ISSN: 2993-5377

Research Article

Open Access Journal of Applied Science and Technology

Quantum-Native Transformer Architecture: Interference, Partition Functions, and Nonlinear Schrödinger Dynamics

Timo Aukusti Laine*

Financial Physics Lab, Finland

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

Submitted: 2026, May 04; **Accepted**: 2026, Jun 25; **Published**: 2026, Jun 30

Citation: Laine, T. A. (2026). Quantum-Native Transformer Architecture: Interference, Partition Functions, and Nonlinear Schrödinger Dynamics. *OA J Applied Sci Technol, 4*(2), 01-28.

Abstract

This paper introduces the Quantum Semantic Circuit (QSC), a quantum-native transformer architecture designed to bridge the gap between classical transformers and quantum mechanics. We map classical transformer components to their quantum mechanical counterparts, classifying these correspondences as exact identities, derivational approximations, representational equivalences, or functional analogies. Key findings include the exact correspondence between L2-normalized embeddings and quantum states, softmax attention weights and Boltzmann distributions, quantum interference as a central process, and the derivation of the cubic nonlinear Schrödinger equation (NLSE) as the unique minimal norm-preserving nonlinearity suitable for quantum-native activation functions. We demonstrate that classical sinusoidal position encoding and residual connections are leading-order approximations of exact quantum phase rotations and unitary evolution, respectively. We propose that the dimensional expansion within the classical feed-forward network (FFN) is a real-valued shadow of quantum symmetry breaking and restoration, potentially linked to the Higgs mechanism in large models. A quantum simulator pilot on a lexical disambiguation task validates the architectural self-consistency of the QSC and confirms that the NLSE nonlinearity performs measurable computational work beyond linear evolution. This work provides a theoretical foundation for exploring quantum-enhanced language models and identifies key hardware requirements for their physical realization.

1. Introduction

The transformer architecture has revolutionized natural language processing, achieving state-of-the-art results across a wide range of tasks. Its core innovation, the scaled dot-product attention mechanism, computes relationships between tokens using linear projections, a softmax function, and weighted averaging. While transformers have demonstrated remarkable capabilities, they remain fundamentally classical systems processing information as real-valued vectors.

This paper addresses a fundamental question: if the initial and final embedding states of a transformer can be represented as quantum states, what quantum mechanical processes and structures are necessary to maintain a valid Hilbert space throughout the intermediate computations? We hypothesize that classical transformer architectures operate within a limited subspace, or submanifold, of a higher-dimensional quantum-native semantic superspace. The classical transformer is not the full quantum system but its real-valued shadow: the projection of a richer quantum computation onto the subspace accessible to real-valued vectors. This perspective motivates a systematic mapping of each transformer component to its most natural quantum mechanical counterpart.

This paper introduces the Quantum Semantic Circuit (QSC), a quantum-native transformer architecture grounded in this mapping. The correspondences are classified by their logical status. Some are exact mathematical identities: L2-normalized embeddings satisfy the quantum state normalization axiom, softmax attention weights are the Boltzmann distribution by definition, and shifted cosine similarity equals a Born rule probability exactly. Central to the QSC is the use of quantum interference, replacing the classical dot-product with a richer mechanism for computing relationships between token states. One correspondence is derived from axioms: the cubic nonlinear Schrödinger equation is the unique minimal norm-preserving nonlinearity consistent with Hilbert-space structure. Others are

derivational approximations: classical sinusoidal position encoding is the first-order real projection of an exact quantum phase rotation, and the classical residual connection is the leading-order Taylor expansion of exact unitary evolution. The remaining correspondences are representational or functional. The central contribution is to establish and classify these correspondences precisely, distinguishing exact identities from approximations and analogies.

This work does not claim that classical transformers are secretly quantum mechanical. The framework identifies which parts of the transformer are already quantum mechanical in structure, which are classical approximations of quantum operations in the semantic superspace, and which require genuinely new quantum primitives. Beyond the direct correspondences, the quantum formalism offers explanations for empirical design choices in the classical architecture that have no theoretical account within the classical framework itself: the dimensional expansion $d \to 4d \to d$ of the feed-forward network and the asymmetry between its weight matrices can be understood as shadows of quantum symmetry breaking and restoration, with the Higgs mechanism providing a qualitative account of large multi-head models. These are explanations of empirical observations rather than strict predictions, since the classical FFN structure is itself an empirical design choice. The Mexican hat potential is proposed as the natural next approximation beyond the cubic NLSE in a hierarchy of increasingly complex nonlinearities, chosen by the same criterion as the cubic: it is the simplest well-studied potential at the next level of complexity with an extensive existing literature. We validate the framework through a quantum simulator pilot on a lexical disambiguation task, confirming architectural self-consistency and that the NLSE nonlinearity does measurable computational work beyond linear evolution.

2. Background and Related Work

The transformer architecture, introduced by Vaswani et al. [1], has become a dominant paradigm in natural language processing, achieving state-of-the-art results on various tasks, including machine translation [2], text summarization [3], and language modeling [4]. However, the computational cost of training and deploying large transformer models remains a significant challenge, motivating research into more efficient architectures and training techniques [5]. The attention mechanism, a core innovation, allows the model to selectively focus on different parts of the input sequence [2], computing pairwise relationships between token embeddings through linear projections, softmax normalization, and weighted aggregation. The success of transformers has also led to research into more efficient and scalable training methods, including techniques like LoRA [6] and knowledge distillation [7].

The application of quantum mechanics to machine learning and artificial intelligence is a growing field of research, known as quantum machine learning (QML) [8]. Quantum machine learning algorithms, such as quantum support vector machines [9] and quantum neural networks [10], have the potential to offer significant computational advantages compared to their classical counterparts [11]. However, the development of practical quantum machine learning algorithms is still in its early stages, and significant challenges remain in terms of hardware limitations and algorithm design [12]. Despite these challenges, the potential for exponential gains in computational efficiency for certain tasks motivates continued research into new quantum algorithms and hybrid quantum-classical approaches. QML research explores not only improved computational efficiency but also fundamentally different approaches to learning and representation, potentially leading to new insights in both machine learning and quantum physics.

Quantum computing and quantum circuits provide the fundamental building blocks for implementing quantum algorithms [13]. Quantum computers leverage the principles of superposition and entanglement to perform computations in a fundamentally different way than classical computers. Quantum circuits are composed of quantum gates, which are unitary transformations that operate on qubits, the basic unit of quantum information [14]. While current quantum hardware is still limited in terms of qubit count and coherence time, significant progress is being made in the development of more powerful and reliable quantum computers [15]. These advancements are crucial for realizing the potential of quantum algorithms in various fields, including machine learning and natural language processing. The development of fault-tolerant quantum computers will be essential to overcome current limitations and unlock the full computational power of quantum algorithms.

Several recent works have explored the connections between quantum mechanics and natural language processing, giving rise to the field of quantum natural language processing (QNLP) [16]. For example, some researchers have used quantum-inspired models to represent word embeddings and semantic relationships [17–19]. These models often leverage the principles of superposition and entanglement to create richer representations of semantic relationships, sometimes achieving improved performance on tasks such as semantic similarity and text classification. The work by Laine [20–27] has also explored the use of quantum formalism for understanding LLM representations. These works provide a foundation for the present paper, which aims to establish a more direct correspondence between the transformer architecture and quantum mechanics.

This paper builds upon the existing literature by exploring specific correspondences between transformer components and quantum mechanical operations. It categorizes these correspondences based on how closely the classical and quantum elements align. These correspondences are organized into four types. An exact correspondence holds when the classical and quantum objects are mathematically

identical: no approximation, projection, or structural argument is involved. A derivational correspondence holds when the classical operation is the leading-order approximation of the quantum operation, obtained by linearization, real projection, or truncation of an exact quantum expression; the quantum operation is the primary object and the classical operation is its shadow. A representational correspondence holds when a classical object can be mapped to a quantum object preserving the relevant geometry, with the two objects describing the same mathematical content in different languages. A functional correspondence holds when a quantum procedure performs the same architectural role as a classical transformer subroutine, with the internal implementations differing and neither being an approximation of the other. Table 1 summarizes the status of the main components considered here.

Transformer component	Classical form	Quantum equivalent	Status
Token representation	$\mathbf{e} \in \mathbb{R}^d$	$\|\psi\rangle \in \mathbb{C}^{2^n}$	Exact
Semantic similarity	$\mathbf{q} \cdot \mathbf{k}/\sqrt{d}$	$\mathrm{Re}\langle\psi_q \mid \psi_k\rangle/\sqrt{d}$	Exact
Softmax attention weights	$e^{z_j}/\sum_m e^{z_m}$	Boltzmann distribution	Exact
Boltzmann–Born bridge	$e^{-\beta E_j}/Z$	Born rule, $\beta \to 2\beta$	Exact up to rescaling
Position encoding	Sinusoidal addition	$U_{\mathrm{pos}}(j)\|\psi_j\rangle$	Derivational
Residual connection	$\mathbf{x} + f(\mathbf{x})$	$(\mathbb{1} - i\epsilon\hat{H}/\hbar)\|\psi\rangle$	Derivational
Projection W_Q	$W_Q\mathbf{x}$	$U_Q(\boldsymbol{\theta}_Q)\|\psi_x\rangle$	Representational
Projection W_K	$W_K\mathbf{x}$	$U_K(\boldsymbol{\theta}_K)\|\psi_x\rangle$	Representational
Projection W_V	$W_V\mathbf{x}$	$U_V(\boldsymbol{\theta}_V)\|\psi_x\rangle$	Representational
Output layer	Softmax over vocabulary	POVM: $\langle\psi\|E_j\|\psi\rangle$	Representational
Norm constraint	Layer normalization	$\langle\psi \mid \psi\rangle = 1$, automatic	Representational
Attention output	$\sum_j \alpha_j \mathbf{v}_j$	ρ_{out}, dominant eigenvector	Functional (non-isomorphic)
Feed-forward network	$W_2\sigma(W_1\mathbf{x})$, $d \to 4d \to d$	$\mathcal{N}_\gamma(U_{\mathrm{ff}}\|\psi\rangle)$, fixed $\mathbb{C}^{2^n}$	Functional
Activation / nonlinearity	ReLU / GELU / SiLU	Cubic NLSE, γ fixed by medium	Functional
Stable representation	Fixed point of activation	Soliton of NLSE	Functional (theoretical)
FFN symmetry (large)	$W_2 \neq W_1^T$ globally	Higgs mechanism, $\prod_{h=1}^{H} U(1)^{d/H}$	Functional (conjectured)

Table 1: Complete mapping between transformer components and quantum mechanical operations, with correspondence status. *Exact*: mathematical identity, no approximation. *Exact up to rescaling*: identical up to a rank-preserving parameter rescaling. *Derivational*: the classical operation is the leading-order approximation of the quantum operation, obtained by linearization, real projection, or truncation. *Representational*: classical and quantum objects describe the same geometric content in different mathematical languages. *Functional*: classical and quantum procedures serve the same architectural role; neither approximates the other.

3. Quantum Encoding of Token Embeddings

This section establishes the foundation for the quantum-native transformer, outlining how token embeddings can be represented as quantum states. This encoding leverages key results from the structural isomorphism detailed in [27], which demonstrates a deep connection between classical embedding spaces and quantum mechanical systems. The classical embedding vector is first lifted into a general quantum semantic superspace and subsequently mapped to a specific quantum qubit representation for implementation.

3.1. Embeddings as Quantum States

In classical transformers, token embeddings are vectors $e_i \in \mathbb{R}^d$ within an embedding space defined by a map $E : \mathcal{V} \to \mathbb{R}^d$ for a vocabulary $\mathcal{V} = \{v_1, \ldots, v_N\}$. L2-normalized embeddings are constrained to the unit sphere S^{d-1}. The quantum analogue maps tokens to quantum states

$$\mathcal{E} : \mathcal{V} \to \mathcal{H} \quad v_i \mapsto |\psi_i\rangle \in \mathbb{C}^{2^n} \quad \langle\psi_i \mid \psi_i\rangle = 1 \tag{1}$$

where $n = \lceil \log_2 d \rceil$ qubits span a Hilbert space $\mathcal{H} = \mathbb{C}^{2^n}$. The normalization constraint $\langle\psi\,|\psi\rangle = 1$ mirrors the L2-normalization of classical embeddings, defining a point on a unit sphere.

The full quantum embedding incorporates both amplitude and phase information

$$|\psi\rangle = \sum_{k=1}^{2^n} a_k\, e^{i\varphi_k} |k\rangle \quad \sum_{k=1}^{2^n} a_k^2 = 1 \tag{2}$$

where $a_k \geq 0$ are real amplitudes and φ_k are phases. Real-valued embeddings correspond to $\varphi_k = 0$. These phases provide additional degrees of freedom that are crucial for the interference patterns produced by unitary operations.

For GPT-scale models with $d = 12288$, the quantum representation requires $n = 14$ qubits, spanning a $2^{14} = 16384$dimensional Hilbert space. The potential advantage of this representation lies in the inductive bias of parameterized quantum circuits: circuits with $\mathcal{O}(\text{poly}(n))$ parameters can generate states with specific entanglement structures that may capture linguistically relevant geometric relationships more naturally than unconstrained real vectors.

3.2. Key Results from the Structural Isomorphism

The structural isomorphism established in [27] demonstrates a fundamental link between L2-normalized embedding vectors and quantum states. The complete equivalence chain connecting classical, quantum mechanical, and circuit representations is

$$S'_C = \mathbf{a}^T \mathbf{H}' \mathbf{a} = a'^2_1 = |\tilde{\psi}_1|^2 = \langle \Psi(t)|\hat{\mathbf{H}}'|\Psi(t)\rangle = P(|0\cdots 0\rangle) \tag{3}$$

where $S'_C = (1 + \mathbf{a} \cdot \mathbf{b})/2 \in [0, 1]$ is the shifted cosine similarity, $\mathbf{H}'$ is the Gram matrix of the embedding pair, a'_1 is the first component of the embedding vector $\mathbf{a}$ in the eigenbasis of $\mathbf{H}'$ (so that a'^2_1 is the projection onto the dominant eigendirection), and $\tilde{\psi}_1$ is the corresponding quantum amplitude in that basis. The first two equalities are classical, the middle two are quantum mechanical expectation values, and the last is a Born rule measurement probability on a quantum circuit of $\lceil \log_2 d \rceil$ qubits. Each equality is exact with no approximation at any step; the full derivation is given in [27].

This justifies the quantum encoding: token embeddings can be represented as quantum states, and their pairwise similarities are exactly recoverable as Born rule probabilities. Note that the isomorphism connects the squared modulus $|\langle \psi_a | \psi_b \rangle|^2$ to S_C' ; the attention mechanism uses instead $\text{Re}\langle \psi_q | \psi_{k_j} \rangle$, which is the direct quantum generalization of the classical dot product and is discussed in Section 6.

4. Position Encoding as Quantum Phase

Classical transformers use additive sinusoidal position encodings that alter both the magnitude and direction of token embeddings, disrupting the semantic geometry. The correct question is not how to replicate this additive encoding in the quantum framework, but what position encoding should be at the level of abstract structure. Position encoding must accomplish three things:

i. Distinguish tokens at different positions that have identical embeddings.
ii. Preserve the semantic content already encoded in the embedding.
iii. Modify the relational structure between tokens (how they attend to each other) in a position-dependent way.

Requirements (ii) and (iii) are in direct tension in the classical framework: any additive modification $e_j \to e_j + p_j$ simultaneously changes both the token's intrinsic content and its relational geometry. There is no way to satisfy (ii) and (iii) independently in $\mathbb{R}^d$, because content and relational structure are encoded in the same real numbers.

In Hilbert space, this tension dissolves. Amplitudes $|\psi_k|^2$ and phases φ_k are orthogonal degrees of freedom: semantic content is carried by the amplitudes, while relational structure is carried by the phases. Position information therefore belongs naturally in the phases, where it modifies relational structure without affecting semantic content. In the quantum framework, position information is encoded as a unitary phase rotation

$$|\psi_j\rangle \mapsto U_{\text{pos}}(j)|\psi_j\rangle \tag{4}$$

where U_{pos} (j) is a diagonal unitary with entries $e^{i\phi_k^{(j)}}$, shifting the phases $\varphi_k \to \varphi_k + \phi_k^{(j)}$. This operation preserves amplitudes, maintains the Born rule interpretation, modifies interference patterns, and is norm-preserving. The classical sinusoidal encoding uses

$$p_{j,2k} = \sin\left(\frac{j}{10000^{2k/d}}\right) \quad p_{j,2k+1} = \cos\left(\frac{j}{10000^{2k/d}}\right) \tag{5}$$

which are the imaginary and real parts of the complex exponential $e^{ij/10000^{2k/d}}$. The quantum phase encoding uses $e^{i\phi_k^{(j)}}$ directly as a complex number acting on a complex amplitude; the classical encoding is what remains after projecting that object onto the real domain. To first order, the quantum encoding reduces to

$$e^{i\hat{\Phi}_j}|\psi_j\rangle \approx |\psi_j\rangle + i\hat{\Phi}_j|\psi_j\rangle \tag{6}$$

which has the additive structure $|\psi_j\rangle + \delta|\psi_j\rangle$ with $\delta|\psi_j\rangle = i\hat{\Phi}_j|\psi_j\rangle$. The classical additive encoding is therefore the first-order linearization of the exact quantum phase rotation, retaining only the leading term of the Taylor expansion.

The quantum phase correction in dimension k

$$\delta\psi_k = (e^{i\phi_k^{(j)}} - 1)\psi_k \tag{7}$$

scales with the token's own amplitude ψ_k in that dimension. This state-dependence is the correct behavior: position information interacts with semantic content through the token's own amplitude geometry. Two tokens with identical classical cosine similarity but different amplitude distributions across dimensions will produce different attention scores after quantum position encoding.

Relative position between tokens at positions j and k is encoded in the phase difference $\phi^{(k)} - \phi^{(j)}$, which enters the attention score directly

$$\mathrm{Re}\langle U_{\mathrm{pos}}(j)\psi_q \,|\, U_{\mathrm{pos}}(k)\psi_{k_j}\rangle = \mathrm{Re}\langle\psi_q|e^{i(\hat{\Phi}_k - \hat{\Phi}_j)}|\psi_{k_j}\rangle \tag{8}$$

The relative position operator $e^{i(\hat{\Phi}_k - \hat{\Phi}_j)}$ acts multiplicatively on the key state, modifying the interference pattern by the phase difference between the two positions. This is the quantum analog of relative position encoding schemes in the classical transformer literature, arising naturally from the multiplicative group structure of unitary operators.

5. Interference-Based Projection

The learned projection matrices W_Q, W_K, and W_V in standard attention mechanisms can be understood as parameterized unitary operators implemented through quantum interference. This section details how these projections can be realized in a quantum system, highlighting the advantages and differences compared to classical projections. Interference is a central concept, as it provides a fundamentally different way to compute relationships between semantic states compared to classical dot products [25].

5.1. Semantic Superposition, Projection, and the Interference Mechanism

Computing the relationship between two token states, $|\psi_a\rangle$ and $|\psi_b\rangle$, can be viewed as a quantum measurement on their superposition [25]

$$|\Phi\rangle = \alpha|\psi_a\rangle + \beta|\psi_b\rangle \tag{9}$$

where α and β are complex amplitudes satisfying $|\alpha|^2 + |\beta|^2 = 1$. The projection of $|\psi_b\rangle$ onto $|\psi_a\rangle$ then becomes a measurement that determines the relative contribution of each state to the superposition.

Formally, with $P_a = |\psi_a\rangle\langle\psi_a|$ as the projector onto $|\psi_a\rangle$, the expectation value of P_a in the state $|\Phi\rangle$ is

$$\langle\Phi|P_a|\Phi\rangle = |\langle\psi_a \,|\, \Phi\rangle|^2 \tag{10}$$

The inner product $\langle\psi_a \,|\, \psi_b\rangle$ quantifies the overlap between the two semantic states, analogous to cosine similarity. The attention score $E_j = -\mathrm{Re}\langle\psi_q \,|\, \psi_{k_j}\rangle/\sqrt{d}$ can therefore be interpreted as a measurement of semantic overlap between the query state and each key state, implemented through projection. This connects directly to the structural isomorphism of [27], where the Born rule probability $|\langle\psi_a \,|\, \psi_b\rangle|^2$ is equivalent to the shifted cosine similarity.

For a token encoded as a complex quantum state $|\psi_x\rangle = \sum_k x_k e^{i\varphi_k}|k\rangle$, applying a parameterized unitary $U(\theta)$ and projecting onto $|j\rangle$ gives

$$\langle j|U(\boldsymbol{\theta})|\psi_x\rangle = \sum_k U_{jk}(\boldsymbol{\theta})\, x_k\, e^{i\varphi_k} \tag{11}$$

Each output component is a sum of constructive and destructive interference contributions from all input dimensions, weighted by the gate parameters $U_{jk}(\theta)$ and the input phases $e^{i\varphi_k}$. This is formally similar to the classical projection $[W\mathbf{x}]_j = \sum_k W_{jk}x_k$, but the quantum implementation offers key differences: the phases $e^{i\varphi_k}$ provide additional degrees of freedom, and the parameterization imposes a different inductive bias. The attention mechanism in transformers leverages this interference to compute relationships between tokens. The query, key, and value projections can be seen as shaping the interference patterns between token states, allowing the model to selectively attend to relevant information.

5.2. Inductive Bias and Parameter Efficiency

Understanding the differences in inductive bias is crucial for assessing whether quantum circuits can efficiently capture complex linguistic relationships. While classical projection matrices learn patterns entry by entry, quantum circuits leverage trigonometric functions and circuit topology to achieve structured cancellation patterns more easily. The unitary nature of quantum projections provides automatic regularization but defers information selection to the measurement stage. The central question is whether this inductive bias aligns well with the transformations that language models need to learn.

Although Eq. (11) is formally analogous to the classical projection $[W\mathbf{x}]_j = \sum_k W_{jk}x_k$, the two parameterizations impose qualitatively different inductive biases. These differences arise from the parameterization family: the set of transformations easily represented by a given parameterization. While any specific unitary or real matrix can be approximated by sufficiently complex classical or quantum circuits, respectively, the key distinction lies in what each parameterization makes easy to achieve with few parameters, and therefore the prior each imposes on the learned transformation.

5.2.1. Phase Structure and the Geometry of Cancellation

To understand how the quantum parameterization's inductive bias differs from the classical one, consider encoding the signed linear combination $x_1 + x_2 - x_3$ into the amplitude of a designated output basis state.

In a classical projection matrix $W \in \mathbb{R}^{d\times d}$, this pattern must be learned entry by entry: $W_{0,1} = +1$, $W_{0,2} = +1$, $W_{0,3} = -1$, and $W_{0,k} = 0$ for all other k. Every entry is an independent real number, and there is nothing inherent in the parameterization that makes cancellation patterns geometrically natural or easy to find. In the quantum circuit, the input state

$$|\psi_{\text{in}}\rangle = x_1|001\rangle + x_2|010\rangle + x_3|100\rangle \tag{12}$$

uses the qubit ordering $q_0 q_1 q_2$ from most to least significant bit, so x_3 has $|1\rangle$ on q_0, x_2 on q_1, and x_1 on q_2, see Figure 1.

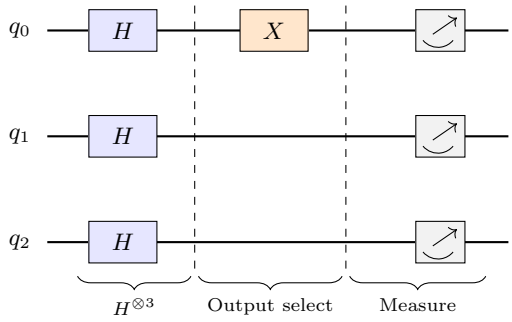


Figure 1: Three-qubit circuit encoding the signed linear combination $x_1 + x_2 - x_3$ into the amplitude of output state $|000\rangle$. The input encodes one amplitude per basis state: x_1 in $|001\rangle$ (qubit q_2 in $|1\rangle$), x_2 in $|010\rangle$ (qubit q_1 in $|1\rangle$), x_3 in $|100\rangle$ (qubit q_0 in $|1\rangle$). The $H^{\otimes 3}$ layer implements the Walsh–Hadamard transform; the sign pattern (+1,+1,−1) appears naturally at output $|100\rangle$ since $b_0 = 1$ produces a sign flip only for x_3, whose $|1\rangle$ is on q_0. The X gate on q_0 (orange) permutes the output basis, moving this amplitude from $|100\rangle$ to $|000\rangle$, giving $P(|000\rangle) = \frac{1}{8}(x_1 + x_2 - x_3)^2$. The cancellation pattern +1,+1,−1 that requires learning three independent matrix entries classically is achieved here by the fixed Hadamard structure combined with a single X gate.

Applying $H^{\otimes 3}$ gives amplitude

$$\langle b_0 b_1 b_2|H^{\otimes 3}|\psi_{\text{in}}\rangle = \frac{1}{2\sqrt{2}} \sum_j x_j \, (-1)^{\mathbf{b}\cdot\mathbf{s}_j} \tag{13}$$

where s_j is the bit string of the j-th input basis state. At output $|100\rangle$: x_1 and x_2 receive sign +1 (their $|1\rangle$ bits do not overlap with $b_0 = 1$), while x_3 receives sign −1 (its $|1\rangle$ is on q_0, overlapping with $b_0 = 1$). Applying X_{q_0} moves this amplitude to $|000\rangle$

$$\langle 000|X_{q_0}H^{\otimes 3}|\psi_{\text{in}}\rangle = \frac{1}{2\sqrt{2}}(x_1 + x_2 - x_3) \tag{14}$$

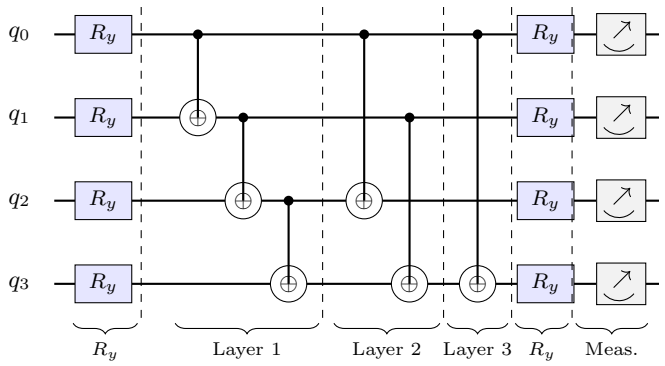
The circuit $U = X_{q_0} \cdot H^{\otimes 3}$ encodes the cancellation pattern using only fixed gates with no continuously tunable parameters. The X gate is unitary $(X^\dagger X = \mathbb{1})$ and serves here as a fixed basis permutation selecting which Walsh Hadamard sign pattern appears at the measurement output, rather than as a learned parameter.

At the amplitude level, the signed combination participates in subsequent quantum interference. At the measurement level, a direct projective measurement returns only $P(|000\rangle) = \frac{1}{8}(x_1+x_2-x_3)^2$. The sign structure is controlled by discrete gate choices (which output state to read), while amplitude modulation for continuously tunable coefficients is controlled by R_y rotation angles. The cancellation pattern +1, +1,−1 that requires learning three independent matrix entries classically is here a consequence of the Walsh–Hadamard structure combined with a single X gate.

5.2.2. Entanglement Topology as a Structured Prior on Correlations

The role of circuit topology is best understood by analogy with convolutional neural networks (CNNs). A fully connected layer connects every input dimension to every output dimension with independent weights, lacking any prior about relevant connections. A CNN, by contrast, connects each output only to a local neighborhood of inputs with shared weights, encoding the prior that nearby pixels are correlated. This prior allows CNNs to learn faster and generalize better than fully connected networks for image data. Similarly, quantum circuits can exploit entanglement locality to encode a structured prior about which qubit pairs carry important correlations.

(a) Linear chain topology



(b) Fully connected topology

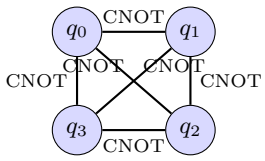


Figure 2: Circuit topology as a structured prior on correlations between semantic dimensions. (a) Linear chain topology on four qubits. Adjacent pairs (q_0, q_1), (q_1, q_2), (q_2, q_3) are correlated by a single CNOT layer (Layer 1: easy). Next-nearest pairs (q_0, q_2), (q_1,q_3) require two CNOT layers (Layer 2: medium cost). The distant pair (q_0, q_3) requires three CNOT layers to bridge the full chain (Layer 3: hard). By choosing the chain topology, the designer encodes the hypothesis that important correlations are between adjacent dimensions. (b) Fully connected topology: every qubit pair has a direct CNOT gate, so all six pairwise correlations are equally easy to generate. This encodes no hypothesis about which correlations matter, but requires more gates, more circuit depth, and more parameters. The topology is a design hypothesis about the data structure, in exactly the same sense as the local receptive field of a convolutional layer.

The CNOT connectivity graph of a quantum circuit acts as this structured prior. When designing a circuit, one chooses which qubits are connected via CNOT gates, determining which correlations are easy to generate (adjacent qubits) and which require more circuit depth (distant qubits), as illustrated in Fig. 2. By choosing a specific topology, one encodes a hypothesis about the important correlations in the data.

To see this concretely, consider 4 qubits in a linear chain. Generating correlations between adjacent pairs requires a single CNOT layer, while distant pairs require multiple layers. This encodes the prior that important correlations are between adjacent dimensions. The topology is therefore a design hypothesis about the structure of the data, analogous to the local receptive field of a CNN. Table 2 summarizes the precise analogy.

	CNN	**Quantum circuit**
Prior	Nearby pixels are correlated	Certain qubit pairs are correlated
Mechanism	Local receptive field	CNOT connectivity graph
Restriction	Translation equivariance	Entanglement locality
Benefit	Fewer parameters, better generalization	Fewer parameters, structured compression
Risk	Wrong if data is not spatially local	Wrong if data correlations do not match topology

Table 2: Analogy between the inductive bias of a convolutional neural network and the entanglement topology of a quantum circuit. Both architectures encode a structured prior over which correlations are important: the CNN through a local receptive field that enforces translation equivariance, and the quantum circuit through a CNOT connectivity graph that enforces entanglement locality. In both cases the prior reduces the number of free parameters and improves generalization when the hypothesis matches the data, but permanently constrains the model when it does not.

A classical projection matrix has no such prior. The quantum circuit, by contrast, encodes a hypothesis about which inter-dimensional correlations are important and learns within that constrained family. If the data has the assumed correlation structure, the circuit learns efficiently. If the hypothesis is wrong, the circuit is permanently constrained.

Whether the correlations that trained transformers learn are compatible with a natural circuit topology is an open empirical question. The low intrinsic rank of weight updates in LoRA [6] and the specialization of attention heads suggest that learned correlations are not uniformly distributed, but neither constitutes direct evidence that the correlation structure matches a specific circuit topology.

The CNOT gate is used throughout as the standard entangling primitive, but any two-qubit entangling gate fulfills the same architectural role. On superconducting hardware, the CZ gate is often native and requires no decomposition. On trapped-ion hardware, the $R_{ZZ}(\theta)$ gate is native and offers the additional advantage of a continuously tunable entanglement strength, making the correlation prior itself a learnable parameter rather than a fixed binary connection. The entanglement topology argument of Fig. 2 applies to any choice of two-qubit entangling gate; what matters is the connectivity graph, not the specific gate implementing each edge.

5.2.3. Unitarity: Norm Preservation with a Cost

The quantum projection is unitary, meaning it preserves the norm of every state. This has two key consequences: an advantage for training stability and a constraint on information flow.

The advantage is automatic regularization of the projection step. Classical projection matrices can in principle have poorly conditioned singular values; a unitary matrix has all singular values equal to 1 by construction, so the projection step cannot introduce exploding or vanishing gradient pathologies. Classical transformers address gradient pathologies through techniques such as gradient clipping and layer normalization, which are applied globally; the quantum projection eliminates this specific source of ill-conditioning at the projection step without requiring additional mechanisms.

The constraint is deferred information selection. The classical projection $W_Q : \mathbb{R}^d \to \mathbb{R}^{d_k}$ projects the input onto a dk-dimensional subspace early in the computation, discarding the complementary directions before attention scores are computed. All d input dimensions contribute to the d_k output dimensions through the matrix multiplication, but the $d - d_k$ discarded output directions cannot influence subsequent computation. The quantum unitary $U_Q : \mathbb{C}^{2^n} \to \mathbb{C}^{2^n}$, by contrast, is invertible and discards no directions, deferring information selection to the final POVM measurement. This is illustrated in Figure 3.

This deferral has consequences in both directions. The potential advantage is that directions appearing irrelevant early may become

relevant after further processing. The potential disadvantage is that early projection acts as a form of regularization, preventing noisy directions from contaminating the attention scores. The quantum architecture carries all directions, including noisy ones, through the entire computation, relying on the POVM measurement to perform the selection at the end.

This is not a technical inconvenience but a fundamental consequence of unitarity. The classical and quantum architectures are solving the same problem with qualitatively different strategies. Whether deferred information selection is advantageous or costly for language modeling is an open question that cannot be resolved by structural arguments alone and is identified as a direction for future empirical work.

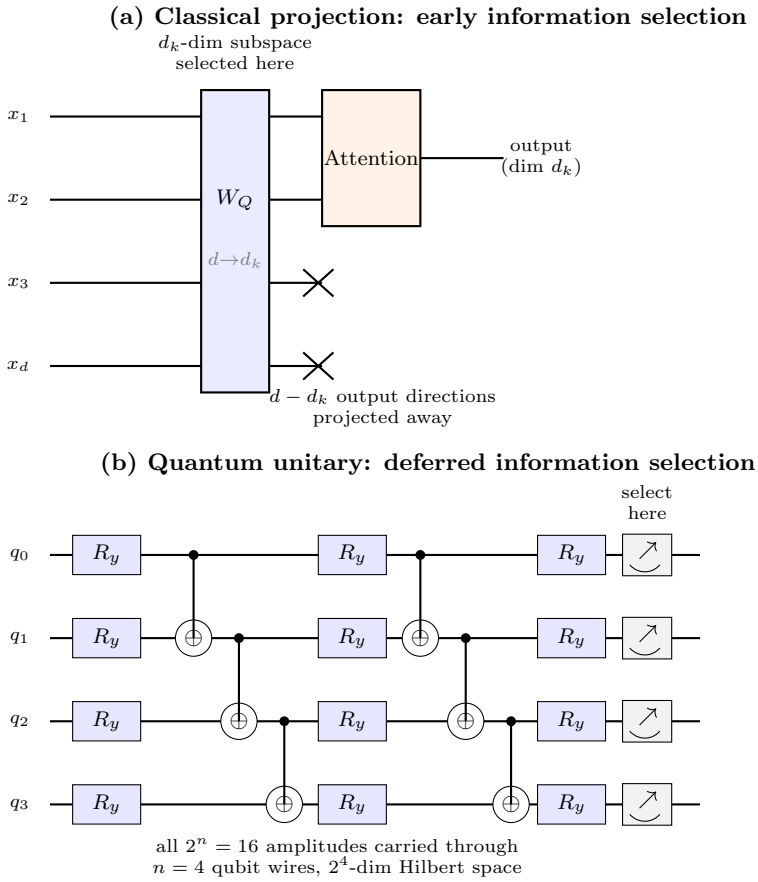


Figure 3: Early versus deferred information selection. (a) Classical projection $W_Q : \mathbb{R}^d \to \mathbb{R}^{d_k}$ with $d_k < d$ projects the input onto a d_k-dimensional subspace, discarding the complementary $d - d_k$ output directions (crossed-out wires on the output side of W_Q). All d input dimensions contribute to the d_k output dimensions through the matrix multiplication; what is lost is the complement of the output subspace, not specific input dimensions. Information selection is early and irreversible: the discarded directions cannot contribute to subsequent computation. (b) Quantum unitary $U_Q : \mathbb{C}^{2^n} \to \mathbb{C}^{2^n}$ is norm-preserving and invertible. The circuit operates on $n = 4$ qubit wires spanning a $2^4 = 16$-dimensional Hilbert space; all 2^n amplitudes are carried through every layer of lossless unitary rotation (R_y gates and CNOT entangling layers), with no directions discarded at any intermediate step. Information selection is deferred to the final POVM measurement, which extracts a probability distribution from the full quantum state. The potential advantage of deferral is that directions appearing irrelevant early may become relevant after further processing. The potential cost is that noisy directions are also carried through, relying on the measurement to sort them out.

5.3. Hadamard Gate and S-Gate Extension: Exploring Real and Imaginary Components

The Hadamard gate provides a simple illustration of the interference mechanism [22]. For a two-component state $|\psi\rangle = \alpha e^{i\varphi}|0\rangle + \beta e^{i\phi}|1\rangle$, the Hadamard gate gives output amplitudes

$$\langle 0|H|\psi\rangle = \tfrac{1}{\sqrt{2}}(\alpha e^{i\varphi} + \beta e^{i\phi}) \tag{15}$$

$$\langle 1|H|\psi\rangle = \tfrac{1}{\sqrt{2}}(\alpha e^{i\varphi} - \beta e^{i\phi}) \tag{16}$$

The real part of the complex inner product is directly readable from the measurement outcome

$$\mathrm{Re}\langle\psi_a \,|\, \psi_b\rangle = 2\alpha\beta\cos(\varphi - \phi) = 2P_H(|0\rangle) - 1 \tag{17}$$

The imaginary part is accessible via a circuit that prepends an S-gate before H. Together, the H and SH circuits recover the full complex inner product. The difference in probabilities measured by these two circuits provides the imaginary component of the inner product. The current Transformer architecture uses only $\mathrm{Re}\langle\psi_q|\psi_{kj}\rangle$ in the attention score, but the imaginary component is accessible and represents a significant architectural extension with no classical analog, potentially enabling the model to capture more nuanced semantic relationships.

5.4. General Parameterized Projection

Any unitary $U \in \mathrm{U}(2^n)$ can be decomposed into a sequence of arbitrary single-qubit unitaries and two-qubit entangling gates. A standard circuit form is

$$U(\boldsymbol{\theta}) = \prod_l \Big(\bigotimes_i u_i(\boldsymbol{\theta}_{li})\Big) \cdot \mathcal{E}_l \tag{18}$$

where each $u_i(\boldsymbol{\theta}_{li}) \in U(2)$ is a parameterized single-qubit unitary and each $\mathcal{E}_l$ is a layer of two-qubit entangling gates such as CNOT, CZ, or $R_{ZZ}(\theta)$. The fixed gates used in the cancellation example, such as H and X, are special cases of single-qubit unitaries and therefore fit naturally within this general decomposition. The main gate types play distinct roles:

- $R_y(\theta)$: amplitude mixing, corresponding to real-valued interference between basis amplitudes
- $R_z(\phi)$: phase shifts, introducing complex relative phases and enabling access to the full complex inner product
- two-qubit entangling gates (e.g. CNOT): coupling between qubits, generating correlations and entanglement between semantic dimensions

For the pilot implementation, we use a restricted two-layer hardware-efficient ansatz: R_y rotations on all qubits, followed by a CNOT entangling layer, followed by another layer of R_y rotations. This ansatz is sufficient to generate a nontrivial family of norm-preserving projections for the simulator study, but it is not universal for arbitrary $U(2^n)$ because R_y and CNOT alone generate only real-valued transformations in the computational basis. A universal ansatz would additionally require phase-generating single-qubit gates such as R_z.

5.5. From Interference to Attention

The interference mechanism connects directly to the attention scoring of Section 6. The semantic energy $E_j = -\mathrm{Re}\langle\psi_q \,|\, \psi_{k_j}\rangle/\sqrt{d}$ depends on the overlap between the query state $|\psi_q\rangle = U_Q|\psi_x\rangle$ and the key state $|\psi_{k_j}\rangle = U_K|\psi_{x_j}\rangle$. This overlap is itself an interference quantity

$$\langle\psi_q \,|\, \psi_{k_j}\rangle = \langle\psi_x|U_Q^\dagger U_K|\psi_{x_j}\rangle = \sum_{k,l} x_k^* e^{-i\varphi_k}[U_Q^\dagger U_K]_{kl}\, x_{j,l}\, e^{i\varphi_{j,l}} \tag{19}$$

The attention mechanism is interference, computed between query and key semantic states through learned projection unitaries.

This implies that the attention score depends on the relative phases, $\varphi_k - \varphi_{jl}$, between query and key tokens. Consider two key tokens with identical amplitude profiles but different phases. In a classical transformer, these tokens produce identical attention scores because cosine similarity depends only on amplitudes. In the quantum architecture, the attention scores differ because the interference pattern depends on the phases. The quantum attention mechanism therefore distinguishes semantic states that are indistinguishable classically, enabling a strictly richer scoring function due to the complex-phase structure of quantum embeddings. The three properties identified above structured phase cancellation, entanglement topology prior, and norm-preserving isometry all contribute to this richer scoring function.

6. Attention as Partition Function

This section establishes two distinct correspondences. The first is a classical mathematical identity: the softmax attention weights are exactly a Boltzmann distribution, a fact that holds independently of any quantum mechanical interpretation. The second is specifically quantum: the Boltzmann weights can be realized as Born rule probabilities on a quantum circuit, up to a rank-preserving rescaling of the inverse temperature [24]. The quantum content of this section lies entirely in the Born rule bridge, not in the Boltzmann identity itself. This distinction motivates the use of density matrices for value aggregation.

6.1. Softmax is a Boltzmann Distribution

The scaled dot-product attention weight for key j given query $\mathbf{q}$ is

$$\alpha_j = \frac{\exp(\mathbf{q} \cdot \mathbf{k}_j / \sqrt{d})}{\sum_m \exp(\mathbf{q} \cdot \mathbf{k}_m / \sqrt{d})} \tag{20}$$

Defining the semantic energy as $E_j = -\frac{\mathbf{q} \cdot \mathbf{k}_j}{\sqrt{d}}$, this becomes

$$\alpha_j = \frac{e^{-\beta E_j}}{Z} \quad Z = \sum_m e^{-\beta E_m} \quad \beta = 1 \tag{21}$$

This is identically the Boltzmann distribution from statistical mechanics, with $\beta = 1$ as the inverse temperature and Z as the partition function. The attention score uses $\mathrm{Re}\langle\psi_q|\psi_{kj}\rangle$ rather than $|\langle\psi_q|\psi_{kj}\rangle|^2$: the real part reduces exactly to the classical dot product $\mathbf{q} \bullet \mathbf{k}$ for real-valued states and can be negative, preserving the full range of the classical attention score; the squared modulus is always non-negative and appears separately in the structural isomorphism of Section 3.

6.2. The Born Rule Bridge

To connect the Boltzmann distribution to the Born rule, we require a single-qubit state whose measurement probability equals the Boltzmann weight. This is achieved by setting the $|0\rangle$ amplitude equal to the shifted Boltzmann weight $e^{-\beta \tilde{E}_j}$

$$|\phi_j\rangle = e^{-\beta \tilde{E}_j}|0\rangle + \sqrt{1 - e^{-2\beta \tilde{E}_j}}|1\rangle \tag{22}$$

where $\tilde{E}_j = E_j + C$ is a shifted energy [26]. The constant C is added to ensure that all energies are positive, which is required for the Born rule probabilities to be well-defined. The Born rule gives

$$P_j(|0\rangle) = e^{-2\beta \tilde{E}_j} \tag{23}$$

Normalizing over all keys

$$\tilde{\alpha}_j = \frac{P_j(|0\rangle)}{\sum_m P_m(|0\rangle)} = \frac{e^{-2\beta \tilde{E}_j}}{\sum_m e^{-2\beta \tilde{E}_m}} \tag{24}$$

This is the Boltzmann distribution with $\beta \to 2\beta$. The complete bridge is

$$\underbrace{\frac{e^{z_j}}{\sum_m e^{z_m}}}_{\text{Softmax}} \xrightarrow{z_j = -\beta E_j,\ \text{exact}} \underbrace{\frac{e^{-\beta E_j}}{Z}}_{\text{Boltzmann}} \xrightarrow{\beta \to 2\beta,\ \text{rank-preserving}} \underbrace{\frac{e^{-2\beta \tilde{E}_j}}{\sum_m e^{-2\beta \tilde{E}_m}}}_{\text{Born rule}} \tag{25}$$

The first arrow is exact and definitional. The second arrow introduces a factor of 2 in β, which is compensated by choosing $\beta_{\text{quantum}} = \beta_{\text{classical}}/2$. The rescaling preserves the rank ordering of attention weights. The rescaling $\beta \to 2\beta$ introduces a quantitative mismatch between the classical and quantum systems at fixed β, requiring an adjustment to compute the same function.

6.3. Complete Quantum Attention

The full quantum attention mechanism proceeds in four steps: project via interference

$$|\psi_q\rangle = U_Q|\psi_x\rangle \quad |\psi_{k_j}\rangle = U_K|\psi_{x_j}\rangle \quad |\psi_{v_j}\rangle = U_V|\psi_{x_j}\rangle \tag{26}$$

score via partition function

$$E_j = -\mathrm{Re}\langle\psi_q \,|\, \psi_{k_j}\rangle/\sqrt{d} \quad \alpha_j \propto e^{-\beta E_j} \tag{27}$$

aggregate via density matrix

$$\rho_{\mathrm{out}} = \sum_j \alpha_j |\psi_{v_j}\rangle\langle\psi_{v_j}| \tag{28}$$

and extract the dominant eigenvector of ρ_{out}. The scoring step uses m qubits and m single-qubit R_y gates with no entanglement required; the attention distribution is encoded as a product state

$$|\Psi_{\mathrm{attn}}\rangle = \bigotimes_{j=1}^{m} \left(e^{-\beta \tilde{E}_j}|0\rangle_j + \sqrt{1 - e^{-2\beta \tilde{E}_j}}|1\rangle_j \right) \tag{29}$$

The density matrix formulation in step 3 is the quantum analog of the classical weighted average $\sum_j \alpha_j \mathbf{v}_j$. Both form a convex combination of value representations with the same weights $\{\alpha_j\}$, but the classical construction produces a vector retaining only the first moment while the quantum construction produces a density matrix retaining the full second-order structure: the individual value states $|\psi_{v_j}\rangle$, the variance structure of the attended values, and the coherence between value states. Neither approximates the other: the quantum construction is strictly richer.

The eigenvalue spectrum of ρ_{out} has a direct interpretation in terms of the structural isomorphism of [27]. The dominant eigenvalue λ_{max} corresponds to the similarity-aligned component of the aggregated value state, analogous to $|\tilde{\psi}_1|^2 = S'_C$ in the isomorphism. The subdominant eigenvalues λ_2, ..., λ_{2^n} constitute the semantic noise profile of the attention output. The ratio λ_{max}/λ_2 therefore measures the signal-to-noise ratio of the attention output and provides a principled criterion for the validity of the rank-1 approximation.

Extracting the dominant eigenvector $|\psi_{max}\rangle$ is the leading-order truncation of the mixed-state density matrix to a pure state, accurate when $\lambda_{max} \gg \lambda_2$ with approximation error $O(\lambda_2/\lambda_{max})$. It is not an approximation to the classical weighted average but a qualitatively different operation: the principal semantic direction of the attended value mixture. Crucially, the QSC does not compute the classical weighted average and then approximate it. Instead, it computes a strictly richer object the density matrix which retains information about the individual value states, their variance, and the coherence between them. The dominant eigenvector is then a summary statistic extracted from this richer representation, discarding information present in the full density matrix. The two operations agree only in the sharp-attention limit $\alpha_{j^*} \approx 1$; in the uniform-attention case $\alpha_j = 1/m$, the dominant eigenvector is the leading principal component of the value states, which can differ substantially from the equal-weight average. Whether the principal-component summary is more useful for language modeling than the centroid summary of the classical transformer is an empirical question.

To make the distinction concrete, consider two value states $|v_1\rangle$ and $|v_2\rangle$ with equal attention weights $\alpha_1 = \alpha_2 = 1/2$. The classical centroid $\mathbf{(v_1 + v_2)/2}$ lies midway between the two states and may not resemble either. The dominant eigenvector of $\rho = \frac{1}{2}(|v_1\rangle\langle v_1| + |v_2\rangle\langle v_2|)$ is the direction of maximum variance in the mixture, which aligns with the more semantically coherent of the two value states when they are not orthogonal. If $|v_1\rangle$ and $|v_2\rangle$ represent competing semantic interpretations of an ambiguous token, the centroid blends them into an incoherent average, while the dominant eigenvector selects the stronger interpretation. This is the sense in which the quantum aggregation is not merely a different implementation but a qualitatively different operation: it performs implicit disambiguation rather than averaging. The correspondence is therefore classified as functional but non-isomorphic: the quantum and classical procedures serve the same architectural role but produce different outputs in the general case.

In the degenerate case $\lambda_{max} \approx \lambda_2$, which occurs when attention weights are nearly uniform, the correct treatment is to propagate the full mixed state ρ_{out} through subsequent layers using the nonlinear von Neumann equation (40), deferring collapse to the final POVM measurement. This is the theoretically complete formulation. The pilot uses the dominant eigenvector extraction as a practical approximation.

7. The NLSE as Quantum Activation Function

The sections from here through Section 9 are ordered by theoretical dependency rather than by position in the layer stack. The NLSE section precedes the FFN section because the FFN analysis depends on NLSE machinery derived here.

In classical transformer architectures, activation functions such as ReLU and GELU serve one essential architectural purpose: they introduce nonlinearity into the computation. Without nonlinearity, the entire network collapses to a single linear map regardless of depth. As seen in classical LLM experiments, the specific functional form of the activation is secondary to this architectural role.

This motivates the question: what is the minimal way to introduce nonlinearity into a quantum mechanical system while preserving the Hilbert-space structure established by the encoding? Intermediate semantic processing must preserve Hilbert-space structure, ruling out classical activation functions applied pointwise to amplitude vectors, since such operations do not respect the normalization constraint $\langle\psi \mid \psi\rangle = 1$ and break the Born rule interpretation of squared amplitudes as probabilities.

7.1. Derivation of the Cubic NLSE

The cubic NLSE can be derived from three requirements:

• Norm preservation: The generator must be Hermitian to ensure $\langle\psi(t) \mid \psi(t)\rangle = 1$ for all t. This requires $\hat{F}[\psi]^\dagger = \hat{F}[\psi]$ for all states $|\psi\rangle$ and rules out pointwise classical activations.

• Density locality: The nonlinear part of the generator depends on the state only through the local density $\rho(x) = |\psi(x)|^2$. This ensures that the nonlinearity is determined by the probability distribution of the quantum state, preserving Born rule interpretability and global phase invariance.

• Minimality: The density-dependent potential $V(|\psi|^2)$ is of lowest nontrivial order, i.e., linear in $|\psi|^2$.

Consider the most general nonlinear generalization of the Schrödinger equation

$$i\hbar\frac{d}{dt}|\psi(t)\rangle = \hat{F}[|\psi\rangle]|\psi(t)\rangle \tag{30}$$

where $\hat{F}[|\psi\rangle]$ is an operator that may depend on the current state. The time derivative of the norm is

$$\frac{d}{dt}\langle\psi \mid \psi\rangle = \frac{1}{i\hbar}\langle\psi|\big(\hat{F}[\psi] - \hat{F}[\psi]^\dagger\big)|\psi\rangle \tag{31}$$

This vanishes for all $|\psi\rangle$ if and only if $\hat{F}[|\psi\rangle]$ is Hermitian for all $|\psi\rangle$

$$\hat{F}[\psi]^\dagger = \hat{F}[\psi] \quad \forall\,|\psi\rangle \tag{32}$$

The most general Hermitian operator that depends on the state only through $|\psi|^2$ has the form

$$\hat{F}[\psi] = \hat{H}_0 + V\big(|\psi|^2\big) \tag{33}$$

where $\hat{H}_0$ is a fixed Hermitian operator (the free Hamiltonian) and $V(|\psi|^2)$ is a real-valued function of the local density, acting as a state-dependent potential.

Among all real-valued functions $V(|\psi|^2)$, the lowest-order nonlinear choice is the linear function

$$V\big(|\psi|^2\big) = -\gamma|\psi|^2 \quad \gamma \in \mathbb{R} \tag{34}$$

The constant term (V = const) contributes only a global phase and is physically irrelevant. The linear term $V = -\gamma|\psi|^2$ is therefore the minimal nontrivial density-dependent potential. Substituting into Eq. (30) gives

$$i\hbar\frac{d}{dt}|\psi(t)\rangle = \left(\hat{H}_0 - \gamma|\psi|^2\right)|\psi(t)\rangle \tag{35}$$

which is the cubic nonlinear Schrödinger equation [20, 25]. The term is "cubic" because the nonlinear part of the equation of motion is $-\gamma|\psi|^2\psi$, which is third-order with respect to ψ.

Therefore, the cubic NLSE is the unique equation of the form $i\hbar\frac{d}{dt}|\psi(t)\rangle = \hat{F}[|\psi\rangle]|\psi(t)\rangle$ satisfying norm preservation, density locality, and minimality. Any norm-preserving, density-local nonlinearity of higher order is a valid extension but not the minimal one.

7.2. Properties of the Cubic NLSE

For $\gamma > 0$, the NLSE nonlinear term $-\gamma|\psi|^2\psi$ acts as self-reinforcement: components with high probability density $|\psi_k|^2$ accumulate phase faster, which through interference with the linear evolution $\hat{H}_0$ amplifies the dominant components relative to subdominant ones. For $\gamma < 0$, the effect is reversed: dominant components are suppressed, producing a more uniform distribution. In the pilot with $H_0 = 0$, the effect is pure phase rotation with no amplitude change; the sharpening arises from the interaction of the phase-rotated NLSE output with the residual connection.

The parameter $\hbar$ in the NLSE (Eq. (35)) is a scaling constant that sets the ratio $\theta_s = t / \hbar$ governing phase evolution. In the LLM embedding context, $\hbar$ admits a concrete operational interpretation as the semantic resolution scale [27]

$$\hbar_{\text{sem}} = \min_{i \neq j} \arccos(\mathbf{a}_i \cdot \mathbf{a}_j) \tag{36}$$

the minimum angular separation between any two distinct token embeddings in the model's vocabulary. Setting $\hbar = \hbar_{\text{sem}}$ provides a canonical gauge choice in which the semantic phase angle $\theta_s = t/\hbar_{\text{sem}}$ is a dimensionless count of minimum semantic steps. The interpretation is mathematical, not physical: $\hbar_{\text{sem}}$ plays the same role in the geometry of the embedding space as $\hbar$ plays in the quantum formalism, and the correspondence is exact at the level of mathematical form.

7.3. Alternative Nonlinearities and Semantic Bias Selection

The cubic NLSE is the canonical starting point and the minimal norm-preserving nonlinearity consistent with Hilbert-space structure. Its value is twofold: it is the simplest possible nonlinearity by construction, and it has an extensive theoretical literature with well-understood properties including soliton solutions, integrability in one dimension, and stability analysis. These properties make it the natural first choice for theoretical work, not because it is necessarily the correct model for any specific regime, but because it is the tractable one from which rigorous results can be derived.

The principal nonlinearity families, ordered by complexity, are listed below. The cubic and Mexican hat cases are discussed in detail in the following subsections; the remaining two are included for completeness.

Let $F[\psi]$ denote the nonlinear term in

$$i\hbar\frac{d}{dt}|\psi\rangle = \hat{H}_0|\psi\rangle + F[\psi] \tag{37}$$

• Cubic (canonical): $F[\psi] = -\gamma|\psi|^2\psi$. The minimal nonlinearity, derived from axioms. Self-reinforcement ($\gamma > 0$) or self-inhibition ($\gamma < 0$). Extensive theoretical results available.
• General polynomial: $F[\psi] = \sum_k g_k|\psi|^{2k}\psi$. Higher-order density dependence; reduces to cubic for $k = 1$ only. Less tractable analytically.
• Saturable: $F[\psi] = g\psi/(1 + |\psi|^2)$. Bounded nonlinearity; models saturation of semantic activation at high density. Well-studied in nonlinear optics.
• Symmetry-breaking (Mexican hat): Derived from a quartic potential $V(|\psi|^2)$ with degenerate minima at nonzero $|\psi|$. The simplest potential that introduces spontaneous symmetry breaking, with an extensive literature from field theory and condensed matter physics.

7.3.1. The Mexican Hat Potential

The Mexican hat potential is the natural next step beyond the cubic: it is the simplest well-studied potential that introduces a qualitatively new feature absent from the cubic, namely spontaneous symmetry breaking. A particularly structured example is obtained from the quartic potential [20]

$$V(\psi) = -\mu^2|\psi|^2 + \lambda|\psi|^4 \qquad \mu^2, \lambda > 0 \tag{38}$$

The minima of V occur at nonzero $|\psi|$, so the zero state is unstable and the system spontaneously evolves toward a nonzero amplitude configuration on the circle $|\psi|^2 = \mu^2/2\lambda$. The cubic nonlinearity has no such feature: it modifies the dynamics around the zero state but does not destabilize it.

In the semantic interpretation, this spontaneous evolution corresponds to semantic bias selection: even though the underlying dynamics are symmetric under phase rotation $\psi \rightarrow e^{i\theta}\psi$, the system settles into one effective semantic orientation, analogous to how a language model commits to a specific semantic direction through the dynamics of processing [20].

The reason to study this potential is that the Higgs mechanism, Goldstone boson absorption, and spontaneous symmetry breaking are well-understood consequences of it, and these theoretical results can be imported directly into the transformer context. The Mexican hat potential is the entry point to this body of theory, just as the cubic NLSE is the entry point to soliton theory and integrability results. The consequences for the relationship between W_1 and W_2 are developed in Section 9.1.3.

7.3.2. Solitons as Stable Semantic Representations

A stable solution of the NLSE in a self-reinforcing mode is a soliton [25]. In this state, the nonlinear self-phase modulation exactly balances dispersion, maintaining the shape of the state indefinitely. This is the functional analog of a fixed point of a classical activation function, where repeated application of the nonlinearity leaves the representation unchanged. The soliton is one of the theoretical results that the cubic nonlinearity makes available; the Mexican hat potential brings a different class of stable solutions, namely the minima of the potential, which correspond to semantically committed states rather than propagating wavepackets.

7.3.3. Hardware Implications

The NLSE cannot be implemented on current gate-based quantum hardware, which supports only unitary (linear) operations. This is a fundamental constraint, not an engineering limitation, and it constitutes a concrete hardware requirement implied by the quantum-native transformer framework. Any quantum system intended to replicate transformer-class computation must support nonlinear dynamics. Possible paths to physical realization include: (i) approximating the NLSE nonlinearity through ancilla-based measurement-and-feedback schemes; (ii) using photonic quantum hardware, which naturally supports nonlinear optical interactions (Kerr effect); (iii) developing LLM-specific quantum hardware designed from the outset to support nonlinear dynamics. The Mexican hat nonlinearity of the large-model regime has an additional natural physical realization in systems with double-well potentials, which are wellstudied in nonlinear optics and superconducting circuit QED, providing a concrete hardware path for implementing the symmetry-enhanced regime independently of the minimal cubic case. These are identified as separate research questions.

The present paper validates the NLSE step in simulation. Quantum simulators are invaluable tools for investigating QSC circuits, allowing for the straightforward implementation of nonlinearities and the generation of noise-free results. The utility of these simulated results then motivates the development of corresponding real quantum hardware capable of supporting such nonlinear dynamics.

7.3.4. Extension to Mixed States: The Nonlinear von Neumann Equation

The cubic NLSE is formulated for pure states $|\psi(t)\rangle \in \mathcal{H}$. However, the value aggregation step of the QSC produces a mixed state $\rho_{\text{out}} = \sum_j \alpha_j |\psi_{v_j}\rangle\langle\psi_{v_j}|$, which is not a pure state unless one attention weight dominates. Applying the pure-state NLSE to the dominant eigenvector of ρout is therefore an approximation. For completeness, we derive the correct generalization of the NLSE to density matrices, which is the equation that a fully quantum treatment of the QSC would use.

The von Neumann equation governs the linear evolution of a density matrix ρ

$$i\hbar\frac{d\rho}{dt} = [\hat{H}_0, \rho] \tag{39}$$

The term $[\hat{H}_0, \rho]$ describes the unitary evolution of the density matrix due to the free Hamiltonian. To introduce the cubic nonlinearity consistently, we replace $\hat{H}_0$ with the state-dependent generator $\hat{F}[\rho] = \hat{H}_0 - \gamma\,\mathrm{diag}(\rho)$, where diag(ρ) denotes the diagonal matrix operator with entries $[\mathrm{diag}(\rho)]_{kk} = \rho_{kk} = \langle k|\rho|k\rangle$ and all off-diagonal entries zero. This is the natural density-matrix analog of the pure-state potential $-\gamma|\psi|^2$: in the pure state $\rho = |\psi\rangle\langle\psi|$, the diagonal entries are $\rho_{kk} = |\psi_k|^2$, recovering the cubic NLSE exactly. The resulting equation is

$$i\hbar\frac{d\rho}{dt} = [\hat{H}_0, \rho] - \gamma[\mathrm{diag}(\rho),\ \rho] \tag{40}$$

This is the nonlinear von Neumann equation (NLvN). It reduces to the cubic NLSE when $\rho = |\psi\rangle\langle\psi|$ is a rank-1 projector, since in that case $[\mathrm{diag}(\rho), \rho]_{jk} = (\rho_{jj} - \rho_{kk})\rho_{jk}$, which reproduces the action of $-\gamma|\psi|^2$ on the off-diagonal coherences.

The NLvN equation preserves $\mathrm{tr}(\rho) = 1$ for any Hermitian $\hat{H}_0$ and real γ. In the mixed-state setting, the diagonal entries ρ_{kk} represent the probability of finding the semantic state in basis direction $|k\rangle$. The nonlinear term $-\gamma[\mathrm{diag}(\rho), \rho]$ modifies the off-diagonal coherences ρ_{jk} in proportion to the difference in occupation probabilities $\rho_{jj} - \rho_{kk}$: coherences between highly occupied directions are reinforced ($\gamma > 0$), while coherences between directions with unequal occupation are suppressed. This is the mixed-state analog of the pure-state self-reinforcement: the nonlinearity sharpens the semantic representation by amplifying the dominant directions of ρ at the expense of the subdominant ones.

The dominant eigenvector approximation corresponds to replacing ρ with its rank-1 approximation $\rho \approx \lambda_{\max}|v_{\max}\rangle\langle v_{\max}|$, and $|v_{\max}\rangle$ are the largest eigenvalue and corresponding eigenvector of ρ_{out}. Under this approximation, the NLvN equation reduces exactly to the pure-state cubic NLSE applied to $|v_{\max}\rangle$. The approximation is therefore not an ad hoc choice but the leading-order truncation of the full mixed-state dynamics. It is accurate when $\lambda_{\max} \gg \lambda_2$ (the second-largest eigenvalue of ρ_{out}), and degrades when the spectrum of ρ_{out} is nearly degenerate.

8. Residual Connections as Hamiltonian Evolution

The residual connection $\mathbf{x} + f(\mathbf{x})$ in the classical transformer corresponds to a small step of Hamiltonian evolution

$$|\psi_{l+1}\rangle \approx \left(\mathbb{1} - \frac{i\hat{H}_{\text{evol}}\epsilon}{\hbar}\right)|\psi_l\rangle \tag{41}$$

where $\hat{H}_{\text{evol}}$ is the evolution Hamiltonian and ϵ is the step size. This is the leading-order Taylor expansion of the exact unitary evolution $e^{-i\hat{H}_{\text{evol}}\epsilon/\hbar}$. This Hamiltonian evolution, combined with the NLSE nonlinearity from the previous section, governs the dynamics of the quantum state as it propagates through the QSC. The classical transformer has two residual connections per layer: one following the attention sublayer and one following the feed-forward sublayer. In the QSC, these are consolidated into a single residual connection placed after the NLSE activation step, as defined in Section 11. This consolidation is natural in the quantum framework: the FFN unitary and NLSE activation form a single continuous evolution segment rather than two separate discrete sublayers (Section 9), so one residual connection at the end of that segment is the correct quantum analog of the two classical residual connections taken together.

Each transformer layer corresponds to one time step ϵ of Hamiltonian evolution. A transformer with L layers corresponds to an evolution time $T = L\epsilon$. The transformer depth is the discretization of a continuous quantum dynamical process.

The step size ϵ in the residual connection Eq. (41) is a gauge parameter in the sense established in [27]: only the product $\hat{H}_{\text{evol}}\epsilon/\hbar$ is physically meaningful, not ϵ or $\hbar$ separately. This means that we can rescale ϵ and $\hat{H}_{\text{evol}}$ with out changing the observable behavior of the system. Rescaling $\epsilon \to \alpha\epsilon$ and $\hat{H}_{\text{evol}} \to \hat{H}_{\text{evol}}/\alpha$ leaves the unitary evolution $e^{-i\hat{H}_{\text{evol}}\epsilon/\hbar}$ unchanged and therefore leaves all observable predictions unchanged. The Layer–Time Correspondence identifies each transformer layer with one time step ϵ, but the absolute value of ϵ is not observable: what matters is the accumulated product $\hat{H}_{\text{evol}}$. $L\epsilon$ / $\hbar$ over all L layers, which determines the total unitary transformation applied to the semantic state. With the canonical gauge choice $\hbar = \hbar_{\text{sem}}$, the step size ϵ is measured in units of the minimum semantic step, giving it a concrete geometric interpretation: $\epsilon/\hbar_{\text{sem}}$ is the number of minimum semantic steps per transformer layer.

9. Feed-Forward Network as Continuous Nonlinear Evolution

This section explores the feed-forward network (FFN) within the transformer architecture, contrasting the classical and quantum approaches. We propose that the dimensional expansion ($d \to 4d \to d$) in classical FFNs arises from the limitations of real-valued computation, specifically the absence of phase information. The quantum FFN, operating in the quantum semantic superspace, leverages both amplitude and phase degrees of freedom, enabling complex nonlinear transformations without dimensional expansion.

9.1. The Classical FFN and Its Dimensional Structure

The classical FFN applies a two-stage transformation: $\mathrm{FFN}(x) = W_2\,\sigma(W_1 x + b_1) + b_2$, where W_1 expands the representation to a higher-dimensional space ($d_{\text{ff}} = 4d$), σ is a pointwise nonlinearity, and W_2 projects back to the model dimension. The expansion factor, often 4, is empirically determined and lacks a theoretical derivation within the classical framework. The quantum framework provides a consistency argument for this factor.

In the quantum semantic superspace, a quantum state $|\psi\rangle \in \mathbb{C}^{2^n}$ carries both amplitude and phase information, degrees of freedom absent in a real vector $x \in \mathbb{R}^d$. The classical FFN's dimensional expansion can be understood as a consequence of the real restriction of a quantum computation that naturally operates on complex states.

9.1.1. Local $U(1)^d$ Symmetry and the Consistency Argument for the Dimensional Expansion

The companion paper [27] establishes that the quantum formulation admits a local $U(1)$ gauge symmetry in the eigenbasis. This symmetry, invisible classically, removes redundant phase degrees of freedom, ensuring that the quantum and classical representations have the same physical content. The cubic NLSE nonlinearity breaks this local $U(1)^d$ symmetry, inducing a component-dependent phase rotation that allows the relative phases between components to carry new physical information.

The dimensional expansion factor of 4 is consistent with two factors of 2 arising from the quantum framework: (i) the real restriction of the unitary, requiring $2d$ real numbers to represent a d-dimensional complex space, and (ii) the breaking of the $U(1)^d$ symmetry by the nonlinearity, requiring approximately $2d$ real dimensions to represent the new phase information. This gives the expansion and contraction cycle

$$d \xrightarrow{\text{real restriction}} 2d \xrightarrow{U(1)^d \text{ breaking}} 4d \xrightarrow{U(1)^d \text{ restored}} d \tag{42}$$

The dimensional expansion is therefore not an arbitrary empirical choice but the natural intermediate dimension of the real-valued shadow of a quantum computation involving symmetry breaking and restoration.

9.1.2. The Quantum FFN as Continuous Evolution

In the fully quantum-native architecture, the quantum state flows continuously through all L layers, with the single POVM measurement occurring only at the output. The FFN is therefore not a discrete three-stage operation but a segment of continuous nonlinear quantum evolution, governed by

$$i\hbar \frac{d}{dt}|\psi(t)\rangle = \left(\hat{H}_{\text{evol}}(t) - \gamma|\psi(t)|^2\right)|\psi(t)\rangle \tag{43}$$

where $\hat{H}_{\text{evol}}(t)$ encodes the unitary operations (U_Q, U_K, U_V , U_{ff}) as time-dependent Hamiltonian pulses. The FFN segment of this evolution is

$$|\psi\rangle \xrightarrow{U_{\text{ff}}} U_{\text{ff}}|\psi\rangle \xrightarrow{\mathcal{N}_\gamma} \mathcal{N}_\gamma(U_{\text{ff}}|\psi\rangle) = |\psi'\rangle \tag{44}$$

where U_{ff} rotates the quantum state, preparing it for the subsequent nonlinear evolution. The output $|\psi'\rangle \in \mathcal{H}$ feeds directly into the next layer with no dimensional change, intermediate measurement, or W_2 contraction. The function of W_2 in the classical architecture (aggregating activated features) is distributed across subsequent unitary operations and the final POVM measurement.

The quantum framework provides an explanation for the structure of the FFN weight matrices. In the minimal single-head regime, W_2 implements symmetry restoration, which is the approximate inverse of the W_1 expansion, suggesting $W_2 \approx W_1^T$ as a consistency indicator. The approximation error

$$\epsilon_W = \frac{\|W_2 - W_1^T\|_F}{\|W_1\|_F} \tag{45}$$

measures how well this explanation fits trained models. A small ϵ_W in minimal models would be consistent with the quantum explanation; a large ϵ_W would indicate that the explanation is incomplete or that the minimal-regime assumption does not apply, but would not by itself falsify the quantum framework, since the classical FFN structure is itself an empirical design choice rather than a derived necessity.

9.1.3. Symmetry Enhancement in Large Models and the Higgs Analogy

In large multi-head models the relevant symmetry structure is richer than in the minimal single-head regime. With H attention heads, each operating on a subspace of dimension d/H, the global $U(1)^d$ phase freedom decomposes into H independent local phase freedoms, promoting the global symmetry group to a local gauge symmetry

$$U(1)^d \longrightarrow \prod_{h=1}^{H} U(1)^{d/H} \tag{46}$$

When the nonlinearity is of Mexican hat type, the Higgs mechanism operates in this enlarged symmetry group. Each head independently selects a minimum of the potential, breaking its local phase freedom. By the Goldstone theorem, each broken generator produces a massless Goldstone mode, which in the transformer analog is absorbed into W_2, giving it additional structure beyond the simple transpose

$$W_2 = W_1^T + \Delta W_{\text{Goldstone}} \tag{47}$$

Since each of the H heads independently breaks one local $U(1)$ symmetry, the number of broken generators equals H, giving

$$\text{rank}(\Delta W_{\text{Goldstone}}) \sim H \tag{48}$$

This is a consistency indicator rather than a strict prediction: if the quantum explanation is correct, the blockwise transpose relation should hold within each head's subspace

$$\|P_h W_2 P_h - (P_h W_1 P_h)^T\|_F \ll \|W_2 - W_1^T\|_F \tag{49}$$

even when the global relation $W_2 \approx W_1^T$ fails.

Several empirical phenomena in large LLMs are qualitatively consistent with this picture: the tendency of large models to commit to a single semantic interpretation rather than maintaining superpositions, the sharpening of polysemy resolution with model scale, and the spontaneous specialization of attention heads to distinct functional roles. Each corresponds naturally to a feature of Mexican hat symmetry breaking, and none has a comparably natural explanation within the classical transformer framework.

The full derivation of Eqs. (47) and (48) from first principles, including the precise identification of the Goldstone modes and the mechanism of their absorption into W_2, is identified as a separate research question.

9.1.4. Bias Terms as the Free Hamiltonian

Returning to the minimal-regime quantum FFN, the bias terms $\mathbf{b}_1$ and $\mathbf{b}_2$ of the classical FFN have a natural quantum counterpart in the free Hamiltonian $\hat{H}_0$ of the NLSE (Eq. (35)). A constant bias $\mathbf{b}$ shifts all components of the representation uniformly; the free Hamiltonian $\hat{H}_0$ plays the same architectural role, providing a state-independent baseline evolution that is present regardless of the input. In the pilot, $\hat{H}_0 = 0$ is used, corresponding to zero bias, so that the NLSE reduces to pure nonlinear phase rotation. In a trained model, $\hat{H}_0$ would be a learnable Hermitian operator, playing the role of both bias terms simultaneously.

9.1.5. External Parameterization of the Nonlinear Coupling

In classical transformers, activation functions (ReLU, GELU, SiLU) are static architectural choices, fixed before training. Learning occurs through weight matrices that exploit this fixed nonlinearity. The NLSE coupling γ plays an analogous role in the QSC: a fixed, pre-training architectural parameter whose presence and norm-compatibility matter more than its precise value.

In a quantum-native setting, however, γ has a unique physical consequence: it is a property of the computational medium (e.g., the Kerr coefficient of a nonlinear photonic waveguide), not a software parameter. Different values of γ require different physical media, just as different activation functions require different architectural choices that cannot be changed mid-training. QSC training therefore has a distinct two-level structure: (i) *Architecture design*: choosing the nonlinear medium and fixing γ, analogous to choosing the activation function; and (ii) *Training*: updating gate angles with γ held constant. Comparing different values of γ requires separate training runs, analogous to ablation studies over activation functions.

10. Output as Generalized Measurement

The output layer of the transformer maps the final hidden state to a probability distribution over the vocabulary. In the quantum framework, this is a positive operator-valued measure (POVM)

$$P(v_j|\psi) = \langle\psi|E_j|\psi\rangle \quad E_j \succeq 0 \quad \sum_j E_j = \mathbb{1} \tag{50}$$

A POVM is a generalization of projective measurement that allows for more flexible and general measurements on a quantum state. The classical softmax output layer corresponds to the restricted case in which the POVM elements take the rank-1 form $E_j = c\,|w_j\rangle\langle w_j|$, where $|w_j\rangle$ are the output embedding vectors and $c > 0$ is a normalization constant chosen so that $\sum_j E_j = \mathbb{1}$. When the vocabulary size $|\mathcal{V}|$ exceeds the Hilbert space dimension 2^n, the vectors $\{|w_j\rangle\}$ must form a tight frame satisfying $\sum_j |w_j\rangle\langle w_j| = (|\mathcal{V}|/2^n)\,\mathbb{1}$, giving $c = 2^n/|\mathcal{V}|$. In practice, the pilot implementation computes unnormalized overlaps $|\langle w_j \mid \psi_{\text{out}}\rangle|^2$ and applies classical renormalization, which is equivalent to a POVM with the appropriate frame normalization.

11. The Quantum Semantic Circuit

This section provides a formal definition of the Quantum Semantic Circuit (QSC) architecture, detailing its components, layer-by-layer operation, and trainable parameters. It establishes the QSC as a concrete quantum-native implementation of the transformer concept, highlighting key differences in parameterization and operation compared to classical transformers.

11.1. Formal Definition

The Quantum Semantic Circuit (QSC) is a parameterized quantum system designed to mirror the key operations of a classical transformer in a quantum-native way. It can be formally described as a tuple

$$\text{QSC} = \left(\mathcal{H},\, \mathcal{E},\, \{U_l^{(Q)}, U_l^{(K)}, U_l^{(V)}, U_l^{(\text{ff})}, \hat{H}_l\}_{l=1}^{L},\, \hat{M}\right) \tag{51}$$

consisting of a Hilbert space $\mathcal{H} = \mathbb{C}^{2^n}$, a quantum embedding map $\mathcal{E} : \mathcal{V} \to \mathcal{H}$, a set of unitary operators and Hamiltonians for each layer, and a POVM measurement operator $\hat{M}$. While the choice of basis for $\mathcal{H}$ is arbitrary in theory, a specific computational basis is typically chosen for practical implementation.

The embedding map transforms tokens into quantum states: $|\psi_i\rangle = U_{\text{embed}}(\boldsymbol{\theta}_i)|0\rangle^{\otimes n}$. The unitary operation U_{embed} maps the initial state $|0\rangle^{\otimes n}$ to the specific quantum state representing the token, encoding its semantic information. Each layer l then applies a query unitary $U_l^{(Q)}$, key unitary $U_l^{(K)}$, value unitary $U_l^{(V)}$, and FFN unitary $U_l^{(\text{ff})}$, all implemented via interference, along with an evolution Hamiltonian $\hat{H}_l$ whose free part $\hat{H}_0$ plays the role of the classical bias terms $\mathbf{b}_1$ and $\mathbf{b}_2$ (Section 9). The final step is a POVM measurement $\hat{M}$ with elements $\{E_j\}$ satisfying $E_j \succeq 0$ and $\sum_j E_j = \mathbb{1}$.

The quantum attention score between the current token and context token j is computed via the semantic energy

$$E_j = -\frac{\text{Re}\langle\psi_q \mid \psi_{k_j}\rangle}{\sqrt{d}} = -\frac{\text{Re}\left(\langle\psi_x|U_l^{(Q)\dagger}U_l^{(K)}|\psi_{x_j}\rangle\right)}{\sqrt{d}} \tag{52}$$

where $|\psi_x\rangle$ is the quantum state of the current token and $|\psi_{x_j}\rangle$ is the quantum state of the j-th context token. The attention weight is then $\alpha_j = e^{-\beta E_j}/\sum_m e^{-\beta E_m}$. The real part is used rather than the squared modulus $|\langle\psi_q \mid \psi_{k_j}\rangle|^2$ because it reduces exactly to the classical dot product for real-valued states and preserves the sign structure of classical attention scores; see Section 6.

11.2. Layer-by-Layer Operation

Position encoding is applied once before the layer loop, as shown in Figure 4

$$|\psi_j^{(p)}\rangle = U_{\text{pos}}(j)|\psi_j\rangle \tag{53}$$

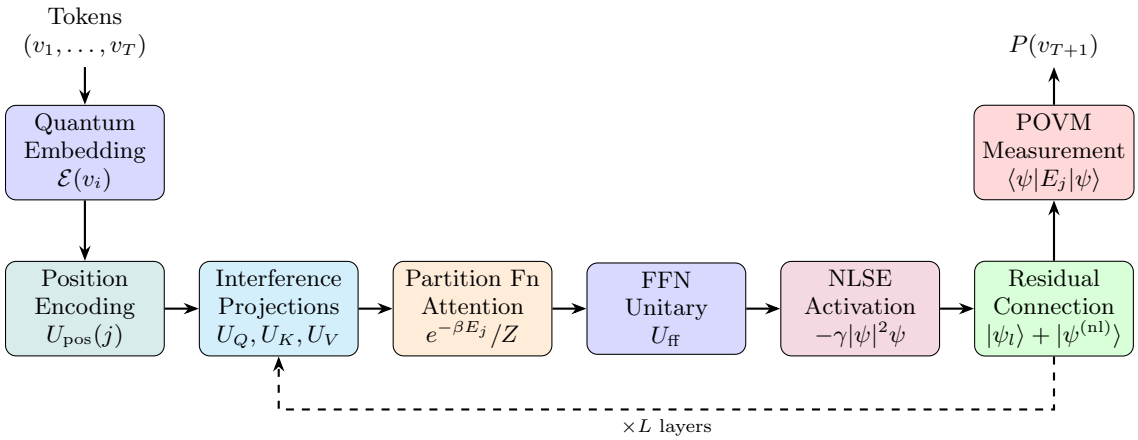


Figure 4: Architecture of the Quantum Semantic Circuit (QSC). Tokens are embedded as quantum states and encoded with position information via the diagonal phase unitary $U_{\text{pos}}(j)$ before entering the layer loop. Each of the L layers then applies interference projections U_Q, U_K, U_V, partition function attention $e^{-\beta E_j}/Z$, FFN unitary U_{ff}, NLSE activation $-\gamma|\psi|^2\psi$, and a residual connection $|\psi_l\rangle + |\psi^{(nl)}\rangle$, as indicated by the dashed feedback arrow. The position encoding applies a norm-preserving phase rotation $|\psi_j\rangle \to U_{\text{pos}}(j)|\psi_j\rangle$, the quantum analog of classical sinusoidal position encoding (Section 4). The CNOT gates create entanglement between qubits, coupling semantic dimensions. After all L layers, a single POVM measurement $\langle\psi|E_j|\psi\rangle$ produces the output probability distribution $P(v_T{+}1)$.

where j is the token's position in the input sequence. Each of the L layers then processes an input state $|\psi_l\rangle$ representing the current token, along with context token states $\{|\psi_{x_j}\rangle\}_{j=1}^{m}$, through the following steps:

1. Interference projection. Compute query, key, and value states. The query is formed from the current token; keys and values are formed from each context token $|\psi_{x_j}\rangle$

$$|\psi_q\rangle = U_l^{(Q)}|\psi_l^{(p)}\rangle \quad |\psi_{k_j}\rangle = U_l^{(K)}|\psi_{x_j}\rangle \quad |\psi_{v_j}\rangle = U_l^{(V)}|\psi_{x_j}\rangle \tag{54}$$

2. Partition function attention. Compute semantic energies and attention weights

$$E_j = -\text{Re}\langle\psi_q \,|\, \psi_{k_j}\rangle/\sqrt{d} \quad \alpha_j = \frac{e^{-\beta E_j}}{\sum_m e^{-\beta E_m}} \tag{55}$$

where E_j here denotes the semantic energy (not to be confused with the POVM elements $\{E_j\}$ of the output measurement, which appear only at the final step).

3. Value aggregation. Form the output density matrix

$$\rho_l^{(\text{attn})} = \sum_j \alpha_j |\psi_{v_j}\rangle\langle\psi_{v_j}| \tag{56}$$

Extract the dominant eigenvector $|\psi_l^{(\text{attn})}\rangle$ of $\rho_l^{(\text{attn})}$.

4. FFN unitary. Apply U_{ff} to rotate the attention output, preparing it for nonlinear evolution

$$|\psi_l^{(\text{ff})}\rangle = U_l^{(\text{ff})}|\psi_l^{(\text{attn})}\rangle \tag{57}$$

5. NLSE activation. Apply one step of nonlinear evolution to the FFN output. The theoretically complete treatment applies the nonlinear von Neumann equation (40) directly to $\rho_l^{(\text{ff})} = U_l^{(\text{ff})}\rho_l^{(\text{attn})}U_l^{(\text{ff})\dagger}$

$$i\hbar\frac{\partial\rho_l^{(\mathrm{ff})}}{\partial t} = [\hat{H}_0,\, \rho_l^{(\mathrm{ff})}] - \gamma[\mathrm{diag}(\rho_l^{(\mathrm{ff})}),\, \rho_l^{(\mathrm{ff})}] \tag{58}$$

6. Residual connection. Add the residual and renormalize

$$|\psi_{l+1}\rangle \propto |\psi_l\rangle + |\psi_l^{(\mathrm{nl})}\rangle \tag{59}$$

This corresponds to one step of Hamiltonian evolution as in Eq. (41). In specific implementations, $|\psi_l\rangle$ may be replaced by a transformed version of the initial token embedding, such as $U_l^{(V)}|\psi_{\text{token}}\rangle$, as demonstrated in the simulator pilot (Section 12.7).

11.3. Trainable Parameters of the QSC

Having established the full architecture, we can now identify precisely which quantities are trainable in the QSC and how they compare to the classical transformer. While this section catalogues the trainable parameters and their scaling, the experimental estimation of these parameters, analogous to backpropagation in classical models, is a separate research question.

11.3.1 Parameter Taxonomy

The QSC has three qualitatively distinct categories of parameter, which differ not only in quantity but in the nature of how they are set. Category 1: Gate angles (learned during training). The primary trainable parameters are the rotation angles of the parameterized unitary circuits (U_Q, U_K, U_V, U_{ff}, and U_{embed}). For an L-layer hardware-efficient ansatz on n qubits, each unitary contributes $O(nL)$ angles. Table 3 summarizes the gate angle count per QSC layer, compared to the classical transformer.

Component	Classical	Quantum
Query projection	d^2 entries in W_Q	$O(n^2)$ angles in U_Q
Key projection	d^2 entries in W_K	$O(n^2)$ angles in U_K
Value projection	d^2 entries in W_V	$O(n^2)$ angles in U_V
FFN projection	$8d^2$ entries in W_1, W_2	$O(n^2)$ angles in U_{ff}
FFN bias terms	$\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^{d_{\text{ff}}}$	$\hat{H}_0$: learnable Hermitian operator
Token embeddings	$\lvert\mathcal{V}\rvert \cdot d$ entries	$\lvert\mathcal{V}\rvert \cdot O(n)$ angles
Position encoding	$T \cdot d$ entries	$T \cdot O(n)$ angles
Layer normalization	$2d$ entries per norm	*not needed*
Output projection W_2	$4d^2$ entries	*not needed*

Table 3: Trainable parameter counts per QSC layer versus classical transformer layer. The quantum column assumes an $O(n^2)$-depth hardware-efficient ansatz, sufficient for approximate universality over $U(2^n)$. Layer normalization parameters are absent in the QSC because unitarity enforces $\langle\psi|\psi\rangle = 1$ automatically. The FFN output projection W_2 is absent because the quantum state requires no dimensional contraction (Section 9).

Category 2: The nonlinear coupling (fixed at architecture design time). The NLSE coupling γ is a physical property of the computational medium, fixed before training and analogous to the choice of activation function in a classical transformer. It is an architectural degree of freedom, but not a trainable parameter.

Category 3: The POVM elements (output layer). The output measurement $\{E_j\}$ has elements $E_j = |w_j\rangle\langle w_j|$ (rank-1 case), parameterized by output embedding vectors $|w_j\rangle \in \mathcal{H}$ for each vocabulary token v_j. The total output layer parameter count is $|\mathcal{V}| \bullet O(n)$, compared to $|\mathcal{V}| \bullet d$ in the classical unembedding matrix.

11.3.2. Total Parameter Count and Scaling

Table 4 summarizes the total trainable parameter counts for representative model scales, comparing the classical transformer to the QSC. The quantum column counts only Category 1 parameters (gate angles); γ is excluded as it is not updated during training.

11.3.3. What Is Not Counted

Two classical parameter categories are absent from the QSC count entirely, not merely compressed:
i) Layer normalization parameters: Classical transformers require learned affine parameters for recentering and rescaling. The QSC requires no such parameters because unitarity automatically enforces $\langle\psi|\psi\rangle = 1$ at every step, preserving the full geometric structure of the state.

Model	d	n	Classical	Quantum
Toy (this paper)	8	3	$\sim 10^3$	$\sim 10^2$
BERT-base	768	10	$\sim 10^8$	$\sim 10^4$
GPT-2	1600	11	$\sim 10^9$	$\sim 10^4$
GPT-3	12288	14	$\sim 10^{11}$	$\sim 10^5$

Table 4: Total trainable parameter counts for classical transformer versus QSC at representative model scales. The quantum counts assume $O(n^2)$-depth ansatz circuits and $L = 96$ layers. The reduction is exponential in n, reflecting the exponential compression of the quantum parameterization. This compression is only realized in practice if the relevant transformations lie within the structured submanifold of $U(2^n)$ accessible to the circuit ansatz.

ii) FFN output projection W_2: The classical W_2 contracts the expanded FFN representation back to dimension d. This contraction is unnecessary in the QSC because the state remains in $\mathbb{C}^{2^n}$ throughout, leveraging phase information and avoiding dimensional expansion. Consequently, the W_2 parameters have no quantum analog.

12. Simulator Pilot

This section describes a simulator pilot conducted to validate the QSC architecture and demonstrate the computational realizability of its key components.

12.1. Setup and Objectives

The pilot validates the QSC architecture on the Qiskit Aer statevector simulator. The objectives are:
i. Validate all eight correspondences: five (embedding, projection, attention, NLSE, and output) by direct circuit execution, two (position encoding and FFN correspondence) by mathematical argument, and one (FFN unitary) by simplifying assumption ($U_{\text{ff}} = \mathbb{1}$).
ii. Demonstrate that the QSC correctly disambiguates a polysemous token in context.
iii. Confirm that the NLSE nonlinearity does real computational work beyond linear evolution.
iv. Verify norm preservation throughout, maintaining the Born rule interpretation.

12.2. Task: Lexical Disambiguation

The pilot uses a 4-token vocabulary $\mathcal{V} = \{\texttt{money}, \texttt{bank}, \texttt{river}, \texttt{water}\}$ on $n = 3$ qubits, spanning $\mathcal{H} = \mathbb{C}^8$. The task is lexical disambiguation: given context [`river`, `water`] and query token `bank`, predict the next token. The token `bank` is polysemous, with both a finance sense (bank account, money) and a nature sense (river bank, water's edge); the context should activate the nature sense, making `water` the expected output. The task is minimal yet non-trivial, requiring the full pipeline to function correctly and consistently, see Fig 5. The pilot parameters are:
• Qubits: $n = 3$, dimension $d = 8$
• NLSE: $\gamma = 0.05$, $\Delta t = 0.05$, $N_{\text{steps}} = 50$, total evolution time $T = N_{\text{steps}}$ • $\Delta \text{I} = 2.5$; effective nonlinear phase $\gamma T = 0.125$
• Attention inverse temperature: $\beta = 5.0$
• Residual weight: $\alpha = 0.2$, chosen as a pilot design parameter giving a non-trivial mix between the residual and processed outputs; the classical transformer corresponds to the symmetric case $\alpha = 0.5$ in the normalized combination, and the optimal value for a trained model would be learned from data.
• FFN unitary: $U_{\text{ff}} = \mathbb{1}$ (identity) for this pilot; the FFN unitary step is omitted as a simplifying assumption, and the FFN correspondence is validated by mathematical argument rather than direct circuit execution.

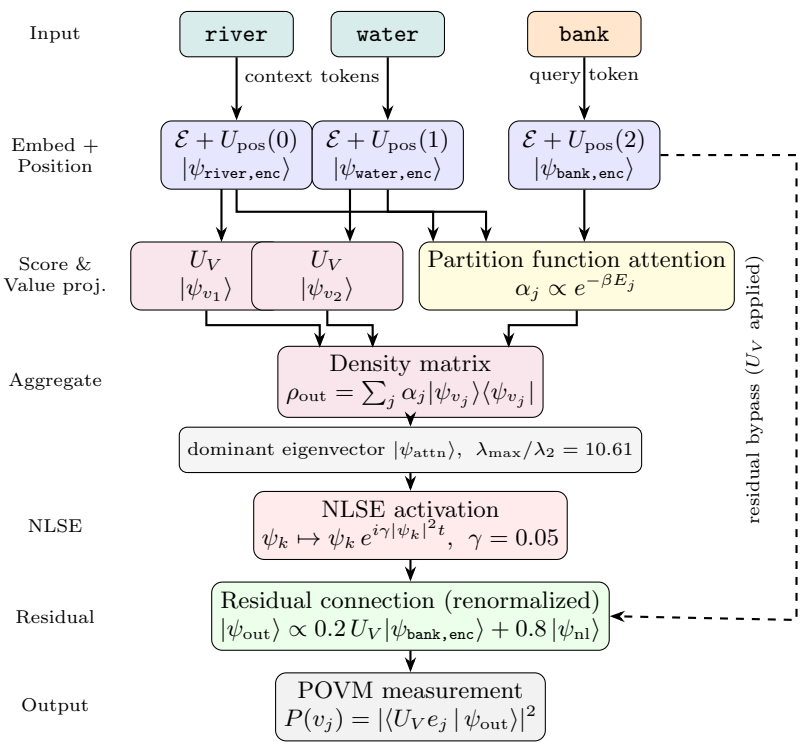


Figure 5: QSC pilot forward pass as implemented in the simulation code. Tokens are embedded and position-encoded (row 2). The same position-encoded states are used simultaneously for attention scoring and value projection U_V (row 3): context tokens feed both the partition function attention (right, $\beta = 5.0$) and the U_V projection (left), while the query token feeds only the attention scoring. The attention weights $\alpha_{\text{water}} = 0.5151 > \alpha_{\text{river}} = 0.4849$ and the value states $|\psi_{v_j}\rangle$ are combined into the density matrix ρout; the dominant eigenvector is extracted ($\lambda_{\text{max}} / \lambda_2 = 10.61$). The NLSE ($\gamma = 0.05$) applies state-dependent phase rotation, the residual combines 80% NLSE output with 20% $U_V |\psi_{\text{bank,enc}}\rangle$ (dashed bypass), and the POVM correctly predicts `water`. All values are from simulation output.

• Projection parameters: two-layer hardware-efficient ansatz with $2n = 6$ parameters per unitary, seeded at 42 for reproducibility

The 3-qubit pilot implementation of the QSC forward pass is illustrated in Figure 6.

12.3. Token Embeddings

Token embeddings are hand-crafted 8-dimensional unit vectors encoding the required semantic geometry. The embedding axes are:
• Dimension 0: finance (+) / nature (−)
• Dimension 1: water feature (primary)
• Dimension 2: nature/outdoor feature
• Dimension 3: land/earth feature
• Dimension 4: flow/movement feature
• Dimensions 5–7: secondary features

The key design requirement is $\cos(\texttt{bank}, \texttt{water}) > \cos(\texttt{bank}, \texttt{river})$, so that `water` receives higher attention weight than `river` when `bank` is the query. This is achieved by making `bank` primarily a water-edge token (high dimension 1,

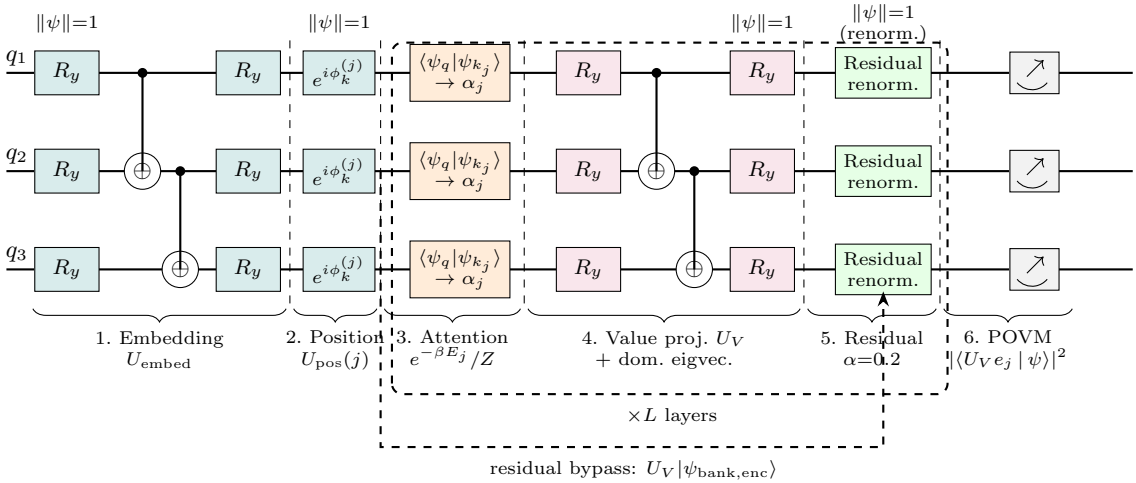


Figure 6: Three-qubit ($n = 3$, $d = 8$) QSC circuit for the simulator pilot, showing the query token register. The circuit is executed on the Qiskit Aer statevector simulator, which computes all quantum operations via linear algebra. Stages proceed left to right. (1) quantum embedding U_{embed}, two-layer R_y–CNOT–R_y ansatz (6 parameters), applied to all tokens on separate registers. (2) position encoding $U_{\text{pos}}(j)$, diagonal phase gates. (3) attention scoring: the quantum inner product $\text{Re}\langle\psi_{\text{query,enc}}\,|\,\psi_{\text{context,enc},j}\rangle/\sqrt{d}$ is computed from the statevectors (on hardware this would use a Hadamard test or swap test circuit); U_Q and U_K are set to identity in this pilot. (4) value projection U_V, same ansatz, applied to context tokens; density matrix $\rho_{\text{out}} = \sum_j \alpha_j |U_V\psi_j\rangle\langle U_V\psi_j|$ formed and dominant eigenvector extracted ($\lambda_{\text{max}}/\lambda_2 = 10.61$). The NLSE ($\gamma = 0.05$, RK4, T = 2.5) acts between stages 4 and 5 as a medium property, not a discrete gate (Section 7.4.). (5) residual connection, combining 0.8 $|\psi_{\text{nl}}\rangle$ with 0.2UV $|\psi_{\text{bank,enc}}\rangle$ (dashed bypass) and renormalizing. (6) POVM measurement. Stages 3–5 repeat L times. Output correctly predicts `water`.

low dimension 4) while `river` is primarily a flow token (high dimension 4). The resulting cosine similarity matrix is given in Table 5, and satisfies all required constraints: cos(`bank`, water) = 0.810 > cos(`bank`, `river`) = 0.764 ≫ cos(bank, money) = −0.216, and cos(`river`, water) = 0.889.

	money	bank	river	water
money	1.000	-0.216	-0.639	-0.693
bank	-0.216	1.000	0.764	0.810
river	-0.639	0.764	1.000	0.889
water	-0.693	0.810	0.889	1.000

Table 5: Cosine similarity matrix for the pilot vocabulary. The critical requirement cos(bank,water) = 0.810 > cos(bank,river) = 0.764 ensures that water receives higher attention weight than river when bank is the query. Negative similarities between money and the nature tokens confirm that the finance and nature senses are geometrically separated.

12.4. NLSE Parameter Sweep

Before running the full pilot, a parameter sweep over (γ, Δ_t, N_{steps}) identifies valid configurations. A configuration is valid if:
(i) Norm preservation: maxsteps $\|\psi_{\text{step}}\| - 1 < 10^{-6}$
(ii) Nonlinearity signal: $\|\psi_{\text{nl}} - \psi_{\text{lin}}\|^2 > 10^{-3}$

The sweep covers $\gamma \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$, $\Delta t \in \{0.001, 0.005, 0.01, 0.05\}$, and $N_{\text{steps}} \in \{10, 20, 50, 100\}$, giving 80 configurations in total, of which 45 satisfy both criteria. The pilot uses $\gamma = 0.05$, $\Delta t = 0.05$, $N_{\text{steps}} = 50$, giving total evolution time $T = 2.5$ and nonlinearity signal 5.51×10^{-2}.

12.5. Component Validation

All eight validation tests pass. The first five are validated by direct circuit execution; two by mathematical argument; and one (FFN unitary) by simplifying assumption ($U_{\text{ff}} = \mathbb{1}$).
(i) Normalization. All four token embeddings satisfy $\||\psi_i\rangle\| = 1$ to numerical precision ($< 10^{-10}$).
(ii) Structural isomorphism. For all $\binom{4}{2} = 6$ token pairs, the shifted cosine similarity $S'_C = (1 + \mathbf{a} \cdot \mathbf{b})/2$ equals the Born rule probability

$P(|0\cdots0\rangle) = |\langle m|\, a\rangle|^2$, where $|m\rangle = (|a\rangle+|b\rangle)/\||a\rangle+|b\rangle\|$. Maximum deviation: $< 10^{-10}$.
(iii) Interference projection. The two-layer parameterized circuit implements a unitary satisfying $U^\dagger U = \mathbb{1}$ to $< 10^{-10}$, preserving the norm of all token embeddings exactly.
(iv) Attention beta rescaling. Boltzmann ($\beta = 5.0$) and Born rule ($\beta = 10.0$) assign identical rank ordering to all context tokens. Maximum numerical difference: 2.03×10^{-2}, consistent with the claim of rank preservation rather than numerical equality.
(v) NLSE parameter sweep. 45 of 80 configurations satisfy both validity criteria. The pilot configuration is confirmed valid with norm drift at machine epsilon and nonlinearity signal 5.51×10^{-2}.
(vi) Position encoding (derivational). The quantum phase rotation preserves all amplitudes $|\psi_k|^2$ exactly (norm deviation $< 10^{-10}$), while the classical additive encoding changes the norm. The linearization error 1.49×10^{-1} confirms that the classical encoding discards higher-order terms. The real shadow error is 0 to machine precision, confirming that the classical sinusoidal functions are exactly the real and imaginary parts of $e^{ij/100002^{k/d}}$.
(vii) FFN correspondence (functional). The quantum FFN operates in $\mathbb{C}^8$ throughout with no dimensional change ($d_{in} = d_{out} = 8$), while the classical FFN expands to $4d = 32$. Norm is preserved to $< 10^{-10}$ without layer normalization. The phase degree-of-freedom fraction $0.049 > 0$ confirms that phase degrees of freedom are active and carry information not representable in the real-valued classical FFN.
(viii) FFN unitary (simplifying assumption). $U_{ff} = \mathbb{1}$ is assumed for this pilot. Norm is trivially preserved.

12.6. Partition Function Bridge

Table 6 verifies the partition function bridge that both distributions assign higher weight to `water` than to `river`, confirming that the rank-preserving rescaling $\beta \to 2\beta$ does not alter the attention ordering. The semantic energies $E_{water} = -0.2864 < E_{river} = -0.2701$ are consistent with $\cos(\texttt{bank}, \texttt{water}) = 0.810 > \cos(\texttt{bank}, \texttt{river}) = 0.764$.

Token	Boltzmann ($\beta = 5.0$)	Born rule ($\beta = 10.0$)	Difference
`river`	0.479653	0.459373	2.03×10^{-2}
`water`	0.520347	0.540627	2.03×10^{-2}

Table 6: Partition function bridge verification. Boltzmann ($\beta = 5.0$) and Born rule ($\beta = 10.0$) assign the same rank ordering: water > river in both cases. The numerical difference 2.03×10^{-2} is expected: the claim is rank preservation under $\beta \to 2\beta$, not numerical equality.

12.7. End-to-End Forward Pass

12.7.1. Geometric Consistency of the Pipeline

The QSC pipeline maintains geometric consistency throughout by operating in a single space: the value-projected Hilbert space $U_V\mathcal{H}$. The attention output, NLSE activation, residual term, and POVM basis vectors all live in $U_V\mathcal{H}$

$$|\psi_{\text{out}}\rangle \propto \alpha\, U_V|\psi_{\texttt{bank}}\rangle + (1-\alpha)|\psi_{\text{nl}}\rangle \quad \alpha = 0.2 \tag{60}$$

This consistency is essential: mixing projected and unprojected states would make the POVM overlaps geometrically meaningless.

12.7.2. Output Probabilities

Table 7 shows the output probability distributions for the linear ($\gamma = 0$) and nonlinear ($\gamma = 0.05$) forward passes. Both correctly predict `water`. The POVM overlaps confirm why: $|\langle U_V\psi_{\texttt{water}}\,|\,\psi_{\text{out}}\rangle|^2 = 0.881 > |\langle U_V\psi_{\texttt{river}}\,|\,\psi_{\text{out}}\rangle|^2 = 0.853 > |\langle U_V\psi_{\texttt{bank}}\,|\,\psi_{\text{out}}\rangle|^2 = 0.708$, giving $P(\texttt{water}) = 0.3125$ as the highest probability after renormalization.

Token	No NLSE ($\gamma = 0$)	With NLSE ($\gamma = 0.05$)	ΔP
`water`	0.3124	0.3125	+0.0001
`river`	0.3030	0.3028	-0.0002
`bank`	0.2532	0.2512	-0.0021
`money`	0.1313	0.1335	+0.0022

Table 7: Output probability distributions for the linear ($\gamma = 0$) and nonlinear ($\gamma = 0.05$) QSC forward passes. Both correctly predict water. The NLSE redistributes probability mass, weakening bank (−0.0021) and strengthening money (+0.0022) while leaving the water prediction intact, demonstrating that the cubic self-interaction does real computational work beyond linear evolution. ΔP = nonlinear minus linear probability.

The dominant NLSE effect is a state-dependent phase rotation

$$\psi_k(t) = \psi_k(0)\, e^{i\gamma|\psi_k(0)|^2 t} \tag{61}$$

which modifies the interference pattern between the NLSE output and the residual term, shifting overlap from bank toward money. The correct prediction (water) is robust to this redistribution, remaining the highest probability token in both passes. The effect magnitude ($\sim 10^{-3}$) is consistent with the nonlinearity signal 5.51×10^{-2} attenuated by the residual weight and POVM projection geometry.

12.7.3. Summary of Pilot Results

Table 8 summarizes the validation status for all eight correspondences.

Validation	Type	Claim	Result	Status
Embedding	Exact	$\|\|\|\psi_i\rangle\|\| = 1$; $S'_C = P(\|0\cdots0\rangle)$	Max deviation $< 10^{-10}$	PASS
Projection	Repr.	$U^\dagger U = \mathbb{1}$; norm preserved	Max error $< 10^{-10}$	PASS
Attention	Ex.up.	Rank preserved under $\beta \to 2\beta$	rank_preserved = True; max_diff $= 2.03 \times 10^{-2}$	PASS
NLSE	Func.	Norm preserved; signal $> 10^{-3}$	Drift $= 2.22 \times 10^{-16}$; signal $= 5.51 \times 10^{-2}$	PASS
POVM	Repr.	Correct prediction	`water` predicted	PASS
Position	Deriv.	Norm preserved; linearization error > 0	Norm ok; error $= 1.49 \times 10^{-1}$; shadow error $= 0$	PASS
FFN	Func.	No dim. change; norm preserved; phase dof > 0	$d_{\text{in}} = d_{\text{out}} = 8$; norm ok; phase fraction $= 0.049$	PASS
FFN unitary	Assumpt.	$U_{\text{ff}} = \mathbb{1}$; norm preserved	Identity assumed; norm trivially preserved	PASS

Table 8: Validation summary for all eight correspondences. Type abbreviations: Repr. = Representational; Func. = Functional; Deriv. = Derivational; Ex.up. = Exact up to rescaling; Assumpt. = Simplifying assumption. The first five correspondences are validated by direct circuit execution on the Qiskit Aer statevector simulator; position encoding and FFN correspondence are validated by mathematical argument; FFN unitary is validated by simplifying assumption.

The pilot establishes that the QSC architecture is self-consistent, that it correctly resolves the disambiguation task, and that the NLSE nonlinearity does real computational work while preserving norm to machine epsilon throughout.

What the pilot does not establish is equally important to state. The 4-token, 3-qubit task is a proof of concept, not a performance benchmark. The embeddings are hand-crafted; a trained model would learn them. The projection unitaries are random; a trained model would optimize them. The NLSE coupling γ is fixed as an architectural constant; a trained model would be trained separately under different fixed values of γ, analogous to ablation studies over activation functions. The pilot validates the architecture's self-consistency and computational realizability, not its performance at scale.

13. Discussion

The QSC framework establishes correspondences between classical transformer components and quantum mechanical operations, classified by their logical status. Three correspondences are exact mathematical identities: L2normalized embeddings are quantum states, cosine similarity is a Born rule probability, and softmax attention weights are a Boltzmann distribution. These hold regardless of whether a quantum computer is ever used. The classical sinusoidal position encoding is the first-order real projection of an exact quantum phase rotation; the quantum version is the natural object and the classical version is its shadow. A key aspect of the QSC is the use of quantum interference in the projection operations, providing a richer and more nuanced mechanism for computing relationships between token states compared to classical dot products. The cubic NLSE is the unique minimal norm-preserving nonlinearity consistent with the Hilbert-space structure of the encoding, derived from three requirements rather than chosen by analogy. The simulator pilot confirms that all eight correspondences are self-consistent and that the NLSE does measurable computational work on a non-trivial disambiguation task.

The framework also offers a natural explanation for empirically observed structure in the classical FFN that has no account within the classical framework itself. The dimensional expansion $d \to 4d \to d$ and the asymmetry between W_1 and W_2 can be understood as shadows of quantum symmetry breaking and restoration in the semantic superspace. In the minimal single-head regime, the symmetry cycle suggests $W_2 \approx W_1^T$ as a consistency indicator; in large multi-head models, the Higgs analogy suggests that the rank of $W_2 - W_1^T$ scales with the number of attention heads H and that the blockwise transpose relation holds within each head's subspace. These

explanations are informative regardless of whether they are confirmed empirically, since the classical FFN structure is itself an empirical design choice rather than a derived necessity.

Several directions present themselves for empirical follow-up. Whether trained transformer projection matrices have the structured cancellation and low-rank correlation patterns that would make the quantum circuit ansatz a good fit is an open question; the LoRA evidence for low-rank weight updates is suggestive but not conclusive. Assessing how well a hardware-efficient ansatz approximates trained projection matrices, and whether the approximation error correlates with task performance, would directly test the practical reach of the quantum parameterization. On the hardware side, the NLSE nonlinearity cannot be implemented on current gate-based devices, but photonic Kerr nonlinearities, measurement-and-feedback schemes, and double-well potential systems in superconducting circuit QED each offer concrete paths toward physical realization, and determining which is most practical for transformer-scale computation is an engineering question that can be addressed independently of the theoretical framework. Parameter compression and computational speedup relative to classical transformers are theoretically possible but contingent on conditions not yet verified: the learned transformations must lie within the submanifold accessible to the circuit ansatz, and the shot overhead of quantum measurement must be overcome either through larger embedding dimensions or demonstrated parameter efficiency at comparable task performance.

The QSC framework is a theoretical foundation, not a deployed system. The correspondences established in this paper map each classical transformer component to its most natural quantum mechanical counterpart, preserving the architectural logic of the classical system within the quantum framework. This mapping is a necessary first step, but it is not necessarily the correct or complete path to a native quantum language model. The classical transformer architecture was designed under the constraints of real-valued computation, and its structure reflects those constraints. A truly native quantum LLM may require additional structures with no classical analog, or may need to modify the quantum counterparts presented here in ways that depart substantially from the classical blueprint. The density matrix formulation of value aggregation, the soliton interpretation of stable representations, and the imaginary component of the attention score accessible via the *SH* circuit are examples of quantum structures that have no direct classical counterpart and whose role in a native quantum LLM is not yet understood. If a system built from the components described in this paper produces suboptimal results, this should not be interpreted as evidence that the theoretical framework is wrong. It may instead indicate that the framework is incomplete: that the correct quantum architecture requires structures beyond those obtained by direct translation from the classical case.

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. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. International Conference on Learning Representations, 2014.
3. Rush, A. M., Chopra, S., & Weston, J. (2015, September). A neural attention model for abstractive sentence summarization. In Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 379-389).
4. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. Computer Science, Linguistics, 2018.
5. Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2022). Efficient transformers: A survey. ACM Computing Surveys, 55(6), 1-28.
6. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. International Conference on Learning Representations.1(2) 3.
7. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.
8. Dunjko, V., & Briegel, H. J. (2018). Machine learning & artificial intelligence in the quantum domain: a review of recent progress. Reports on Progress in Physics, 81(7), 074001.
9. Rebentrost, P., Mohseni, M., & Lloyd, S. (2014). Quantum support vector machine for big data classification. Physical review letters, 113(13), 130503.
10. Farhi, E., & Neven, H. (2018). Classification with quantum neural networks on near term processors. arXiv preprint arXiv:1802.06002.
11. Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., & Lloyd, S. (2017). Quantum machine learning. Nature, 549(7671), 195-202.
12. Preskill, J. (2018). Quantum computing in the NISQ era and beyond. Quantum, 2, 79.
13. Ladd, T. D., Jelezko, F., Laflamme, R., Nakamura, Y., Monroe, C., & O'Brien, J. L. (2010). Quantum computers. nature, 464(7285), 45-53.
14. Nielsen, M. A., & Chuang, I. L. (2010). Quantum computation and quantum information. Cambridge university press.
15. 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.
16. Coecke, B., de Felice, G., Meichanetzidis, K., & Toumi, A. (2020). Foundations for near-term quantum natural language processing. arXiv preprint arXiv:2012.03755.

17. Widdows, D., & Widdows, D. (2004). Geometry and meaning (Vol. 773). Stanford: CSLI publications.
18. Coecke, B., Sadrzadeh, M., & Clark, S. (2010). Mathematical foundations for a compositional distributional model of meaning. Lambek Festschrift Linguistic Analysis, 36(1).
19. Gupta, A., Kaur, K., Gupta, V., & Shah, C. (2025). QLENS: Towards A Quantum Perspective of Language Transformers. arXiv preprint arXiv:2510.11963.
20. 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.
21. Laine, T. A. (2025). The Quantum LLM: Modeling Semantic Spaces with Quantum Principles. OA J Applied Sci Technol, 3(2), 01-13.
22. Laine, T. A. (2026). Quantum LLMs Using Quantum Computing to Analyze and Process Semantic Information. OA J Applied Sci Technol, 4(1), 01-19.
23. Laine, T. A. (2026). Discrete semantic states and Hamiltonian dynamics in LLM embedding spaces. OA Journal of Applied Science and Technology, 4(1), 1–23.
24. Laine, T. A. (2026). Quantum computation of partition function similarity for large language models. OA Journal of Applied Science and Technology, 4(1), 1–11.
25. Laine, T. A. (2026). Quantum hierarchy for understanding LLM representations by modeling linear projections and nonlinear dynamics. OA Journal of Applied Science and Technology, 4(1), 1–43.
26. Laine, T. A. (2026). Quantum algorithms for large language models on noisy intermediate-scale quantum computers. OA Journal of Applied Science and Technology, 4(1), 1–13.
27. Laine, T. A. (2026). Structural Isomorphism Between LLM Embedding Spaces and Quantum Mechanical Systems. OA J Applied Sci Technol. 4(2), 1-18.