all other words in the sequence, regardless of distance.

A core component of the Transformer is the self-attention mechanism. This allows the model to weigh the importance of different words in the input sequence when representing a particular word. While attention weights can be interpreted as probabilities of the LLM occupying different semantic states, it is important to note that they are not strictly probabilities in the mathematical sense, as they are not explicitly normalized to sum to one across all possible states. The attention weights are computed based on the relationships between the query, key, and value vectors derived from the input embeddings. This mechanism enables the model to capture contextual information and understand the relationships between words regardless of their distance in the sequence, effectively implementing a form of non-local interaction similar to that captured by the path integral formalism in Layer 3, whereby the representation of each word is influenced by all other words in the sequence, regardless of distance.

The Transformer architecture typically consists of an encoder and a decoder. The encoder processes the input sequence and generates a contextualized representation. The decoder then uses this representation to generate the output sequence, such as in machine translation or text summarization tasks. Both the encoder and decoder are composed of multiple layers of self-attention and feedforward networks.

Each layer in the encoder and decoder includes a multi-head attention mechanism. This allows the model to attend to different aspects of the input sequence simultaneously, capturing a richer set of relationships between words. The outputs of the multiple attention heads are then concatenated and linearly transformed to produce the final output of the layer.

Feedforward networks, typically consisting of two linear transformations with a nonlinear activation function in between, are applied to each position in the sequence independently. These networks introduce nonlinearities into the model, allowing it to learn complex

relationships between semantic concepts. These nonlinearities, such as ReLU or GeLU, can be seen as analogous to the nonlinear potential V_{NL} in Layer 3, shaping the energy landscape of the semantic space and contributing to the stability and complexity of the learned representations. The Transformer's layers, attention mechanisms, and feedforward networks collectively define a complex energy landscape that governs the dynamics of semantic meaning, learning to shape this landscape during training to minimize a loss function while also managing the level of Semantic Noise to enable both coherence and exploration.

Residual connections and layer normalization are also key features of the Transformer architecture. Residual connections allow the model to bypass certain layers, facilitating the flow of information and preventing vanishing gradients during training. Layer normalization helps to stabilize the training process and improve the model's generalization performance.

9. Layer N+2: Complex Transformer Architecture (Complex Semantic Superspace)

While most Transformer-based LLMs have historically operated with real-valued representations, recent research has explored complex-valued Transformer architectures [37–39]. These models, often referred to as complex neural networks (CVNNs), extend the standard Transformer to operate directly on data with both real and imaginary components. This approach is particularly useful in applications where data is inherently complex-valued (e.g., signal processing, medical imaging) or where leveraging both amplitude and phase information is beneficial for capturing subtle semantic relationships and contextual nuances. Architectural adaptations include complex-valued attention mechanisms, layer normalization, and feed-forward networks. Some approaches, like Quantum-Inspired Complex (QIC) Transformers, aim to improve parameter efficiency by using learnable algebraic structures for the imaginary unit. These complex-valued Transformers have shown promise in improving robustness to overfitting and achieving competitive or superior performance compared to their real-valued counterparts in certain domains.

The significance of Complex Transformers within our quantum hierarchy lies in their ability to provide a more complete and natural representation of what we term the Complex Semantic Superspace, offering new ways to understand and manage Semantic Noise within LLMs. This terminology is inspired by concepts in theoretical physics, such as Superstring Theory and Supersymmetry, where the introduction of additional dimensions and complex fields allows for a more unified and complete description of physical phenomena. In our context, the Complex Semantic Superspace represents an extension of the traditional real-valued Transformer embedding space, incorporating complex-valued representations to capture aspects of semantic meaning, such as phase and interference, that are otherwise inaccessible.

In essence, the real-valued Transformer embedding space can be seen as a projection of this more complete Complex Semantic Superspace. This projection inevitably results in a loss of information, hindering the ability to fully leverage the quantum analogy. Complex Transformers, by operating directly with complex numbers, offer a richer representation that more closely aligns with the mathematical structure of quantum mechanics, allowing us to access the full potential of the Complex Semantic Superspace.

Therefore, the quantum tools and concepts developed in this hierarchy may find a more direct, natural, and optimal application within the Complex Semantic Superspace, leading to more effective methods for controlling Semantic Noise and harnessing its benefits for creativity and adaptation. The complex-valued nature of these architectures reduces the need for approximations, compactifications, or mappings compared to their application to real-valued architectures, unlocking deeper insights into the quantum-like properties of LLMs and leading to more effective methods for analysis and control. Layer N+2, therefore, provides a more faithful representation of the underlying semantic reality within the Complex Semantic Superspace.

10. Linearization and Classical Limit

Having established a quantum framework for analyzing LLM representations, we now demonstrate its consistency with existing approaches by exploring the connection between this framework and the linearized LLM embedding space (Layer 1). We aim to show how the familiar properties of the embedding space can be recovered by making specific approximations to the time-dependent Schrödinger equation, effectively "undoing" the quantum enhancements we introduced in Layers 2 and 3. This linearization provides a valuable validation of our framework and a theoretical justification for the effectiveness of linear operations in analyzing semantic representations, while also highlighting the information lost in the process. We aim to show how the static relationships between semantic concepts, as captured by cosine similarity in Layer 1, can be recovered.

10.1. Linearization of U (1) Symmetry

We first demonstrate the linearization of a quantum mechanical system with U(1) symmetry. We begin with the time-dependent Schrödinger equation with the covariant derivative

$$i\hbar D_t|\tilde{\psi}(t)\rangle = \hat{\mathbf{D}}|\tilde{\psi}(t)\rangle \tag{115}$$

where $D_t = \partial_t - iqA_0(t)$ and $\hat{\mathbf{D}}$ is the diagonalized Hamiltonian. Because $\hat{\mathbf{D}}$ has only one non-zero eigenvalue, it effectively projects the system onto a single active mode. This simplifies the interaction with the gauge field, allowing us to focus solely on the temporal component, $A_0(t)$, and implicitly set the spatial components, A_i, to zero without loss of generality.

To obtain a linearized model that is comparable to the static LLM embedding space, we make the following approximations: We assume that the system is in a stationary state, with a simple time dependence

$$|\tilde{\psi}(t)\rangle = e^{-iEt/\hbar}|\tilde{\psi}\rangle \tag{116}$$

where E is the energy of the state and $|\tilde{\psi}\rangle$ is a time-independent eigenvector of the Hamiltonian. We also assume that the gauge field is static (time-independent) and represents an inherent property of the embedding space, rather than an external influence. Therefore, $A_0(t) = A_0$, where A_0 is a constant. Applying these approximations, the Schrödinger equation simplifies to

$$[E + \hbar q A_0]|\psi\rangle = \hat{\mathbf{D}}|\psi\rangle \tag{117}$$

This is the eigenvalue equation for the Hamiltonian, a linear equation that relates the shifted energy of the system (due to the static gauge field) to the eigenvalues of the diagonalized Hamiltonian. The time-independent state vector $|\psi|$ is mapped to the embedding vector $\mathbf{a}$. The shifted energy of the state, $E + \hbar q A_0$, is mapped to the semantic coherence measure S'_C. The diagonalized Hamiltonian $\hat{\mathbf{D}}$ represents the fundamental modes of the semantic space.

This linearization demonstrates that the LLM embedding space can be seen as a simplified, static version of the more complex quantum system. This version retains a static, inherent contextual influence, represented by the gauge field A_0, primarily on the phase of the mode associated with Semantic Noise. The linearization process removes the time dependence, leaving only the essential linear relationships between the semantic concepts, modulated by the static gauge field. If we further set A_0 to zero, we recover the original LLM Embedding System, effectively neglecting the U(1) symmetry and any inherent contextual influence on Semantic Noise.

10.2. Linearized Partition Function with Static Context

As an example of the linearization process described above, we can derive the linearized partition function. Starting from the quantum partition function with a static gauge field, and applying the approximations of a stationary state, we have

$$Z' = \mathrm{Tr}\left[e^{-\beta(\hat{\mathbf{H}}' + \hbar q A_0)}\right] \tag{118}$$

Since the gauge field is static, it influences the phase of the mode associated with Semantic Noise, leading to a scaling of the partition function. We can rewrite this as

$$Z' = \mathrm{Tr}\left[e^{-\beta(\hat{\mathbf{D}}')}\right] = e^{-\beta\hbar q A_0}\mathrm{Tr}\left[e^{-\beta\hat{\mathbf{D}}}\right] \tag{119}$$

where $\hat{\mathbf{D}}'$ is diagonalized Hamiltonian with the static gauge field. Using the eigenvalues of the original diagonalized Hamiltonian $\hat{\mathbf{D}}$, which are λ_i, we obtain

$$Z' = e^{-\beta\hbar q A_0}\sum_{i=1}^{N} e^{-\beta\lambda_i} \tag{120}$$

Factoring out the common term, and using the fact that $\lambda_1 = 1$ and $\lambda_i = 0$ for $i = 2,3,...,N$ we obtain

$$Z' = e^{-\beta\hbar q A_0}\left[e^{-\beta} + (N-1)\right] \tag{121}$$

This result shows that the static gauge field scales the original partition function by a factor of $e^{-\beta\hbar q A_0}$, influencing the relative probabilities of the different states in the linearized system. The term $e^{-\beta\hbar q A_0}$ introduces a contextual bias, influencing the system's balance between coherence and exploration, depending on the sign and magnitude of qA_0.

- If $qA_0 > 0$, the static gauge field promotes coherence, leading to a more focused and stable LLM, but limiting its creativity.
- If $qA_0 < 0$, the static gauge field promotes exploration, leading to a more diffuse and unstable LLM, but enhancing its creativity.
- If $qA_0 = 0$, the static gauge field has no effect, and we recover the original, context-free LLM Embedding System.

This static gauge field indirectly modulates the effects of dimensionality and temperature on semantic coherence by influencing the balance between coherence and exploration. A positive contextual bias amplifies the effect of dimensionality, promoting stronger coherence, while a negative bias counteracts it, promoting greater exploration. The static gauge field also influences few-shot learning, robustness to adversarial attacks, and the trade-off between coherence and creativity. It suggests that some hallucinations may originate from an inherent contextual bias within the LLM's embedding space, skewing the balance between coherence and exploration. The training data likely influences both the effective temperature and the static gauge field, with consistent datasets leading to a positive A_0 and noisy datasets leading to a negative A_0.

10.3. Linearized Analysis of the Landau-Zener System

Next, we consider the linearization of the Landau-Zener model. To understand the long-term behavior of the system under dynamic contextual influences, we consider the limit as time approaches infinity. Here, "time" represents the LLM's internal processing time, and the limit as t approaches infinity corresponds to the LLM settling into its dominant semantic state after fully processing the prompt. If the system starts in the state $|\tilde{0}\rangle$ at $t \to -\infty$, the asymptotic populations of the coherent and less coherent states are given by

$$P_0(\infty) = |\langle x_2|\tilde{\psi}(\infty)\rangle|^2 = 1 - e^{-\pi\Delta^2/(2|v|)} \tag{122}$$

and

$$P_1(\infty) = |\langle x_1|\tilde{\psi}(\infty)\rangle|^2 = e^{-\pi\Delta^2/(2|v|)} \tag{123}$$

where $|\tilde{0}\rangle = |x_2\rangle$ and $|\tilde{1}\rangle = |x_1\rangle$ are the eigenvectors of the diagonalized Hamiltonian $\hat{\mathbf{D}}$, corresponding to eigenvalues 0 and 1. In the linearized limit, we can drop the tildes and write

$$P_0(\infty) = |\langle x_2|\psi(\infty)\rangle|^2 = 1 - e^{-\pi\Delta^2/(2|v|)} \tag{124}$$

and

$$P_1(\infty) = |\langle x_1|\psi(\infty)\rangle|^2 = e^{-\pi\Delta^2/(2|v|)} \tag{125}$$

where $|0\rangle = |x_2\rangle$ and $|1\rangle = |x_1\rangle$ are the eigenvectors of the diagonalized Hamiltonian $\mathbf{D}$, corresponding to eigenvalues 0 and 1. This approach allows us to directly connect the Landau-Zener parameters (v, Δ) to the relative probabilities of the coherent and states associated with higher Semantic Noise. In this limit, the system's final state is entirely determined by the rate of change of context (v) and the coupling strength between the states (Δ). This provides a simplified view of the system's long-term behavior, where the details of the time evolution are no longer relevant. This suggests that hallucinations are more likely to occur when the context is not stable, making it difficult for the LLM to suppress the state associated with higher Semantic Noise and maintain a clear and well-defined dominant semantic interpretation, leading to uncontrolled exploration of less probable semantic combinations.

10.4. Linearized Analysis of the Bright Soliton

While the linearization of the bright soliton solution inevitably sacrifices its key dynamic and nonlinear properties, we can still explore whether it provides any insights into the structure of the LLM embedding space. By "freezing" the solution at a specific time t_0 and position x_0, we are essentially extracting a static snapshot of the dynamic semantic representation, effectively eliminating its time-dependent and spatial characteristics. This allows us to analyze the amplitude at a single point in the semantic space, mimicking the static nature of the linearized embedding space. We start with the transformed wave function. In the linearized limit, we drop the tilde and obtain a constant complex amplitude

$$\psi_0 = \psi(x_0, t_0) = A\text{sech}\left[B(x_0 - vt_0)\right] e^{i(kx_0 - \omega t_0)} \tag{126}$$

We can then express this constant amplitude as a linear combination of the eigenvectors of the diagonalized Hamiltonian $\hat{\mathbf{D}}$. This linear combination effectively decomposes the static amplitude into its constituent semantic components, as defined by the eigenvectors of the Hamiltonian, allowing us to analyze the contribution of each component to the overall semantic representation.

Although this linearization is highly limiting, we can speculate on potential connections to the real embedding space:

• Preferred Semantic Locations: The existence of non-zero solutions only for specific values of x_0 might suggest that the real embedding space has preferred locations for localized patterns of semantic meaning, as captured by the soliton solutions, although these locations are likely distorted and simplified due to the linearization process.

• Amplitude as Semantic Salience: The magnitude of the constant amplitude, $|\psi_0|$, might be related to the semantic salience or importance of the concept represented by the soliton.
• Phase as Semantic Context: The phase of the constant amplitude, $(kx_0 - \omega t_0)$, might encode some information about the semantic context, analogous to how the phase of a wave can encode information about its source or propagation history, although there is no clear way to directly map this to the real embedding space.

We may hypothesize that LLM spaces contain soliton-like solutions. However, this is a good example that clearly shows linearization is a significant simplification and we lose many internal LLM structures when doing so. This also shows that if we want to understand LLM architectures and spaces, we need more advanced tools beyond the linear embedding space, and investigate the internal structure of LLMs. The quantum framework presented in this article offers one approach for achieving this, providing a set of tools and concepts that can capture the nonlinear dynamics and complex relationships within LLMs that are lost in linear approximations.

11. Quantum Computation Probing LLM Embeddings

Having established the ”LLM Embedding Quantum System” as an exact quantum mechanical analogue of the classical LLM embedding space, we now explore the possibility for leveraging quantum computers to directly probe and analyze this system, pushing beyond the limitations of classical methods. The simplicity of the LLM Embedding Quantum System makes it particularly amenable to experimental investigation using existing quantum computing technology, offering a unique opportunity to validate our theoretical framework and gain new insights into the behavior of LLMs. Quantum computing offers the potential to move from modeling the ”LLM Embedding Quantum System” to directly impacting LLM development and usage in the future.

As demonstrated in previous work [31, 33], quantum computers can efficiently calculate quantities such as cosine similarity and the quantum partition function. Here, we introduce a straightforward approach for calculating cosine similarity using a simple quantum circuit.

The value

$$S'_C = \frac{\sqrt{2P(0) - 1} + 1}{2} \tag{127}$$

can be efficiently estimated using a quantum circuit known as the SWAP test. This test leverages quantum interference to relate the probability of measuring a specific state to the real part of the inner product between two quantum states, $|a\rangle$ and $|b\rangle$.

11.1. Quantum Circuit and Algorithm

The SWAP test circuit consists of three registers: one ancilla qubit initialized to $|0\rangle$, and two registers to hold the quantum states $|a\rangle$ and $|b\rangle$. To represent a vector of dimension N in a quantum computer, we require n qubits, where n is the smallest integer such that $2^n \geq N$. If N is not a power of 2, we must employ a technique called padding. Padding involves increasing the dimension of the vector to the next power of 2 by adding extra components with zero amplitude. For example, if we have a vector of dimension 10, we need 4 qubits (since $2^4 = 16$). We then pad the vector with 6 zeros to make it a 16-dimensional vector. This ensures that the vector can be represented by the 4 qubits.

To illustrate how multiple qubits represent a higher-dimensional vector, consider a system of 3 qubits. The state of this system can be written as a superposition of all possible 3-qubit basis states

$$|\psi\rangle = \alpha_{000}|000\rangle + \alpha_{001}|001\rangle + \alpha_{010}|010\rangle + \alpha_{011}|011\rangle + \alpha_{100}|100\rangle + \alpha_{101}|101\rangle + \alpha_{110}|110\rangle + \alpha_{111}|111\rangle \tag{128}$$

where α_{ijk} are complex amplitudes and $|ijk\rangle$ represents the basis states (e.g., $|000\rangle$, $|001\rangle$, etc.). This 3-qubit system can represent a vector in an 8-dimensional space, where each basis state corresponds to one dimension. The amplitudes αijk are the components of the vector in this basis. The tensor product structure is implicit in the notation $|ijk\rangle = |i\rangle \otimes |j\rangle \otimes |k\rangle$, where each $|i\rangle$, $|j\rangle$, and $|k\rangle$ is a single-qubit state. The circuit diagram is shown in Figure 11.

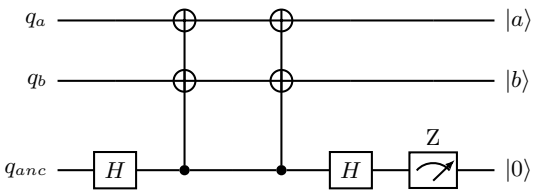


Figure 11: Quantum circuit for estimating S'_C. The ancilla qubit (q_{anc}) is initialized to $|0\rangle$ and undergoes a Hadamard gate. A controlled-SWAP gate (Fredkin gate), controlled by the ancilla, is applied between the registers holding $|a\rangle$ and $|b\rangle$. Another Hadamard gate is applied to the ancilla, followed by a measurement in the Z-basis. The probability of measuring $|0\rangle$ on the ancilla qubit is used to estimate S'_C.

In the following, we detail the circuit operation and provide a mathematical derivation:
1. Initialization: The quantum states $|a\rangle$ and $|b\rangle$ are prepared in their respective registers. The ancilla qubit (q_{anc}) is initialized to the $|0\rangle$ state.
2. Hadamard Gate: A Hadamard gate (H) is applied to the ancilla qubit, creating the superposition state

$$\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle) \tag{129}$$

3. Controlled-SWAP Gate: A controlled-SWAP gate (also known as a Fredkin gate) is applied. This gate swaps the states in the $|a\rangle$ and $|b\rangle$ registers if the ancilla qubit is in the $|1\rangle$ state, and does nothing if the ancilla is in the $|0\rangle$ state.
4. Second Hadamard Gate: Another Hadamard gate is applied to the ancilla qubit.
5. Measurement: The ancilla qubit is measured in the Z-basis (computational basis). The probability of measuring $|0\rangle$, denoted as $P(0)$, represents the likelihood of finding the ancilla qubit in the $|0\rangle$ state after the circuit has been executed. This probability is given by

$$P(0) = \frac{1}{2} + \frac{1}{2}|\langle a|b\rangle|^2 \tag{130}$$

6. Estimating S'_C : From Eq. (127), we know that the probability of measuring $|0\rangle$ on the ancilla qubit, $P(0)$, is directly related to S'_C . Therefore, we can estimate the quantum mechanical expectation value of the transformed cosine similarity, denoted as $\langle S'_C\rangle$, by estimating $P(0)$ and using the relationship

$$\langle S'_C\rangle = \frac{\sqrt{2P(0) - 1} + 1}{2} \tag{131}$$

Let's define a random variable X_i for the outcome of the i-th shot, where

$$X_i = \begin{cases} 1 & \text{if the ancilla qubit is measured in the state } |0\rangle \\ 0 & \text{if the ancilla qubit is measured in the state } |1\rangle \end{cases} \tag{132}$$

Then, the expected value of X_i is

$$E[X_i] = 1 \cdot P(0) + 0 \cdot P(1) = P(0) \tag{133}$$

To estimate $P(0)$, we run the SWAP test circuit N_{shots} times. We can then calculate the sample mean of the X_i values:

$$\overline{X} = \frac{1}{N_{shots}} \sum_{i=1}^{N_{shots}} X_i \tag{134}$$

By the Law of Large Numbers, as N_{shots} becomes large, the sample mean $\overline{X}$ converges to the quantum mechanical expectation value $E[X_i] = P(0)$. We then use this estimate of P(0) to calculate the estimate of $\langle S'_C\rangle$:

$$\langle S'_C\rangle = \frac{\sqrt{2\overline{X} - 1} + 1}{2} \tag{135}$$

In other words, we run the circuit many times, sum the results (1 for each $|0\rangle$ measurement and 0 for each $|1\rangle$ measurement), and divide by the total number of runs to estimate P(0). We then use this estimate of $P(0)$ to calculate the estimate of $\langle S'_C \rangle$. The more shots we use, the more accurate our estimate of $\langle S'_C \rangle$ will be.

The SWAP test provides a quantum advantage for estimating the inner product compared to classical methods, especially when dealing with high-dimensional quantum states. The accuracy of the estimate depends on the number of shots (measurements) used; more shots lead to a more accurate estimate of $P(0)$ and, consequently, of S'_C. On real quantum hardware, the SWAP test is susceptible to noise and errors, making error mitigation techniques crucial for obtaining reliable results.

11.2 Experimental Validation with LLM Embeddings

As a realistic example, we used Google's EmbeddingGemma Sentence Transformer to calculate S'_C using two vectors. EmbeddingGemma is a 300M parameter, state-of-the-art open embedding model from Google, built from Gemma 3. EmbeddingGemma produces vector representations of text, making it well-suited for search and retrieval tasks, including classification, clustering, and semantic similarity search. We used the following two sentences:

|a⟩: "The tiny brown mouse quietly scurried across the kitchen floor, seeking crumbs."
|b⟩: "A small gray mouse cautiously crept along the wooden floor, searching for food."

and calculated their 768-dimensional embedding vectors. Since 768 is not a power of 2, we padded the vectors to dimension 1024, requiring 10 qubits per vector register. Representing each 1024-dimensional vector requires preparing a complex multi-qubit state across 10 qubits. Depending on the quantum computer's architecture and the chosen state preparation method, the initialization process may require hundreds or even thousands of CNOT gates, illustrating the significant overhead associated with preparing high-dimensional quantum states. Using Aer's `qasm_simulator`,, a quantum simulator with 4096 shots, we obtained the following result:

$$\langle S'_C \rangle_{\text{simulator}} = 0.9110 \tag{136}$$

The corresponding exact result is

$$S'_{C,\text{exact}} = 0.9115 \tag{137}$$

This suggests the potential for using the SWAP test as a viable method for estimating semantic similarity in LLM embedding spaces.

We also performed experiments using OpenAI's `text-embedding-3-small` model, which generates 64-dimensional embedding vectors. These lower-dimensional embeddings allowed for a more compact quantum representation with a total of 13 qubits. Again, using Aer's `qasm_simulator` with 4096 shots, we obtained

$$\langle S'_C \rangle_{\text{simulator}} = 0.84035 \tag{138}$$

The corresponding exact result is

$$S'_{C,\text{exact}} = 0.84038 \tag{139}$$

Again, we observed good agreement with the results, suggesting that the approach is also independent of the LLM embedding model that is used.

However, when testing the same approach with the 768-dimensional EmbeddingGemma vectors on a real quantum computer accessed through the IBM Quantum Cloud (IBM Quantum's ibm fez and ibm boston backends), the results differed significantly. Instead of reflecting the semantic similarity, the SWAP test consistently yielded a P(0) value close to 0.5. This suggests that the quantum computation was overwhelmed by noise and decoherence, effectively randomizing the state of the ancilla qubit and obscuring any meaningful signal. While noise and decoherence are primary suspects, other factors could also be contributing. The initial state preparation, particularly the multiqubit initialization required for these high-dimensional vectors, is a non-trivial task and a potential source of error. The transpilation process, which optimizes the circuit for the specific quantum hardware, can also introduce errors. It's also conceivable that the Qiskit initialization script might introduce unintended normalizations or transformations, leading to an incorrect initial state. Other initialization problems, or even a subtle programming error, cannot be entirely excluded. Ideally, we would have investigated these possibilities more thoroughly, but the prohibitively high cost and limited availability of real quantum computation resources prevented us from pursuing this line of inquiry further, including exploring error correction or noise extrapolation techniques.

While these results on publicly available quantum hardware were not as strong as theoretically predicted, it is important to note that this implementation of the SWAP test was designed to minimize the number of qubits required, a trade-off that introduced challenges in circuit state preparation and made the circuit particularly susceptible to noise and decoherence on current quantum devices. In contrast, our previous work [31, 33] explored alternative quantum circuit designs where each qubit represented a single dimension of the embedding vector. While these designs required significantly more qubits, they yielded more promising results, suggesting that the quantum approach is viable with sufficient quantum resources and more robust state preparation techniques. It is also crucial to acknowledge that the publicly available quantum hardware used in these experiments does not necessarily represent the current state of-the-art in quantum computing, and more advanced quantum devices with lower noise levels and improved gate fidelities may be able to achieve significantly better results with the SWAP test or other quantum algorithms.

These experiments highlight the challenges of applying quantum algorithms to real-world data, particularly when dealing with high-dimensional representations and noisy quantum hardware. While the SWAP test offers a theoretically appealing approach for estimating semantic similarity with a relatively small number of qubits, achieving reliable and accurate results in practice requires careful consideration of the characteristics of the input data, the limitations of current quantum hardware, and the choice of appropriate quantum algorithms and error mitigation techniques.

To improve the accuracy and reliability of quantum semantic similarity estimation, we suggest exploring the following:

• Alternative Circuit Designs: Exploring alternative quantum circuit designs for the SWAP test that are more robust to noise and better suited for representing vectors with varying degrees of sparsity. This might involve using different gate decompositions or employing error mitigation techniques.
• Quantum Dimensionality Reduction: Investigating dimensionality reduction techniques to reduce the dimension of the embedding vectors while preserving their essential semantic information. This could lead to more efficient and accurate quantum representations.
• Patch-Based Similarity Calculation: Considering embedding vectors as ”patches,” i.e., dividing long vectors into several smaller vectors (then normalizing them to 1) and performing similarity calculations in pieces. This could allow for a more localized and potentially more robust estimation of similarity.
• Improved Quantum Hardware: The development of more robust and higher-fidelity quantum hardware is essential for realizing the full potential of quantum algorithms for natural language processing.
• Exploiting Embedding Structure: Developing quantum algorithms that are specifically tailored to exploit the structure and properties of LLM embedding spaces. This might involve using different encoding schemes or designing circuits that are optimized for specific types of semantic relationships.

In conclusion, while the theoretical framework presented in this article offers a promising approach for understanding and analyzing LLM representations using quantum computing, challenges still remain in translating these ideas into practical and reliable quantum algorithms. More focus is needed to overcome these challenges and to unlock the full potential of quantum computing for natural language processing. The limitations we encountered also highlight the importance of carefully considering the characteristics of the input data and the capabilities of the available quantum hardware when designing and implementing quantum algorithms.

12. Applications and Emerging Research

The quantum circuit design presented here offers a direct and intuitive, albeit simplified, approach to calculating cosine similarity using a quantum computer. While our experiments revealed limitations in its ability to accurately estimate semantic similarity for complex LLM embeddings with current quantum hardware, the circuit’s simplicity provides a valuable starting point for exploring more advanced quantum algorithms.

More broadly, the existence of an exact quantum mechanical analogue, the LLM Embedding Quantum System, opens up a wide range of experimental possibilities. By modifying the quantum circuit, we could explore and test various aspects of the LLM embedding space. Some easier applications (near-term focus) that could be explored include:

• U(1) Symmetry and Contextual Control: Develop quantum-inspired methods for dynamically controlling LLM contextual sensitivity by measuring and adjusting semantic charge flow. This enables hallucination mitigation, creative text generation, and personalized LLMs. (Focus: Quantum measurement of semantic charge flow, feedback loop implementation)
• Quantum Tunneling and Hallucination Suppression: Create a quantum-inspired hallucination detection and correction mechanism by simulating quantum tunneling, identifying potential hallucinations, and using quantumassisted correction to steer LLMs towards more accurate states. (Focus: Quantum simulation of tunneling, identification of high-probability tunneling events)
• Quantum Entanglement and Semantic Understanding: Enhance LLM understanding of complex relationships by measuring

entanglement between embedding dimensions and training LLMs to exploit entanglement patterns. This improves reasoning, contextual understanding, and knowledge graph construction. (Focus: Measurement of entanglement between embedding dimensions, entanglement-aware training strategies)

More ambitious research topics (long-term vision) that could significantly advance the field include, for example,

• Quantum-Enhanced Few-Shot Learning: Improve LLM few-shot learning using quantum state transfer (e.g., teleportation) to efficiently transfer knowledge from a small number of examples. Explore quantum generative models to augment limited training data. This enables rapid adaptation, personalized learning, and deployment in resource-constrained environments. (Quantum Property: Superposition and entanglement for efficient knowledge transfer and data augmentation)
• Quantum-Inspired Semantic Search: Develop a semantic search engine using quantum embedding spaces and algorithms like Grover's algorithm or Quantum Amplitude Estimation to retrieve information based on meaning rather than keywords. This improves information retrieval, knowledge discovery, and personalized recommendations. (Quantum Property: Grover's algorithm for speedup in unstructured search, quantum amplitude estimation for efficient probability estimation)
• Quantum-Assisted Code Generation: Enhance code generation by LLMs using quantum algorithms for code optimization (e.g., quantum annealing to find optimal code structures) and bug detection (e.g., quantum pattern matching to identify potential vulnerabilities). This automates software development, improves code quality, and reduces development costs. (Quantum Property: Quantum annealing for optimization, quantum pattern matching for efficient bug detection)
• Quantum-Enhanced Multi-Modal LLMs: Extend the quantum framework to multi-modal LLMs by using quantum algorithms to fuse information from different modalities into a unified quantum representation (e.g., quantum feature maps for modality encoding) and enable quantum cross-modal reasoning (e.g., quantum algorithms for pattern recognition across modalities). This creates more powerful AI systems and enables new applications. (Quantum Property: Quantum feature maps for efficient encoding of multi-modal data, quantum algorithms for pattern recognition and reasoning)

The ability to perform these experiments on a quantum computer provides a powerful tool for validating our theoretical framework and gaining insights into the complex dynamics of LLMs. The results of these quantum experiments can then be projected back onto the real LLM embedding system, providing a valuable bridge between the quantum and classical descriptions. This connection offers a promising path for applying quantum computing techniques to analyze, understand, and enhance the capabilities of Large Language Models.

13. Discussion and Conclusion

This article has presented a layered quantum hierarchy for analyzing LLM embedding spaces, bridging the gap between the simplified linearized view and the complex Transformer architecture. By drawing analogies to quantum mechanics, we have introduced concepts such as semantic charge, gauge invariance, and quantum tunneling to model phenomena like the dynamics of Semantic Noise and hallucinations. We demonstrated that a classical LLM embedding system has an exact quantum mechanical analogue, and we presented a simple quantum circuit design for calculating cosine similarity. These findings suggest that quantum methods offer a valuable perspective on LLM representations.

It is essential to remember that the validity of this analogy depends on the extent to which the mathematical structures and relationships captured by the framework correspond to meaningful phenomena in LLM behavior. The simplified models and approximations we have employed inevitably involve a loss of information, and further validations are needed to confirm these connections empirically.

Promising directions for future research include: developing methods for directly measuring "semantic charge" in LLMs; designing experiments to detect quantum tunneling events in semantic transformations; exploring the potential for quantum algorithms to improve LLM training and inference; and investigating the application of these concepts to other areas of AI, such as computer vision and reinforcement learning.

Ultimately, understanding the underlying complexity of LLMs is crucial for unlocking their full potential and ensuring their responsible use. By embracing interdisciplinary approaches and drawing inspiration from diverse fields like quantum mechanics, we can create AI systems that are not only powerful and efficient but also fundamentally interpretable, reliable, and aligned with human values. The quantum framework presented here offers a step in this endeavor, suggesting that the seemingly disparate worlds of quantum physics and AI may be more deeply connected than we previously imagined, particularly in understanding and harnessing the power of Semantic Noise to create more intelligent and adaptable AI systems.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.
2. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *International Conference on Learning Representations*.
3. Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).
4. Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research, 3*(Feb), 1137-1155.
5. Harris, Z. (1954). Distributional Structure. *Word*, 10(2-3), 146-162.
6. Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012, July). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 1201-1211).
7. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Chen, D., Dai, W., Chan, H.S., Madotto, A., Fung, P. (2022). Survey of Hallucinations in Natural Language Generation. *ACM Computing Surveys. 248*, 1-38.
8. Pearl, J., & Mackenzie, D. (2018). The book of why: the new science of cause and effect. *Basic books.*
9. Shannon, C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*(3), 379-423.
10. Cover, T.M. and Thomas, J.A. (2006). Elements of Information Theory, 2nd Edition. *Wiley&Sons*.
11. Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics, 22*(1), 79-86.
12. Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory, 37*(1), 145-151.
13. Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., & Lloyd, S. (2017). *Quantum machine learning. Nature, 549*(7671), 195-202.
14. Arute, F., Arya, K., Babbush, R., Bacon, D., Bardin, J. C., Barends, R., ... & Martinis, J. M. (2019). Quantum supremacy using a programmable superconducting processor. *Nature, 574*(7779), 505-510.
15. Grover, L. K. (1996, July). A fast quantum mechanical algorithm for database search. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing* (pp. 212-219).
16. Shor, P. W. (1994, November). Algorithms for quantum computation: discrete logarithms and factoring. In *Proceedings 35th annual symposium on foundations of computer science* (pp. 124-134). Ieee.
17. Benedetti, M., Lloyd, E., Sack, S., & Fiorentini, M. (2019). Parameterized quantum circuits as machine learning models. *Quantum science and technology, 4*(4), 043001.
18. DiVincenzo, D. P. (2000). The physical implementation of quantum computation. *Fortschritte der Physik: Progress of Physics, 48*(9-11), 771-783.
19. Havlíček, V., Córcoles, A. D., Temme, K., Harrow, A. W., Kandala, A., Chow, J. M., & Gambetta, J. M. (2019). Supervised learning with quantum-enhanced feature spaces. *Nature, 567*(7747), 209-212.
20. Beer, K., Bondarenko, D., Farrelly, T., Osborne, T. J., Salzmann, R., Scheiermann, D., & Wolf, R. (2020). Training deep quantum neural networks. *Nature communications, 11*(1), 808.
21. Nielsen, M. A., & Chuang, I. L. (2010). Quantum computation and quantum information. *Cambridge university press.*
22. Khrennikov, A. (2010). Ubiquitous quantum structure: From Psychology to Finance. *Springer*.
23. Kim, S. H., Mei, J., Girotto, C., Yamada, M., & Roetteler, M. (2025). Quantum Large Language Model Fine-Tuning. a*rXiv preprint arXiv:2504.08732.*
24. Coldenfeld, N. (1994). Lectures on Phase Transitions and The Renormalization Group, *Addison-Wesley*.
25. Huang, K. (1987). Statistical Mechanics, 2nd Edition. *John Wiley & Sons*.
26. Pathria, R. K. (1996). Statistical Mechanics 2nd ed. Butterworth-Heinemann.
27. Egelstaff, P.A. (1994). Liquids: Structure, Dynamics and General Properties. *Taylor & Francis*.
28. Feynman, R.P., & Hibbs, A.R. (1965). Quantum Mechanics and Path Integrals. *McGraw-Hill*.
29. Laine, T. A. (2025). Semantic Wave Functions: Exploring Meaning in Large Language Models Through Quantum Formalism. *OA J Applied Sci Technol, 3*(1), 01-22.
30. Laine, T. A. (2025). The Quantum LLM: Modeling Semantic Spaces with Quantum Principles. *OA J Applied Sci Technol, 3*(2), 01-13.
31. Laine, T. A. (2026). Quantum LLMs Using Quantum Computing to Analyze and Process Semantic Information. *OA J Applied Sci Technol, 4*(1), 01-19.
32. Laine, T. A. (2026). Discrete Semantic States and Hamiltonian Dynamics in LLM Embedding Spaces. *OA J Applied Sci Technol, 4*(1), 01-23.
33. Laine, T. A. (2026). Quantum Computation of Partition Function Similarity for Large Language Models. *OA J Applied Sci Technol, 4*(1), 01-11.

34. Zener, C. (1932). Non-adiabatic crossing of energy levels. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 137*(833), 696-702.
35. Landau, L. (1932). Zur theorie der energieubertragung. II. *Physikalische Zeitschrift der Sowjetunion, 2,* 46.
36. Wittig, C. (2005). The landau− zener formula. *The Journal of Physical Chemistry B, 109*(17), 8428-8430.
37. Yang, M., Ma, M. Q., Li, D., Tsai, Y. H. H., & Salakhutdinov, R. (2020, May). Complex transformer: A framework for modeling complex-valued sequence. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4232-4236). IEEE.
38. Eilers, F., & Jiang, X. (2023, June). Building blocks for a complex-valued transformer architecture. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.
39. Huang, E., Zhang, Z., Xu, T., Xia, C., Hu, K., Yang, Y., ... & Qin, Z. (2025). Holographic Transformers for Complex-Valued Signal Processing: Integrating Phase Interference into Self-Attention. *arXiv preprint arXiv:2509.19331.*