

The Wave Optics of Semantic Computation: Complex Transformers, Phase Interference, and Readout Capacity Limits

Timo Aukusti Laine* 

Financial Physics Lab, Finland

***Corresponding Author**

Timo Aukusti Laine, Financial Physics Lab, Finland.

Submitted: 2026, Aug 18; Accepted: 2026, Sep 25; Published: 2026, Sep 30

Citation: Laine, T. A. (2026). The Wave Optics of Semantic Computation: Complex Transformers, Phase Interference, and Readout Capacity Limits. *OA J Applied Sci Technol*, 4(3), 01-19.

Abstract

We establish a first-principles field theory of semantic computation that grounds language representations in physical wave mechanics, demonstrating that complex-valued vectors define an autonomous classical wave theory. We structure semantic computation into three exact mathematical levels: Level 1 (standard real transformers operating as phase-restricted subspaces), Level 2 (complex transformers operating strictly as classical wave optics governed by Poincaré polarization kinematics and an explicit layer-wise coherence budget), and Level 3 (quantum semantic fields obtained through canonical quantization of the Level 2 phase space without approximation). Level 2 provides a self-contained physical framework that maps natively onto passive, room-temperature photonic integrated circuits (PICs). By formalizing this structure, we derive exact readout capacity bounds across all levels and decompose LLM hallucinations into an unremovable readout capacity floor forced by vocabulary scaling, an architectural gap caused by real-space phase restrictions, and an empirical gap reducible by training. Canonical quantization to Level 3 yields non-classical limits, including a semantic zero-point energy and an entropic attention bandwidth relation. Empirically evaluating Level 2 on FB15k-237 knowledge graph completion across ten random seeds confirms that complex wave transformers consistently outperform matched real baselines in accuracy, Mean Reciprocal Rank, and hallucination reduction while utilizing 4.6% fewer total parameters, proving phase interference to be a parameter-efficient, physically grounded inductive bias for sequence modeling.

1. Introduction

Large language models represent semantic information in high-dimensional vector spaces, yet the field lacks a foundational physical theory governing representation dynamics, internal layer evolution, and structural failure modes such as candidate confusion and hallucinations. Current mitigation strategies treat hallucinations primarily as empirical training defects addressed through post-hoc interventions like retrieval-augmented generation or fine-tuning. However, growing evidence suggests that representation degeneration and candidate confusion have deeper structural and architectural origins.

To move beyond empirical black-box methods and establish a first-principles foundation, we formulate an autonomous physical theory of semantic computation centered on complex transformers. Instead of treating complex representations as speculative quantum analogies or ad-hoc architectural tweaks, we demonstrate that complex arithmetic defines an autonomous classical wave theory. Within this framework, semantic computation structures into three exact mathematical levels. Level 1 represents standard real-valued transformers operating as phase-restricted subspaces that lack relative phase degrees of freedom and cannot perform wave interference. Level 2 introduces complex transformers operating strictly as classical wave optics, where attention functions as a semantic multi-slit experiment, intermediate residual streams act as polarized optical fields, and outputs are deterministic classical intensity readouts. Level 3 defines quantum semantic fields obtained through canonical quantization of the Level 2 phase space without approximations, promoting field quadratures to operators satisfying canonical commutation relations, Hermitian state evolution, and Born-rule measurement collapse.

In this paper, we focus primarily on the second level. Grounding complex transformers in classical wave optics leverages the formal machinery of physical optics, including optical coherence budgets, Poincaré polarization kinematics, and interference dynamics,

showing that constructive and destructive phase interference provides a parameter-efficient inductive bias that improves model accuracy and reduces hallucinations. Because this wave-optical formulation is entirely classical, it maps directly onto room-temperature Photonic Integrated Circuits (PICs), where complex vector arithmetic executes natively and passively at the speed of light. This field theory allows us to derive exact readout capacity bounds and decompose total hallucination rates into an unremovable fundamental floor forced by vocabulary scaling bottlenecks, an architectural gap closable by introducing complex phase degrees of freedom, and an empirical gap reducible through optimization. Extending the framework to the third level via canonical quantization derives structural limits with no classical counterpart, including an irreducible semantic zero-point energy, an entropic attention-bandwidth principle, and a layer-depth bound on state reorientation.

On knowledge graph completion benchmarks across ten independent random seeds, complex transformers operating at the second level consistently outperform matched real-valued baselines in accuracy, ranking metrics, and hallucination reduction while using 4.6% fewer total parameters. These empirical findings confirm that classical wave interference serves as a high-capacity, parameter-efficient, and physically grounded computational foundation for sequence modeling.

2. Background and Related Work

The transformer architecture is central to modern natural language processing, enabling unprecedented performance across tasks ranging from machine translation to multi-step reasoning chains [1]. Despite these advancements, hallucinations have emerged as a systematic failure mode across language tasks, producing content that appears fluent and plausible despite being factually erroneous or incoherent [2,3]. Current efforts to mitigate hallucinations primarily focus on post-hoc interventions, such as retrieval-augmented generation, external fact-checking modules, or fine-tuning on meticulously curated datasets [4]. While these methods offer incremental empirical gains, they treat hallucinations as an empirical training defect or external data shortfall, as opposed to an intrinsic structural limitation of the underlying representational architecture. Theoretical lines of work support the view that candidate confusion has a structural origin. The softmax bottleneck theorem proves that mapping a d -dimensional hidden state to a vocabulary of $V > d$ tokens via softmax induces a severe rank deficiency in the representational capacity of log-probabilities [5]. Calibration and computability bounds show that statistical training incentives guarantee nonzero hallucination rates under finite precision [6,7]. Representation degeneration literature establishes that learned embedding spaces collapse into narrow, anisotropic cones, forcing high pairwise overlaps among tokens and leading to candidate confusion during attention scoring [8,9].

At the same time, a growing volume of research explores the fundamental mathematical structures underlying language model embedding spaces by drawing connections to quantum formalisms. The use of quantum frameworks to model semantic phenomena has a long history, establishing that vector space models of meaning naturally admit quantum-logical structure [10–12]. Recent work shows that the mathematical framework governing normalized embeddings exhibits a structural isomorphism with quantum mechanical systems, where L_2 -normalized embedding vectors map directly to quantum states and transformed cosine similarity corresponds to Born-rule measurement probabilities [13]. Building upon this foundation, a quantum semantic structure positions the real-valued transformer embedding space as a projection of a richer complex semantic super space [14,15]. The analogy between semantic superposition and wave interference highlights the potential for complex representations to capture rich semantic dynamics across transformer components, including attention, feed-forward networks, position encodings, and residual connections, which can be viewed as real-subspace restrictions of complex counterparts [15,16]. Within this framework, time in the Schrödinger equation acts as a gauge variable with local $U(1)$ gauge structure and conserved semantic charges, while hallucinations are modeled as quantum tunneling between coherent semantic states and high-noise distractor states via the Landau-Zener model [13,14].

Complex-valued neural networks possess a wide theoretical foundation and established practical advantages in signal processing and sequence modeling [17,18]. In relational domain learning, complex representations effectively capture asymmetric, anti-symmetric, and structural relations in knowledge graph completion through multiplicative scoring models such as ComplEx and RotatE [19,20]. More recently, researchers have explored complex-valued transformer architectures by extending linear transformations and activation functions to the complex domain, typically employing split real-imaginary weight matrices and phase-preserving activations such as modReLU [21–23]. Our work builds on these architectural advancements while making a physical distinction that separates our approach from prior quantum-inspired literature: complex arithmetic alone does not imply quantum mechanics. Grounding Level 2 explicitly in classical wave optics shows that complex transformers operate entirely as classical wave systems governed by phase interference without invoking quantum measurement postulates. We evaluate our models on standard knowledge graph completion benchmarks, specifically FB15k-237, using standard filtered evaluation protocols [24,25].

3. Three Levels of Semantic Computation

We categorize semantic architectures into three distinct levels based on their underlying mathematical structures. Because classical wave optics utilizes the full complex vector formalism without invoking a measurement postulate, complex arithmetic alone does not imply quantum mechanics.

Level 1 represents the standard real-valued transformer. Representations are real vectors in $\mathbb{R}^d$, carrying no relative phase degrees of freedom and displaying no wave interference.

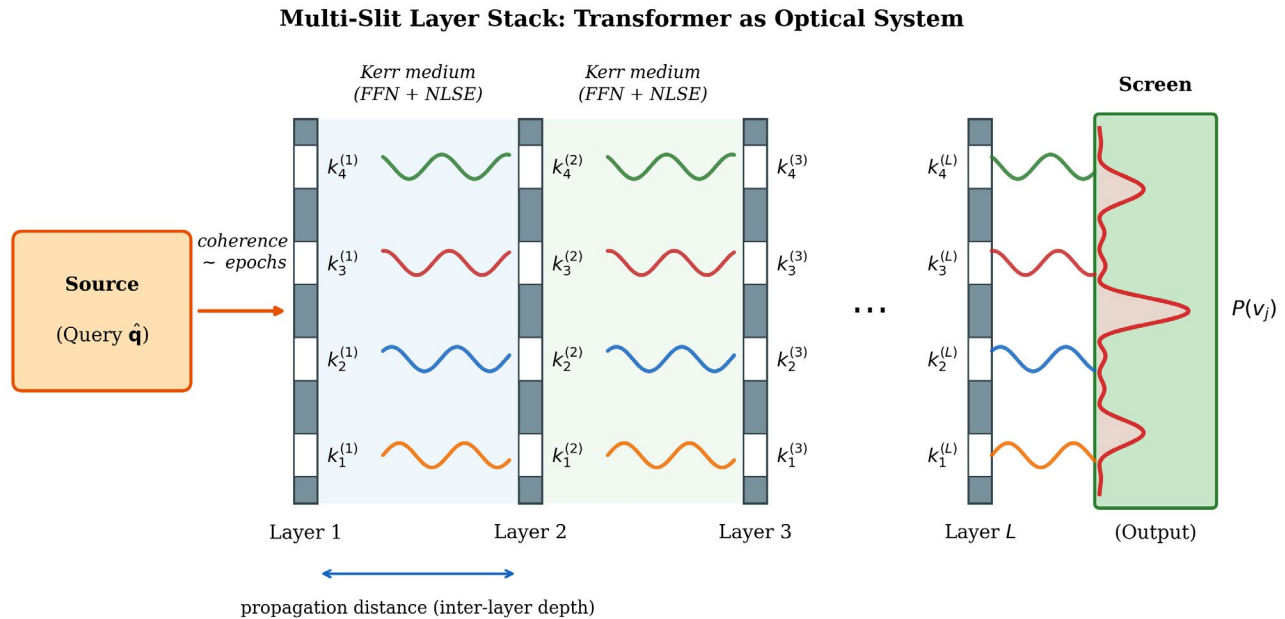


Figure 1: The transformer as a multi-slit optical stack (Level 2). Each layer ℓ acts as a diffraction barrier whose apertures are the key vectors $k_1^{(\ell)}, \dots, k_4^{(\ell)}$. A coherent query $\hat{q}$ superposes across all apertures simultaneously, and the field propagates through a saturable-absorber medium (modReLU) representing the feed-forward block. Mutually coherent amplitudes interfere at the readout screen, where constructive and destructive interference shape candidate weights (an optical reference pattern $P(v_j) \propto [\sin(N\phi)/\sin\phi]^2$ is illustrated). Unlike Level 3 readouts that destructively collapse a single copy to a discrete outcome, the Level 2 readout measures the deterministic intensity fraction $I_j/\sum_k I_k$ across the whole beam without disturbing the state.

Level 2 is the complex transformer, which operates as a classical wave computation. Physically, the attention mechanism acts as a multi-slit optical experiment where query and key fields superpose to form interference fringes that constructively or destructively weight candidate outputs (see Figure 1). Representations are complex vectors in $\mathbb{C}^d$, and attention scores function as two-channel interference terms determined directly by relative phase. Linear maps belong to $GL(d, \mathbb{C})$ as opposed to unitaries, layer normalization operates separately on real and imaginary parts, and the output layer is a real linear functional of the field components. Level 2 stands to quantum theory exactly as Maxwell's electrodynamics stands to quantum optics: complex amplitudes, interference, and coherence operating strictly within classical wave physics.

Level 3 is the quantum semantic field, formally obtained by canonically quantizing the phase space of Level 2. Semantic modes are normal modes of the learned free Hamiltonian, and ladder operators quantize the Level 2 complex amplitudes themselves. On the resulting Fock space, field quadratures satisfy canonical commutation relations, establishing a state-independent Heisenberg bound and an irreducible ground-state energy. States are normalized rays in Hilbert space, state evolution is generated by Hermitian operators, and output readouts execute via Positive Operator-Valued Measures (POVMs) on single state copies.

Incorporating the measurement postulate and Hermitian generators at Level 3 fundamentally alters readout physics compared to classical fields. Measurement collapse renders state uncertainty epistemic by restricting observation of a single copy to a single basis, whereas Level 2 field readouts inspect the complete state vector deterministically. Global phase becomes unobservable under Born-rule probabilities, in contrast to Level 2 readouts that explicitly mix real and imaginary field components. Overcompleteness ($V > d$) shifts from a log-probability rank bottleneck at Levels 1 and 2 to an exact trace ceiling on token confidence at Level 3. Additionally, because the context index and its Fourier-conjugate positional frequency form mutually unbiased bases, sharpening attention in space necessarily spreads it across positional frequencies. Neither c-number commutators nor measurement trade-offs exist at Level 2, where linear maps remain arbitrary elements of $GL(d, \mathbb{C})$.

Placing the components of the quantum semantic circuit [15] into this framework clarifies that several mechanisms assigned to Level 3 in earlier work are properly Level 2 structures. Complex embeddings, interference projections, phase position encodings, residual

update steps, and cubic nonlinear Schrödinger activations are complex-linear or classical Kerr-optical mappings that require no Born rule or c-number commutator generator. Only postselection (supplying the partition function) and POVM readouts are genuinely Level 3, as both act on single copies via the Born rule. Value aggregation operates as a measurement-class jump process that shifts the state discontinuously without performing a Born-rule readout.

4. Hallucination Rate Decomposition

We decompose the total hallucination rate into three distinct components:

$$\text{Hallucination Rate} = \underbrace{\text{Fundamental Floor}}_{\text{readout capacity}} + \underbrace{\text{Architectural Gap}}_{\text{closable by } \mathbb{C}^d} + \underbrace{\text{Empirical Gap}}_{\text{reducible by training}}. \quad (1)$$

The *fundamental floor* is an unremovable error forced by the finite capacity of the readout map when expressing a large vocabulary (V) using a smaller state dimension (d or D). It is a structural property of the final readout mapping, independent of training data quality, optimization algorithms, or internal layer dynamics.

The *architectural gap* represents the performance penalty incurred by restricting computations to real arithmetic ($\mathbb{R}^d$) instead of complex phase space ($\mathbb{C}^d$). A real model lacks relative phase degrees of freedom and cannot perform constructive or destructive interference. This gap can be closed without increasing parameter count by changing the algebra of computation.

The *empirical gap* accounts for conventional engineering errors, including insufficient training data, uncaptured semantic correlations, and distribution shift. It is the sole target of standard data curation, fine-tuning, and retrieval-augmented generation techniques.

5. The Semantic Superspace

The geometry developed in this section is common to all three levels: it characterizes the spatial arrangement of token embeddings independently of the dynamics acting upon them.

5.1. Pairwise Projection Geometry

Token embeddings are represented as unit vectors on S^{d-1} , where pairwise relations define two-dimensional semantic subspaces [13]. For any distinct token pair (i, j) , the angle bisector $\hat{\mathbf{v}}_{ij} = (\mathbf{a}_i + \mathbf{a}_j) / \|\mathbf{a}_i + \mathbf{a}_j\|$ and its orthogonal complement $\hat{\mathbf{w}}_{ij}$ form an orthonormal basis. The rescaled inner product

$$S'_C(i, j) = \frac{1 + \mathbf{a}_i \cdot \mathbf{a}_j}{2} = (\hat{\mathbf{v}}_{ij} \cdot \mathbf{a}_i)^2 \quad (2)$$

measures the squared projection onto the bisector, providing a bounded similarity in $[0, 1]$. The identity is geometric and level-neutral: it is a squared projection at every level, and becomes a Born-rule probability only when read destructively off a single copy, which is the Level 3 readout of Ref. [13] and not a feature of Levels 1 or 2. A normalized state in this subspace is parameterized as

$$|\tilde{\psi}\rangle = \cos \phi |\hat{\mathbf{v}}_{ij}\rangle + e^{i\alpha} \sin \phi |\hat{\mathbf{w}}_{ij}\rangle, \quad (3)$$

where $\cos^2 \phi = S'_C(i, j)$ and $\alpha \in [0, 2\pi)$ represents the relative phase. Because key–query attention scoring relies on pairwise discrimination, this two-dimensional geometry governs token interactions. Real-valued models restrict this relative phase to $|\cos \alpha| = 1$.

5.2. Finite Vocabulary as Confinement

A finite vocabulary of $V < \infty$ tokens distributed over S^{d-1} imposes a non-zero minimum pairwise angular separation

$$\theta_{\min} = \min_{i \neq j} \arccos(\mathbf{a}_i \cdot \mathbf{a}_j) > 0. \quad (4)$$

Because the alignment variance $S'_C(1 - S'_C)$ vanishes at both $S'_C = 1$ and $S'_C = 0$, bounding pairwise separation across the vocabulary requires two constants: $\theta_{\min}$ at the near-identical end and $\epsilon_{\text{ap}} = \min_{i \neq j} S'_C(i, j)$ at the antipodal end. An exhaustive sweep over all 1.26×10^9 token pairs in GPT-2 yields $\sin^2(\theta_{\min}/2) = 1.13 \times 10^{-3}$ and $\epsilon_{\text{ap}} = 0.439$ (see Appendix C for full empirical measurement details). Because near-identical tokens strictly bind the boundary ($\sin^2(\theta_{\min}/2) \ll \epsilon_{\text{ap}}$), the confinement constant simplifies to $\epsilon = \sin^2(\theta_{\min}/2)$. This second-order term sets the per-layer coherence floor for a maximally contested real-valued head.

Overcompleteness ($V > d$) forces non-zero overlaps. V unit vectors in d dimensions satisfy

$$\max_{i \neq j} (\mathbf{a}_i \cdot \mathbf{a}_j)^2 \geq \frac{V-d}{d(V-1)} > 0, \quad (5)$$

proving that perfect pairwise orthogonality is impossible whenever the vocabulary outnumber the embedding dimension.

5.3. Readout Capacity and Vocabulary Scaling

While finite vocabulary confinement forces non-zero spatial overlap among embedding vectors on S^{d-1} , this spatial geometry is not the primary origin of the fundamental error floor. Instead, the binding fundamental error floor is governed by readout capacity; the structural bottleneck created when mapping a d -dimensional hidden state (or D -dimensional Hilbert space) onto a vastly larger vocabulary space $V \gg d$.

This static readout capacity bottleneck manifests differently across computational levels: Levels 1 and 2 (Rank Bottleneck): Mapping a d -dimensional hidden state $\mathbf{H}$ to V logits produces a log-probability matrix of linear rank at most $d+1$. When the true conditional target distribution across the vocabulary has a rank exceeding $d+1$, the network suffers an unremovable representation bottleneck. This forces a structural cross-entropy error floor that scaling training data or parameter depth cannot eliminate.

Level 3 (POVM Trace Ceiling): A Born-rule readout replaces the rank bottleneck with an exact ceiling on the total confidence the readout may distribute across the vocabulary, Eq. (18), which concentrates structural under-confidence primarily onto rare tail entities. Because that ceiling presupposes the measurement postulate, it is stated and derived where the postulate is introduced, in Section 7.3.

We note a formal distinction. Both the rank bottleneck (Levels 1 and 2) and the POVM trace ceiling (Level 3) are capacity limits, set by the state space dimension (d or D) relative to vocabulary size V , and both enter the fundamental floor of Eq. (1).

6. Level 2: The Complex Transformer as Classical Wave Computation

This section describes the complex-valued transformer as an autonomous theory of classical wave computation. It possesses complete wave motion, an exact physical realization in classical polarization optics, an explicit per-layer coherence budget, and a direct mapping onto room-temperature optical hardware.

6.1. Attention as Interference

The attention mechanism acts as a semantic multi-slit experiment, see Figure 2. In standard real attention, a query $\mathbf{q}$ acts as a monochromatic source, key vectors $\mathbf{k}_j$ represent aperture locations, and the scalar score $s_{qj} = d_k^{-1/2} \mathbf{q} \cdot \mathbf{k}_j = d_k^{-1/2} \|\mathbf{q}\| \|\mathbf{k}_j\| \cos \theta_{qj}$, where $\cos \theta_{qj} = \hat{\mathbf{q}} \cdot \hat{\mathbf{k}}_j$, determines the field amplitude transmitted through slit j . Softmax aggregation normalizes these relative field amplitudes into output weights $w_j = e^{s_{qj}} / \sum_k e^{s_{qk}}$, yielding a weighted superposition $|\psi_{\text{out}}\rangle = \sum_j w_j |\psi_{v_j}\rangle$.

In a complex transformer, queries and keys carry complex amplitudes in $\mathbb{C}^d$, and attention is computed using the real part of the Hermitian inner product:

$$s_{qj}^{\mathbb{C}} = \frac{\text{Re}(\mathbf{q}^H \mathbf{k}_j)}{\sqrt{d_k}} = \frac{\mathbf{q}_r \cdot \mathbf{k}_{jr} + \mathbf{q}_i \cdot \mathbf{k}_{ji}}{\sqrt{d_k}}. \quad (6)$$

The score gap between a target key k^* and a competitor k_w splits into two distinct channels:

$$\delta_{qw}^{\mathbb{C}} = \frac{\mathbf{q}_r \cdot (\mathbf{k}_r^* - \mathbf{k}_r^w)}{\sqrt{d_k}} + \frac{\mathbf{q}_i \cdot (\mathbf{k}_i^* - \mathbf{k}_i^w)}{\sqrt{d_k}}. \quad (7)$$

The second term is the interference channel. Note that $\text{Re}(\mathbf{q}^H \mathbf{k})$ equals the $\mathbb{R}^{2d}$ inner product of $[\mathbf{q}_r; \mathbf{q}_i]$ and $[\mathbf{k}_r; \mathbf{k}_i]$. The Level 2 wave structure emerges from J -equivariant weight parameterization and phase-preserving modReLU, which phase-lock the two channels so that relative phase dictates constructive and destructive wave interference.

Single Barrier: Two-Slit Discrimination Plane

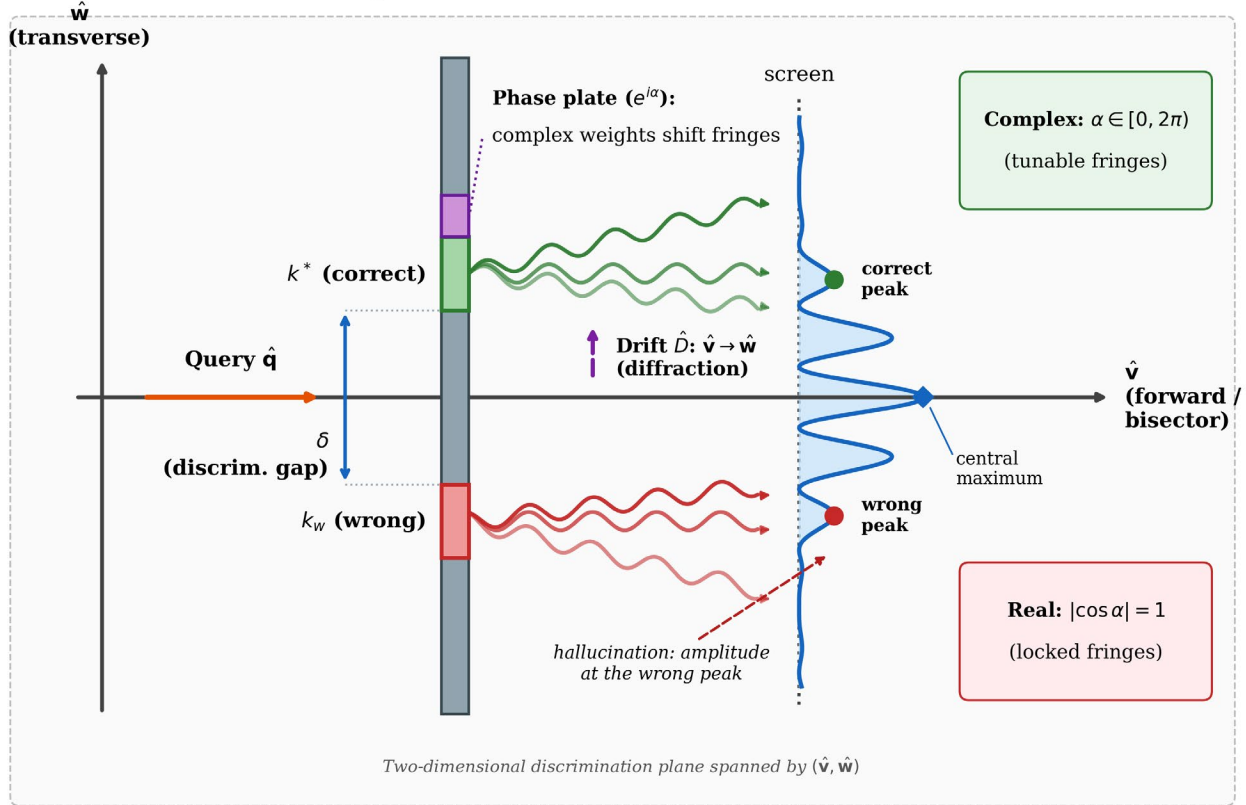


Figure 2: Two-slit discrimination at a single barrier. Detail of one layer in the 2D plane spanned by bisector $\hat{v}$ and transverse axis $\hat{w}$, separating the target key k^* and competitor k_w . A query positioned on the bisector is maximally ambiguous, requiring the drift operator $\hat{D}$ to generate transverse discrimination. Real weights restrict relative phase to $|\cos \alpha| = 1$ (locking interference fringes to two discrete configurations), whereas complex weights enable continuously tunable phase $\alpha \in [0, 2\pi)$ to dynamically steer intensity maxima onto k^* . Residual intensity at k_w represents candidate confusion.

6.2. Kinematics of the Subspace

The pairwise geometry of two-candidate discrimination carries an exact operator dynamics of motion. For any token pair $(|\psi_i\rangle, |\psi_j\rangle)$, the angle bisector $\hat{v}_{ij} = (\mathbf{a}_i + \mathbf{a}_j) / \|\mathbf{a}_i + \mathbf{a}_j\|$ defines a symmetry axis, while its orthogonal complement $\hat{w}_{ij}$ defines a transverse axis. In this basis, the embeddings decompose as $|\psi_i\rangle = a \hat{v}_{ij} + b \hat{w}_{ij}$ and $|\psi_j\rangle = a \hat{v}_{ij} - b \hat{w}_{ij}$, where $a = \cos(\theta_{ij}/2)$ is the projection along the symmetry axis and $b = \sin(\theta_{ij}/2)$ is the transverse component, with $\theta_{ij} = \arccos(\mathbf{a}_i \cdot \mathbf{a}_j)$. Any normalized state in this plane is parameterized by $|\psi\rangle = \cos \phi |\hat{v}_{ij}\rangle + e^{i\alpha} \sin \phi |\hat{w}_{ij}\rangle$, amplitude angle $\phi \in [0, \pi/2]$ and relative phase $\alpha \in [0, 2\pi)$.

Semantic alignment is governed by the orthogonal projection operator onto the bisector, $\hat{A}_{ij} = \hat{v}_{ij} \hat{v}_{ij}^\dagger$. Its expectation value equals the rescaled inner product $\langle \hat{A}_{ij} \rangle = \cos^2 \phi = S'_C(i, j)$. Because $\hat{A}_{ij}$ is a projector ($\hat{A}_{ij}^2 = \hat{A}_{ij}$), its variance is fixed entirely by its expectation value:

$$\sigma_{\text{align}}^2(i, j) = \langle \hat{A}_{ij}^2 \rangle - \langle \hat{A}_{ij} \rangle^2 = S'_C(i, j)(1 - S'_C(i, j)). \quad (8)$$

Within the two-dimensional discrimination plane, Hermitian operators are spanned by the identity and the Pauli operators $\hat{\sigma}_z = \hat{v} \hat{v}^\dagger - \hat{w} \hat{w}^\dagger$, $\hat{\sigma}_x = \hat{v} \hat{w}^\dagger + \hat{w} \hat{v}^\dagger$, and $\hat{\sigma}_y = -i(\hat{v} \hat{w}^\dagger - \hat{w} \hat{v}^\dagger)$, establishing an exact $\text{su}(2)$ Lie algebra where $\hat{A}_{ij} = \frac{1}{2}(\mathbb{1} + \hat{\sigma}_z)$.

Norm-preserving state motion within the plane is generated by the anti-symmetric drift operator $\hat{D}_{ij}(\lambda) = i\lambda(\hat{v}_{ij} \hat{w}_{ij}^\dagger - \hat{w}_{ij} \hat{v}_{ij}^\dagger) = -\lambda \hat{\sigma}_y$. Acting on the pairwise difference vector gives $\hat{D}_{ij}(|\psi_i\rangle - |\psi_j\rangle) = 2i\lambda b \hat{v}_{ij}$, converting transverse separation along $\hat{w}_{ij}$ into alignment along the bisector $\hat{v}_{ij}$. Drift therefore functions as a rotation generator that converts candidate distinctness into alignment. Commuting alignment with drift yields $[\hat{A}_{ij}, \hat{D}_{ij}] = i\lambda \hat{\sigma}_x \equiv i\lambda \hat{C}_{ij}$, where $\hat{C}_{ij} = \hat{\sigma}_x$ represents the shift operator; the complementary partner of alignment in the discrimination algebra.

6.3. The Optical Dictionary

The operator motion of the discrimination plane is realized identically in the polarization optics of a single classical light beam. A monochromatic optical beam with two transverse polarization components is described by a twocomplex-component vector $(E_v, E_w) = (\cos \phi, e^{i\alpha} \sin \phi)$. Its state space is the classical Poincaré sphere, and its observables are the four Stokes parameters:

$$S_0 = |E_v|^2 + |E_w|^2, \quad S_1 = |E_v|^2 - |E_w|^2, \quad S_2 = 2 \operatorname{Re}(\bar{E}_v E_w), \quad S_3 = 2 \operatorname{Im}(\bar{E}_v E_w). \quad (9)$$

These correspond numerically to expectations of $(\mathbb{1}, \hat{\sigma}_z, \hat{\sigma}_x, \hat{\sigma}_y)$.

Under this isomorphism, the semantic alignment operator $\hat{A}_{ij}$ represents an ideal linear polarizer, and its expectation $\langle \hat{A}_{ij} \rangle = \cos^2 \phi$ describes the transmitted optical intensity. The drift generator $\hat{D}_{ij}$ represents a rotator, relative phase α corresponds to the retardance of a wave plate, and the Poincaré sphere vector length specifies the degree of polarization. For a d -dimensional complex representation, this generalizes directly: the classical Stokes vector carries $\mathfrak{su}(d)$ as a classical Lie–Poisson algebra $\{S_a, S_b\} = (2/\hbar_{\text{sem}}) f_{abc} S_c$ with structure constants identical to quantum commutator relations. Non-commutativity in Level 2 models reflects classical precessional dynamics on the Poincaré sphere, requiring no quantum postulates.

6.4. Robertson Uncertainty and Semantic Projectors

It is tempting to look for a Robertson-type uncertainty relation $\Delta A \Delta B \geq \frac{1}{2} |\langle [\hat{A}, \hat{B}] \rangle|$ pairing alignment $\hat{A}$ with a conjugate observable. However, no such relation constrains ΔA , because $\hat{A}$ is a rank-one projector. For any state $|\psi\rangle = a|v\rangle + b|v_\perp\rangle$, the projector variance is Bernoulli: $\Delta A = |a||b|$. For any Hermitian operator $\hat{B}$, evaluating the commutator yields $\frac{1}{2} |\langle [\hat{A}, \hat{B}] \rangle| = |\operatorname{Im}(\bar{a}b \langle B_{v\perp} \rangle)| = \eta \Delta A |B_{v\perp}|$, where $\eta = |\operatorname{Im}(\bar{a}b \langle B_{v\perp} \rangle)| / (|a||b||B_{v\perp}|) \in [0, 1]$ is a pure phase factor. The commutator carries exactly the same state dependence as ΔA itself, so Robertson’s inequality reduces to $\Delta A \Delta B \geq \eta \Delta A |B_{v\perp}|$. Whenever $\Delta A \neq 0$, it cancels from both sides, leaving $\Delta B \geq \eta |B_{v\perp}|$, a statement about $\hat{B}$ alone. The relation places no lower bound on ΔA : as $\Delta A \rightarrow 0$, the right-hand side of Robertson’s inequality vanishes at the exact same rate, so the bound is satisfied trivially and never forbids sharp alignment.

Physically, this breakdown reflects the optical nature of Level 2: because a complex transformer represents a classical optical field whose complete state vector is deterministically accessible without measurement collapse, state uncertainties carry no epistemic content.

6.5. Wave Optics Interference Duality

While Level 2 hidden states do not face a quantum measurement principle, individual attention layers are strictly governed by classical wave-optical interference duality. Restricting the layer to the two-candidate subspace spanned by target k^* and competitor k_w , we collect the attention weights and the mutual coherence into the reduced coherency matrix $\mathbf{J} = \begin{pmatrix} w_{k^*} & c \\ \bar{c} & w_{k_w} \end{pmatrix}$, $\operatorname{Tr} \mathbf{J} = 1$, $c = \sqrt{w_{k^*} w_{k_w}} \gamma_{v^* v_w}$,

the semantic analogue of the optical coherency matrix (with weights w_{k^*}, w_{k_w} renormalized on this two-candidate subspace so $\operatorname{Tr} \mathbf{J} = 1$) whose Stokes components are $S_1 = w_{k^*} - w_{k_w}$, $S_2 = 2 \operatorname{Re} c$, and $S_3 = -2 \operatorname{Im} c$. Predictability P and fringe visibility V are the longitudinal and transverse Stokes magnitudes: $P = |S_1| = |w_{k^*} - w_{k_w}|$, $V = \sqrt{S_2^2 + S_3^2} = 2|c|$. Positivity of $\mathbf{J}$ requires $|c|^2 \leq w_{k^*} w_{k_w}$, which directly establishes interference duality:

$$P^2 + V^2 = (w_{k^*} - w_{k_w})^2 + 4|c|^2 \leq (w_{k^*} - w_{k_w})^2 + 4w_{k^*} w_{k_w} = (w_{k^*} + w_{k_w})^2 = 1, \quad (10)$$

with equality if and only if $|\gamma_{v^* v_w}| = 1$ (full mutual coherence).

In the context of language modeling, Predictability $P \in [0, 1]$ quantifies head decisiveness (the degree to which the attention head discriminates between target k^* and distractor k_w), while Fringe Visibility $V \in [0, 1]$ quantifies the phase interference capacity between their respective value states. The duality inequality $P^2 + V^2 \leq 1$ enforces a physical ceiling on transformer dynamics: when a head cleanly separates candidates ($P \rightarrow 1$), visibility is forced to zero ($V \rightarrow 0$) and phase interference plays no further role. Conversely, when candidate confusion is maximal ($P \approx 0$, a 50/50 weight split), the ceiling on fringe visibility reaches its theoretical maximum ($V \leq 1$), opening the maximum capacity for phase interference to act as a corrective mechanism.

6.6. Optical Coherence and Layer Dynamics

A major advantage of grounding Level 2 in classical optics is that established principles of optical coherence theory transfer directly to complex transformers, providing an analytical model for state evolution across deep layers. Softmax aggregation performs a convex,

contractive mixing of value vectors, causing a transformer stack to act as a partially coherent optical train: as layers superpose field components carrying a spread of relative phases, mutual fringe contrast degrades endogenously without requiring external noise.

By importing optical coherence theory, we derive the per-layer coherence transfer factor for a two-candidate attention distribution (see Appendix A3 for the full derivation), $\Delta\mu_{\text{passive}}^{(l)} = 2w_{k^*}^{(l)}w_{k_w}^{(l)}(1 - \text{Re}(\gamma_{v^*v_w}^{(l)}))$. Expressing the two-way softmax identity

$2w_{k^*}^{(l)}w_{k_w}^{(l)} = \frac{1}{2}(1 - P_l^2)$ in terms of predictability $P_l = |\tanh(\delta_{qw}^{(l)}/2)|$ yields the exact factorization:

$$\Delta\mu_{\text{passive}}^{(l)} = \underbrace{\frac{1}{2}(1 - P_l^2)}_{\text{unpredictability budget}} \times \underbrace{(1 - \text{Re}(\gamma_{v^*v_w}^{(l)}))}_{\text{value incoherence}} = \frac{1}{2}(V_l^{\text{max}})^2(1 - \text{Re}(\gamma_{v^*v_w}^{(l)})), \quad (11)$$

where $V_l^{\text{max}} = \sqrt{1 - P_l^2}$ is the maximum fringe visibility allowed by interference duality. Coherence loss factorizes cleanly into how undecided the layer is (V_{max}^2) and how incoherent the mixed value states are ($1 - \text{Re} \gamma$).

Under this optical formulation, complex representations decouple mutual value-state coherence $\text{Re} \gamma \in [-1, 1]$ from spatial vector geometry. While real models tie $\text{Re} \gamma$ to value cosines, complex models tune $\text{Re} \gamma$ via relative phase. Positive coherence preserves value mixtures intact, whereas negative coherence drives destructive interference to suppress competitors. Coherence is thus a consumable budget governed by Eq. (11): the factor $\frac{1}{2}(1 - P_l^2)$ restricts coherence expenditure to contested candidates ($P_l \rightarrow 0$), preventing indiscriminate field degradation across network depth.

6.7. Complex Transformer Architecture

Token and relation representations are extended to complex vectors $\mathbf{z} = \mathbf{x} + i\mathbf{y} \in \mathbb{C}^n$. Complex linear transformations compute matrix–vector products as $\mathbf{W}\mathbf{h} = (\mathbf{W}_r\mathbf{h}_r - \mathbf{W}_i\mathbf{h}_i) + i(\mathbf{W}_r\mathbf{h}_i + \mathbf{W}_i\mathbf{h}_r)$, sharing weight matrices $\mathbf{W}_r, \mathbf{W}_i \in \mathbb{R}^{n \times n}$. Expressed in real coordinates, this restricts the $2n \times 2n$ transformation matrix to the block structure $\begin{pmatrix} \mathbf{W}_r & -\mathbf{W}_i \\ \mathbf{W}_i & \mathbf{W}_r \end{pmatrix}$, which represents the commutant of the complex structure J .

When comparing real and complex models at matched representational width, where a complex model of dimension n and a real model of dimension $2n$ both use $2n$ real components per state, the complex parameterization acts as a restrictive inductive bias. The complex linear maps belong to $\text{GL}(n, \mathbb{C}) \subset \text{GL}(2n, \mathbb{R})$, reducing the real parameter count of linear layers from $4n^2$ to $2n^2$. Complex layers carry exactly half the parameters of matched real layers.

Nonlinear feed-forward sublayers use the phase-preserving modReLU activation function $\sigma_{\mathbb{C}}(z) = \text{relu}(|z| - b) \frac{z}{|z|}$ with bias $b = 0.5$. In the optical dictionary, modReLU acts as a saturable absorber followed by transparent wave propagation. Layer normalization is applied independently to real and imaginary components. Output logits are computed via a real linear functional logits = $\mathbf{W}_{\text{out}}[\mathbf{h}_r; \mathbf{h}_i] + \mathbf{b}_{\text{out}}$ with $\mathbf{W}_{\text{out}} \in \mathbb{R}^{V \times 2n}$, maintaining Level 2 field readout mechanics.

6.8. Hardware Realization on Photonic Integrated Circuits

Because Level 2 computation operates strictly within classical wave optics, complex transformers map directly onto room-temperature Photonic Integrated Circuits (PICs). Unlike Level 3 architectures, which require non-classical light, squeezed or entangled states together with photon-resolving detection, photonic hardware processes complex optical field amplitudes $E = Ae^{i\phi}$ natively. On coherent optical substrates, complex vector arithmetic is the native operational algebra, whereas real arithmetic is an artificially constrained subset. Linear transformations $\mathbf{W}\mathbf{h}$ execute passively at the speed of light through meshes of Mach–Zehnder Interferometers (MZIs) and optical phase shifters. By utilizing the phase degree of freedom $\phi \in [0, 2\pi)$, a $2n^2$ -parameter complex layer halves the physical footprint and weight-storage requirement compared to a matched real layer.

Treating optics as a physical platform for semantic computation transforms hardware characteristics into computational advantages, opening two complementary paths for hardware and architectural advancement. Where specific optical components, such as integrated phase-preserving non-linearities, are currently undergoing active development, their realization represents a straightforward engineering challenge driven by the recent demand for scaling LLMs on optical hardware. Because the standard transformer architecture represents an empirical construct instead of a derived fundamental law, physical optics provides a principled foundation for architectural refinement. For instance, global mathematical operations like explicit softmax normalization can be replaced by functionally equivalent waveoptical mechanisms, such as linear interference attention or Kramers–Kronig optical attention. Aligning transformer architectures with the native physics of optical wave propagation removes artificial digital bottlenecks, enabling ultralow-latency, room-temperature photonic semantic processors.

7. Level 3: Quantum Semantics

Level 2 exhausts the classical wave mechanics of language modeling. Level 3 adds exactly two components: normpreserving evolution generated by a Hermitian operator, and Born-rule measurement on a single copy.

We obtain Level 3 by canonically quantizing the Level 2 phase space. The semantic modes are not postulated: they are the normal modes of the learned free Hamiltonian, and the ladder operators $\hat{a}_k$ are the quantizations of the Level 2 complex amplitudes themselves. On the resulting Fock space the field quadratures obey the canonical commutation relation exactly, and the quantized nonlinear medium, whose classical counterpart is the cubic Kerr term of the Level 2 variant [15], is well defined [26]. A particular trained model realizes a finite-dimensional restriction of this space, in which the commutator is necessarily modified by a state-dependent factor, since no finite-dimensional representation of the canonical commutation relation exists. We therefore state the exact relations first and record the restriction, together with the regime in which it is negligible.

Each added component supplies one static and one dynamical structural limit, yielding four fundamental constraints with no Level 2 counterpart. From the Hermitian generator arise: (i) a static limit given by the semantic uncertainty principle $\Delta X_k \Delta P_k \geq \zeta \hbar_{\text{sem}}/2$ (where $\zeta = 1$ on the full mode space), which enforces a non-zero semantic zero-point energy $E_0 = \frac{1}{2} \hbar_{\text{sem}} \sum_k \omega_k$; and (ii) a dynamical limit given by the layer depth bound $L \geq \Theta/[(\tau/\hbar_{\text{sem}})\overline{\Delta E}_{\text{sem}}]$, setting the minimum network depth required to execute a semantic reorientation. From the measurement postulate arise: (iii) a static limit given by the readout trace ceiling $\sum_j \lambda_{\text{max}}(E_j) \leq D$, placing an exact capacity bound on token confidence; and (iv) a dynamical limit given by the entropic attention bandwidth relation $H(\alpha) + H(\hat{F}) \geq \log m$, trading spatial attention localization against positional-frequency spread.

Removing the Hermitian generator eliminates the former pair, while removing the measurement postulate eliminates the latter. None of these four limits survive at Level 2, where linear maps belong to arbitrary $\text{GL}(d, \mathbb{C})$ with no c -number commutator available, softmax aggregation is contractive, and the classical field is deterministically accessible without measurement collapse. Each limit is stated below and derived in the companion paper [26], with supporting equations collected in Appendix B.

7.1. Semantic Modes, the Uncertainty Principle, and Zero-Point Energy

Level 3 does not postulate a bosonic mode structure; it inherits one from Level 2, by the construction of Appendix B1. A Level 2 representation is a complex vector $\mathbf{z} \in \mathbb{C}^d$, that is, a classical field configuration. Writing its components in terms of real coordinates,

$$z_k = \frac{X_k + iP_k}{\sqrt{2\hbar_{\text{sem}}}}, \quad (12)$$

equips the Level 2 state space with the symplectic form $\Omega = \sum_k dX_k \wedge dP_k$ and the Poisson brackets $\{X_k, P_l\} = \delta_{kl}$, the same classical bracket structure already carried by the Stokes variables of Section 6.3. Canonical quantization, $\{\cdot, \cdot\} \mapsto [\cdot, \cdot]/i\hbar_{\text{sem}}$, directly yields

$$[\hat{X}_k, \hat{P}_l] = i\hbar_{\text{sem}} \delta_{kl} \implies \Delta X_k \Delta P_k \geq \frac{\hbar_{\text{sem}}}{2}, \quad (13)$$

by the Robertson argument. The ladder operators are derived objects,

$$\hat{a}_k = \frac{\hat{X}_k + i\hat{P}_k}{\sqrt{2\hbar_{\text{sem}}}}, \quad [\hat{a}_k, \hat{a}_l^\dagger] = \delta_{kl}, \quad (14)$$

so that $\hat{a}_k$ is precisely the quantization of the k -th Level 2 amplitude z_k , the number operator $\hat{n}_k = \hat{a}_k^\dagger \hat{a}_k$ counts quanta of semantic feature k , and the mode index k labels a normal mode of the learned free Hamiltonian $\hat{H}_0$, which is what makes the decomposition into independent modes available at all.

The coherent state $|\mathbf{z}\rangle$, defined by $\hat{a}_k|\mathbf{z}\rangle = (z_k/\sqrt{\nu})|\mathbf{z}\rangle$, reproduces the Level 2 field configuration in expectation. Level 2 is recovered in the classical limit $\bar{n}_k = |z_k|^2/\nu \gg 1$, where fractional quadrature fluctuations vanish as $\bar{n}_k^{-1/2}$. Here ν denotes the squared Level 2 amplitude carrying one quantum; thus, ν controls the onset of Level 3 corrections. At the convention $\nu = 1$, measuring GPT-2 residual vectors ($d = 768$, $\|\mathbf{h}\| = 239.3$) yields a mean occupation $\bar{n}_k \simeq 75$ and relative fluctuations $\bar{n}_k^{-1/2} \simeq 0.12$, placing typical representations in a many-quantum regime and confining non-classical corrections to low-amplitude modes.

The operators $\hat{X}_k$ and $\hat{P}_k$ represent the in-phase and out-of-phase field quadratures of the k -th semantic amplitude, as opposed to spatial coordinates. As in quantum optics, Eq. (13) forbids simultaneously specifying amplitude magnitude and phase with arbitrary precision. Substituting Eq. (12) into Eq. (13) cancels $\hbar_{\text{sem}}$, leaving a dimensionless noise lower bound $\Delta(\text{Re } z_k)\Delta(\text{Im } z_k) \geq \frac{1}{4}$. The absolute scale of $\hbar_{\text{sem}}$ is instead fixed dynamically by the action-perlayer ratio $\hbar_{\text{sem}}/\tau$ (Section 7.4)

Because the commutator in Eq. (13) is a c-number, its uncertainty bound is state-independent; a property Level 2 cannot provide. In a finite-dimensional model, this c-number commutator is modified by a state-dependent truncation factor ζ , yielding $\Delta X \Delta P \geq \zeta \hbar_{\text{sem}}/2$, where $\zeta \rightarrow 1$ away from state space boundaries.

This quadrature uncertainty directly constrains the system energy. Expressing the free Hamiltonian in normal modes, $\hat{H}_0 = \sum_k \hbar_{\text{sem}} \omega_k (\hat{n}_k + \frac{1}{2})$, bounds the ground-state energy away from zero:

$$E_0 = \frac{\hbar_{\text{sem}}}{2} \sum_k \omega_k, \quad \Delta X_k = \Delta P_k = \sqrt{\hbar_{\text{sem}}/2}. \quad (15)$$

The semantic vacuum saturates Eq. (13): the zero-point energy E_0 is the minimum quadrature uncertainty of the learned free Hamiltonian's normal-mode frequencies ω_k .

Operationally, Eq. (15) imparts an irreducible fluctuation to every semantic amplitude, introducing variance into any downstream readout. Level 3 structural under-confidence thus arises from two distinct mechanisms: the global POVM trace ceiling 7.3 and state preparation limits from zero-point fluctuations. Quadrature asymmetry permits semantic squeezing ($\Delta X_k < \sqrt{\hbar_{\text{sem}}/2}$ at the cost of increased ΔP_k), offering a mechanism to suppress quadrature noise along targeted semantic axes.

7.2. Attention Localization and Positional Bandwidth

Equation (13) constrains the semantic field. A second relation, of entropic as opposed to variance type, constrains the attention mechanism, and it arises from the measurement postulate instead of the generator.

On the positional index register $\mathcal{H}_{\text{idx}} = \mathbb{C}^m$, the attention distribution α_j represents the Born-rule outcome statistics of a measurement in the index basis $\{|j\rangle\}$. Its conjugate is the discrete Fourier basis $\{|p_k\rangle\}$, which is mutually unbiased to the index basis with $|\langle j | p_k \rangle| = m^{-1/2}$. Applying the entropic uncertainty relation yields (Appendix B2)

$$H(\alpha) + H(\hat{F}) \geq \log m, \quad (16)$$

where $H(\hat{F})$ is the outcome entropy in the positional-frequency basis. With attention sharpness defined as $\mathcal{S} = \log m - H(\alpha) \geq 0$, this bound simplifies to $H(\hat{F}) \geq \mathcal{S}$: an attention head that localizes sharply onto specific context tokens is necessarily spread over positional frequencies and cannot simultaneously act as a narrow-band spatial filter.

The bound is determined entirely by the context length m .

Sharpness is also paid for in measurement trials. With the gauge $\tilde{E}_j \geq 0$ of the companion construction, the postselection success probability $\bar{P}$ obeys the exact identity

$$\log \frac{1}{\bar{P}} = \mathcal{S} + 2\beta \langle \tilde{E} \rangle_\alpha \geq \mathcal{S}, \quad (17)$$

so the expected number of preparations grows at least as $e^{\mathcal{S}}$, reduced to $e^{\mathcal{S}/2}$ by amplitude amplification. Attention sharpness is traded against positional bandwidth by Eq. (16) and against measurement cost by Eq. (17).

7.3. The POVM Trace Ceiling

Under a Born-rule POVM readout $\{E_j\}_{j=1}^V$ with $\sum_j E_j = \mathbb{1}_D$, the maximum achievable probability for token j is $\max_{|\psi\rangle} P(v_j|\psi) = \lambda_{\max}(E_j) \leq \text{Tr } E_j$, giving the exact, parameter-free budget

$$\sum_{j=1}^V \lambda_{\max}(E_j) \leq D. \quad (18)$$

Here D is the dimension of the Hilbert space the circuit realizes, and is not the Level 2 width d . Quantizing d semantic modes gives a Fock space of infinite dimension, on which the canonical commutation relation holds exactly and on which this ceiling is vacuous; a trained circuit realizes a finite-dimensional restriction of it, with $D = (N + 1)^d$ for a truncation of each mode at occupation N and $D = d$ for a single-excitation encoding [26]. The ceiling is therefore a statement about a particular finite realization.

Expected accuracy is capped by the cumulative mass of the D most probable tokens, $\mathbb{E}[\text{accuracy}] \leq \sum_{j \in \mathcal{T}_D} p(j)$: frequent head entities may be predicted with high certainty, while rare tail entities remain structurally under-confident. This is the Level 3 form of the readout capacity floor whose Levels 1 and 2 counterpart is the softmax rank bottleneck of Section 5.3: the same counting argument in a different language.

7.4. The Depth Bound

Because Level 3 unitary dynamics are generated by a Hermitian Hamiltonian, state evolution is strictly governed by the quantum speed limit (Appendix B3). The rate of state reorientation $\Theta = \arccos |\langle \psi(0) | \psi(t) \rangle|$ is bounded by the energy spread, $d\Theta/dt \leq \Delta E/\hbar_{\text{sem}}$. Applying the layer–time correspondence ($t = L\tau$, where τ is the effective layer duration) yields a fundamental lower bound on network depth L :

$$L \geq \frac{\Theta}{(\tau/\hbar_{\text{sem}}) \overline{\Delta E}_{\text{sem}}}. \quad (19)$$

Equation (19) dictates that resolving a polysemous token whose contextualized sense lies at angular distance Θ from its base embedding requires a minimum layer depth proportional to Θ and inversely proportional to the mean logit spread $\overline{\Delta E}_{\text{sem}}$.

The architectural ratio $\varkappa_{\text{sem}} \equiv \hbar_{\text{sem}}/\tau$ defines the dimensionless semantic action per layer (nats per radian). Expressed per layer ℓ , Eq. (19) provides an empirical upper bound on $\varkappa_{\text{sem}}$:

$$\varkappa_{\text{sem}} \leq \min_{\ell} \frac{\overline{\Delta E}_{\text{sem}}^{(\ell)}}{\Delta \Theta_{\text{layer}}^{(\ell)}}, \quad (20)$$

where $\Delta \Theta_{\text{layer}}^{(\ell)}$ is the angular reorientation executed in layer ℓ . Because logit spread and angular rotation are directly observable across layers in trained models, $\varkappa_{\text{sem}}$ is an empirically measurable physical property, as opposed to an unconstrained free parameter.

Eq. (19) applies strictly to continuous Hamiltonian evolution. The non-unitary operations within the circuit, specifically postselection and density-matrix state aggregation, act as measurement-class jump processes. These operations reorient semantic states discontinuously without time cost, allowing individual layers to exceed the unitary speed limit at the cost of postselection trial attempts (Eq. (17)).

7.5. Quantum Capabilities Beyond Classical Wave Optics

The same quadrature structure that produces Eq. (13) permits capabilities physically impossible in classical wave optics. A few representative examples follow.

Semantic super-resolution. In classical optics fringe resolution is bounded by the wavelength, imposing a Rayleigh diffraction limit; at Level 2 attention fringe sharpness is correspondingly bounded by $\theta_{\min}$. In quantum optics N -particle entangled states act as a single entity of effective de Broglie wavelength λ / N , producing fringes N times sharper. Entangling context tokens therefore allows attention to resolve candidates lying within the classical resolution limit $\theta_{\min}$, addressing tail-entity confusion without expanding the state dimension.

Heisenberg-scaled long-context precision. Averaging N independent classical signals reduces noise at the standard quantum limit $O(N^{-1/2})$, the scaling of classical and Level 2 context aggregation and a source of degradation over long context windows. Entangled Level 3 aggregation achieves the Heisenberg limit $O(N^{-1})$, extracting weak semantic signals across large contexts with higher fidelity than any classical or complex field model.

8. Experiments

All experiments reported in this paper operate at Level 2. They establish that complex-valued transformers improve performance through wave interference.

Throughout our evaluation, a hallucination is defined operationally as predicting an entity that matches the expected object type for a given relation but is factually incorrect for the subject. Outcomes are partitioned into three mutually exclusive classes: correct predictions (accuracy), hallucinations, and type-invalid errors, which sum to one. All three metrics are reported together, as reducing candidate hallucinations by shifting errors into type-invalid outputs represents an error redistribution.

8.1. Experimental Setup

We evaluate knowledge graph completion on FB15k-237 [24], which contains 272,115 training triples spanning 14,541 entities and 237 relation types. Evaluation is conducted on the full test split of 20,466 triples using filtered ranking for tail prediction ($(h, r, ?)$), matching the autoregressive next-token prediction structure of language models. All models consist of two layers and four attention heads, trained for 100 epochs using AdamW (learning rate 10^{-3} , weight decay 10^{-4} , cosine decay schedule with warmup, and gradient clipping). Benchmark evaluations report means and sample standard deviations across ten independent random seeds; parameter ablations use five seeds to manage computational cost.

To evaluate complex wave computation, we compare against two real-valued models:

1. real-torch ($d=128$): The primary matched-width real baseline, using a standard PyTorch transformer encoder with 128 real components per state.
2. real-shared ($d=128$): An algebraic ablation using the complex block structure with imaginary weights set to zero ($W_i \equiv 0$), decoupling real and imaginary channels.

Models are compared at matched representational width. A complex model at dimension $d = 64$ and real baselines at dimension $d = 128$ both carry 128 real components per entity state. At matched width, complex linear transformations $GL(n, \mathbb{C}) \subset GL(2n, \mathbb{R})$ constrain parameter freedom, reducing layer weights by half. Including shared entity embedding tables (3.7M parameters), complex $d = 64$ (3.97M total, 247k non-embedding) carries 4.6% fewer total and 43.4% fewer non-embedding parameters than real $d = 128$ (4.16M total, 438k non-embedding).

8.2. Knowledge Graph Completion Results

Table 1 presents filtered tail prediction performance on FB15k-237 across ten random seeds. Complex transformers outperform the primary real baseline (real-torch, $d = 128$) across accuracy, Hits@10, Mean Reciprocal Rank (MRR), and hallucination reduction, despite utilizing 4.6% fewer total parameters.

Model	Params	Hallucin.	Accuracy	Other err.	Hits@10	MRR
Real-torch ($d=128$)	4.16M	0.7029 ± 0.0031	0.2773 ± 0.0022	0.0198 ± 0.0017	0.5431 ± 0.0028	0.3659 ± 0.0016
Real-shared ($d=128$)	4.16M	0.7106 ± 0.0022	0.2668 ± 0.0018	0.0227 ± 0.0012	0.5326 ± 0.0025	0.3556 ± 0.0017
Complex ($d=92$)	5.82M	0.6648 ± 0.0025	0.2858 ± 0.0018	0.0494 ± 0.0021	0.5548 ± 0.0021	0.3756 ± 0.0013
Complex ($d=68$)	4.23M	0.6756 ± 0.0029	0.2807 ± 0.0025	0.0436 ± 0.0021	0.5509 ± 0.0027	0.3707 ± 0.0020
Complex ($d=64$)	3.97M	0.6770 ± 0.0022	0.2800 ± 0.0020	0.0430 ± 0.0021	0.5501 ± 0.0024	0.3699 ± 0.0016

Table 1: FB15k-237 filtered tail prediction (ten seeds, full test split). Hallucination, accuracy, and type-invalid error sum to 1.0. Models compared at representational width $d = 64$ ($\mathbb{C}^d$, 128 real components) and matched real baselines $d = 128$ ($\mathbb{R}^{2d}$).

Because all models were evaluated across identical random seeds, we perform paired t-tests to evaluate performance differences against the primary baseline real-torch ($d = 128$) in Table 2, reporting the mean and standard deviation of the paired per-seed differences ($\Delta \pm SD_\Delta$).

Metric	Difference ($\Delta \pm SD_\Delta$)	t-statistic	p-value
Accuracy	$+0.0027 \pm 0.0021$	+4.06	0.0028
MRR	$+0.0040 \pm 0.0018$	+7.02	6.1×10^{-5}
Hallucination	-0.0259 ± 0.0034	-24.07	1.9×10^{-9}
Type-invalid	$+0.0232 \pm 0.0024$	+30.55	3.1×10^{-10}

Table 2: Paired t-test comparisons between complex ($d=64$) and primary real baseline (real-torch, $d=128$) across matching random seeds ($N = 10$, $df = 9$). Δ denotes the per-seed difference (Complex – Real).

Against real-torch, hallucination falls by 0.0259 while type-invalid errors rise by 0.0232: 89.6% of the hallucination reduction represents error redistribution accompanied by a net accuracy gain of +0.0027. We report this breakdown explicitly because the redistribution represents an operational benefit: type-invalid outputs violate relational entity types and are readily caught by downstream schema checks, whereas type-consistent false entities (hallucinations) silently bypass validation. Level 2 wave computation converts a large fraction of undetectable errors into schemadetectable errors while simultaneously improving ranking and accuracy.

8.3. Parameter Ablation Analysis

To confirm that trained imaginary parameters contribute meaningfully to model predictions, we perform parameter ablations across five seeds (Table 3). Each condition modifies 1,045,504 parameters (89.0% entity embeddings, 9.4% layer weights, 1.5% relation embeddings, 0.1% biases).

Condition	Hallucin.	Accuracy	Other err.
Original trained checkpoint	0.6763	0.2793	0.0443
Zero imaginary parameters	0.6070	0.0398	0.3532
Random imaginary (trained scale)	0.5969	0.0371	0.3660
Zero matched real (control)	0.7426	0.1344	0.1229

Table 3: Ablation analysis on complex ($d = 64$) across five seeds. Hallucination, Accuracy, and Other (Type-Invalid) errors sum to 1.0.

Zeroing imaginary parameters causes an 85.8% drop in model accuracy (from 27.93% to 3.98%). We note an apparent non-monotonicity in the hallucination metric under this severe ablation: the recorded hallucination rate drops from 0.6763 to 0.6070. This does not indicate improved factual generation. Because outcomes are partitioned into three mutually exclusive categories summing to 1.0, zeroing imaginary weights causes model outputs to collapse into ungrammatical, type-invalid entities (Type-invalid error increases eightfold from 0.0443 to 0.3532). The apparent reduction in hallucinations reflects a migration of errors from type-consistent false entities into type-invalid garbage. Similarly, randomizing imaginary weights at their trained scale causes an equivalent collapse (86.7% accuracy loss, 0.3660 type-invalid error), confirming that the imaginary channel carries a substantial share of the learned representation.

8.4. Direct Measurement of the Interference Mechanism

Direct instrumentation of trained checkpoints confirms that real and imaginary channels participate actively in attention scoring and readout.

Under isotropic initialization, any arbitrary two-channel decomposition yields a baseline magnitude split near $50.0\% \pm 0.1\%$. In trained complex attention sublayers, the imaginary interference term $\mathbf{q}_i \cdot \mathbf{k}_i$ accounts for 57.0% of the score magnitude $\text{Re}(\mathbf{q}^H \mathbf{k})$. Recomputing attention weights without the imaginary term shifts the attention distribution by 0.1817 in mean absolute probability weight $\frac{1}{m} \sum_j |\alpha_j - \alpha'_j|$ (which for a two-position head coincides with the total-variation distance and is bounded by 1.0).

At the output layer, the imaginary channel accounts for 52.4% of the net logit gap magnitude δ_{qw}^C . Removing imaginary components $\mathbf{h}_i$ at the readout layer alone (while preserving internal complex layers) reduces unfiltered top-1 accuracy from 13.67% to 11.44% (16% relative loss). Comparing this readout degradation against the 86% loss from removing the imaginary channel throughout the model that phase interference acts primarily within intermediate representations.

Table 4 evaluates the mutual value-state coherence $\text{Re}(\gamma)$, and confirms the decoupling predicted in Section 6.6. The real baseline is pinned by geometry: with γ fixed to the spatial cosine of the value vectors, the anisotropic embedding cone biases it positive, giving mean +0.0752 and only 36.9% negative. The complex model, whose $\text{Re}\gamma$ is set by relative phase, is centered on zero (+0.0048) with 49.0% negative, 1.33 times as often, and a markedly heavier destructive tail (−0.4182 against −0.2703 at the fifth percentile). The distribution is what the wave-optical account predicts: not a shift toward coherence preservation, but the removal of the geometric constraint that prevents a real layer from interfering destructively at all.

Model	Mean coherence	Fraction < 0	5th percentile
Complex ($d=64$)	+0.0048	49.0%	−0.4182
Real-shared ($d=128$)	+0.0752	36.9%	−0.2703

Table 4: Mutual value-state coherence $\text{Re}(\gamma)$ across the full FB15k-237 test split

9. Conclusion

In this paper, we established a first-principles field theory of semantic computation, shifting the foundation of language modeling from empirical black-box methods to physical wave mechanics. By separating complex linear dynamics from quantum measurement

postulates, we demonstrated that complex-valued transformers define an autonomous classical wave theory. We structured semantic computation into three exact mathematical levels: standard real transformers operating as phase-restricted subspaces (Level 1), complex transformers operating strictly as classical wave optics (Level 2), and quantum semantic fields obtained through canonical quantization of the Level 2 phase space without approximation (Level 3).

Grounding Level 2 in classical wave optics provides a rigorous physical framework for internal model dynamics. Reinterpreting attention scoring as a semantic multi-slit experiment and residual representations as polarized field transport opens the full analytical toolkit of physical electrodynamics. Concepts such as Poincaré polarization kinematics, wave-plate retardance, and explicit per-layer optical coherence budgets transform intermediate transformer mechanics from opaque statistical approximators into predictable wave phenomena. Within this framework, candidate confusion and hallucinations decompose into an unremovable readout capacity floor forced by vocabulary scaling bounds, an architectural gap caused by real-space phase restrictions, and an empirical gap reducible by training.

Beyond theoretical clarity, Level 2 wave computation creates a direct bridge to next-generation optical hardware. Because complex field operations execute passively and natively at the speed of light on coherent optical substrates, complex transformers map directly onto room-temperature Photonic Integrated Circuits (PICs). Aligning transformer architectures with native wave optics removes artificial digital bottlenecks, enabling ultra-low-latency photonic semantic processors. Empirically, Level 2 complex transformers achieve superior accuracy, Mean Reciprocal Rank, and hallucination reduction on knowledge graph completion benchmarks across ten independent random seeds while utilizing 4.6% fewer total parameters than matched real baselines, proving phase interference to be a high-capacity, parameter-efficient inductive bias.

Formalizing Level 3 via canonical quantization establishes the precise boundary where quantum mechanics governs semantic fields. Promoting classical field quadratures to operators obeying canonical commutation relations derives exact physical limits with no classical counterpart: a semantic zero-point energy, an entropic attention bandwidth relation, a POVM trace ceiling on token confidence, and a quantum speed limit on state reorientation. By replacing speculative quantum analogies with a rigorous quantization of classical wave fields, this work provides a unified, proven theoretical foundation for scaling physical wave architectures and photonic intelligence.

Appendix A: Derivations

1. Single-Competitor Bound

Let $P_{\text{hall}}(q) = \sum_{j \in \mathcal{K}_{\text{wrong}}} w_{qj} = A/(e^{s_{qk^*}} + A)$, where $A = \sum_{j \in \mathcal{K}_{\text{wrong}}} e^{s_{qj}}$. Since $f(A) = A/(e^{s_{qk^*}} + A)$ is strictly monotonically increasing in A , and $A \geq e^{s_{qk_w}}$ for any single competitor k_w , we have:

$$P_{\text{hall}}(q) \geq \frac{e^{s_{qk_w}}}{e^{s_{qk^*}} + e^{s_{qk_w}}} = \sigma(-\delta_{qw}). \quad (\text{A1})$$

Let $N_{\text{eff}}(q) = |\{j \in \mathcal{K}_{\text{wrong}} : \delta_{qj} \leq \delta_{qw}\}|$ count competitors as close as k_w . Monotonicity gives:

$$P_{\text{hall}}(q) \geq \frac{N_{\text{eff}}(q) e^{-\delta_{qw}}}{1 + N_{\text{eff}}(q) e^{-\delta_{qw}}} = \sigma(\log N_{\text{eff}}(q) - \delta_{qw}). \quad (\text{A2})$$

Equation (A1) bounds a probability mass, not an error frequency. The distinction is important: a bound on $\sum_j w_{qj}$ is compared correctly only against the measured conditional mass on the incorrect candidates, and a model may be wrong more or less often than its mass suggests whenever it is miscalibrated.

2. Gap Decomposition

Separating geometric and magnitude contributions in the attention score,

$$\delta_{qw} = \underbrace{\kappa \Delta \cos \theta_{qw}}_{\text{angular}} + \underbrace{\frac{\|\mathbf{q}\|}{\sqrt{d_k}} (\|\mathbf{k}^*\| - \|\mathbf{k}_w\|) \cos \theta_{qk_w}}_{\text{norm}}, \quad (\text{A3})$$

with $\kappa = \|\mathbf{q}\| \|\mathbf{k}^*\| / \sqrt{d_k}$. In a complex model, the score adds the interference channel of Eq. (7). The same decomposition applies at the output readout layer. Let $\mathbf{a}_c$ (for the target candidate k^*) and $\mathbf{a}_w$ (for the competitor distractor k_w) denote the unnormalized unembedding weight vectors of the readout layer. Writing $\cos \theta_c$ and $\cos \theta_w$ for their angles to the final hidden state $\mathbf{h}$, the net logit gap decomposes as:

$$\begin{aligned}\delta_{\text{logit}} &= \mathbf{h} \cdot \mathbf{a}_c - \mathbf{h} \cdot \mathbf{a}_w \\ &= \underbrace{\kappa_{\text{ro}} (\cos \theta_c - \cos \theta_w)}_{\text{angular}} + \underbrace{\|\mathbf{h}\| (\|\mathbf{a}_c\| - \|\mathbf{a}_w\|)}_{\text{norm}} \cos \theta_w,\end{aligned}\quad (\text{A4})$$

with $\kappa_{\text{ro}} = \|\mathbf{h}\| \|\mathbf{a}_c\|$. Two remarks are required before Eq. (A4) is evaluated. The readout carries no $1/\sqrt{d_k}$: that factor is the attention temperature and has no counterpart here, so κ_{ro} is set by the product of the two norms. Substituting the attention κ at the readout understates the angular channel by three orders of magnitude at GPT-2 scale. The split is not unique: taking $\|\mathbf{a}_w\|$ as the reference norm moves a term between the two channels. The identity closes under either convention and the sign of the angular channel follows the sign of $\Delta \cos \theta$ in both, but the individual magnitudes are convention-dependent and should be reported with the convention stated. We use $\kappa_{\text{ro}} = \|\mathbf{h}\| \|\mathbf{a}_c\|$ throughout.

3. Passive Decoherence

In a two-candidate approximation the coherence transfer factor is $\mu^{(l)} = (w_{k^*})^2 + (w_{k_w})^2 + 2w_{k^*}w_{k_w} \text{Re}(\gamma_{v^*v_w}^{(l)})$, giving Eq. (11). For distinct tokens $1 - \text{Re}(\gamma) \geq 1 - \cos \theta_{\min} = 2 \sin^2(\theta_{\min}/2)$ when the value coherence is positive. At a maximally contested head ($w_{k^*} = w_{k_w} = 1/2$), this establishes a strict per-layer coherence loss floor:

$$\Delta \mu_{\text{passive}}^{(l)} \geq \sin^2(\theta_{\min}/2). \quad (\text{A5})$$

We stress the exponent: the per-layer floor is second order in $\theta_{\min}$. Substituting $\theta_{\min}$ itself overstates the floor by $4/\theta_{\min}$, a factor of 59 for GPT-2. With lag-1 correlation ρ across layers, random-walk coherence loss yields effective depth $L_{\text{eff}} = \sqrt{L(1+\rho)/(1-\rho)}$.

Appendix B: Level-3 Formulae

This appendix collects the equations supporting the results stated in Section 7. Their purpose is to make these results checkable; full derivations are provided in the companion paper [26].

1. From the Level 2 Phase Space to Semantic Modes

This appendix supplies the operators $\hat{a}_k, \hat{a}_k^\dagger$ with a definition in terms of Level 2 quantities, so that Eq. (13) is derived rather than postulated.

A Level 2 state is a complex vector $\mathbf{z} \in \mathbb{C}^d$. Regarded as a classical field configuration, its natural phase-space coordinates are the quadratures of Eq. (12),

$$X_k = \sqrt{2\hbar_{\text{sem}}} \text{Re } z_k, \quad P_k = \sqrt{2\hbar_{\text{sem}}} \text{Im } z_k, \quad (\text{B1})$$

with symplectic form $\Omega = \sum_k dX_k \wedge dP_k = i\hbar_{\text{sem}} \sum_k dz_k \wedge d\bar{z}_k$ and Poisson brackets

$$\{X_k, P_l\} = \delta_{kl}, \quad \text{equivalently} \quad \{z_k, \bar{z}_l\} = -\frac{i}{\hbar_{\text{sem}}} \delta_{kl}. \quad (\text{B2})$$

This structure is classical and belongs to Level 2; it is the Lie–Poisson bracket of the Stokes variables written in canonical coordinates.

Canonical quantization promotes X_k, P_k to Hermitian operators with $\{\cdot, \cdot\} \mapsto [\cdot, \cdot]/i\hbar_{\text{sem}}$, giving Eq. (13) directly from Eq. (B2). With $\hat{a}_k$ defined by Eq. (14),

$$[\hat{a}_k, \hat{a}_l^\dagger] = \frac{1}{2\hbar_{\text{sem}}} \left(-i[\hat{X}_k, \hat{P}_l] + i[\hat{P}_k, \hat{X}_l] \right) = \delta_{kl}, \quad (\text{B3})$$

so the bosonic algebra is a consequence of Eq. (B2) and not an independent assumption. Inverting Eq. (14) returns the quadratures in the form used in the main text,

$$\hat{X}_k = \sqrt{\frac{\hbar_{\text{sem}}}{2}} (\hat{a}_k + \hat{a}_k^\dagger), \quad \hat{P}_k = -i\sqrt{\frac{\hbar_{\text{sem}}}{2}} (\hat{a}_k - \hat{a}_k^\dagger). \quad (\text{B4})$$

The correspondence with Level 2 is established by coherent states. Let $|\mathbf{z}\rangle$ satisfy $\hat{a}_k|\mathbf{z}\rangle = (z_k/\sqrt{\nu})|\mathbf{z}\rangle$, where ν represents the squared Level 2 amplitude carrying a single quantum. Then Eq. (B4) yields $\langle \mathbf{z} | \hat{X}_k | \mathbf{z} \rangle = \sqrt{2\hbar_{\text{sem}}/\nu} \operatorname{Re} z_k = X_k/\sqrt{\nu}$ (and similarly for $\hat{P}_k$), while quadrature uncertainties remain state-independent: $\Delta X_k = \Delta P_k = \sqrt{\hbar_{\text{sem}}/2}$.

Setting $\nu = 1$ reproduces the Level 2 classical field configuration exactly in expectation.

A Level 2 representation is therefore the expectation value of a Level 3 coherent state, and the classical wave limit (Level 2) corresponds precisely to the high-occupation limit of quantum fields (Level 3). Defining the mean mode occupation as $\bar{n}_k = |z_k|^2/\nu$, the fractional quantum fluctuations decay as $\Delta n_k/\bar{n}_k = \bar{n}_k^{-1/2} \rightarrow 0$. Thus, the departure of Level 3 quantum field dynamics from Level 2 classical wave optics is governed by the mode occupation number $\bar{n}_k$, rather than the scale parameter $\hbar_{\text{sem}}$ (which cancels out of dimensionless uncertainty ratios). This directly parallels how the photon occupation number per mode—rather than the dimensional units of field quadratures—governs the transition between quantum optics and classical Maxwell electrodynamics.

The mode decomposition requires that the free Hamiltonian be quadratic in the quadratures and positive definite; a linear canonical transformation then brings it to the normal form $\hat{H}_0 = \sum_k \hbar_{\text{sem}} \omega_k (\hat{n}_k + \frac{1}{2})$, and the semantic modes k are by definition its normal modes. A learned $\hat{H}_0$ that is not quadratic admits no global mode structure, and Eqs. (13) and (15) then hold only in the harmonic approximation about a stable configuration, with ω_k the curvatures of $\hat{H}_0$ at that configuration. Anharmonic corrections belong to the nonlinear medium of the quantum semantic circuit and are treated in [26].

Equation (15) follows at once. At the energy minimum, where $\langle \hat{X}_k \rangle = \langle \hat{P}_k \rangle = 0$, the single-mode Hamiltonian $\hat{H}_k = \frac{\omega_k}{2} (\hat{X}_k^2 + \hat{P}_k^2)$ obeys

$$\langle \hat{H}_k \rangle = \frac{\omega_k}{2} (\Delta X_k^2 + \Delta P_k^2) \geq \omega_k \Delta X_k \Delta P_k \geq \frac{\hbar_{\text{sem}} \omega_k}{2}, \quad (\text{B5})$$

using the arithmetic–geometric mean inequality and then Eq. (13). Summing over the normal modes gives Eq. (15). Equality requires $\Delta X_k = \Delta P_k = \sqrt{\hbar_{\text{sem}}/2}$, so the semantic vacuum saturates Eq. (13) and is the $\mathbf{z} = 0$ coherent state: the zero-point energy and the quadrature uncertainty are the same statement. Squeezed states satisfy $\Delta X_k < \sqrt{\hbar_{\text{sem}}/2}$ with ΔP_k correspondingly enlarged, and no longer saturate Eq. (B5).

2. The Attention State and Its Entropy

Let m be the number of attended context positions, $\mathcal{H}_{\text{idx}} = \mathbb{C}^m$ the index register with basis $\{|j\rangle\}$, and $\tilde{E}_j = E_j + C$ the attention energy shifted so that $\min_j \tilde{E}_j = 0$. The companion construction initializes the register in the uniform superposition together with a flag ancilla, applies the multiplexed rotation $\sum_j |j\rangle\langle j| \otimes R_y^{(f)}(\theta_j)$ with $\theta_j = 2 \arccos(e^{-\beta \tilde{E}_j})$, which is unitary because it is block diagonal with blocks in $\text{SU}(2)$, and postselects the flag on $|0\rangle_f$, collapsing the index register into

$$|\alpha\rangle = \sum_{j=1}^m \sqrt{\alpha_j} |j\rangle, \quad \alpha_j = \frac{e^{-2\beta \tilde{E}_j}}{Z'}, \quad Z' = \sum_{j=1}^m e^{-2\beta \tilde{E}_j}, \quad (\text{B6})$$

with success probability $\bar{P} = Z'/m$. Two features follow from the collapse: the inverse temperature appears as 2β because the Born rule squares the flag amplitude, so the rescaling $\beta \rightarrow 2\beta$ is derived; and the normalization Z' is supplied by the postselection itself, the only classical information extracted being a herald bit.

Equation (B6) makes α the Born outcome statistics of a measurement in the index basis. Consider the discrete Fourier basis, $|p_k\rangle = m^{-1/2} \sum_{j=0}^{m-1} e^{2\pi i j k/m} |j\rangle$, which is mutually unbiased to the index basis, giving $c = \max_{j,k} |\langle j | p_k \rangle| = m^{-1/2}$. The entropic uncertainty relation for mutually unbiased bases [27] then gives

$$H(\alpha) + H(\hat{F}) \geq 2 \log(1/c) = \log m, \quad (\text{B7})$$

with $H(\alpha) = -\sum_j \alpha_j \log \alpha_j$ and $H(\hat{F})$ the outcome entropy in the Fourier basis, equivalently $H(\hat{F}) \geq \mathcal{S}$ with $\mathcal{S} = \log m - H(\alpha)$. Only mutual unbiasedness is required. The Fourier basis is the eigenbasis of the cyclic positional shift $\hat{T} = \sum_j |j+1 \bmod m\rangle\langle j|$, which motivates calling the conjugate variable a positional frequency; the bound does not depend on that identification and is unaffected by the choice of boundary condition on the context window.

This distribution directly constrains the measurement cost. From Eq. (B6), $H(\alpha) = \log Z' + 2\beta\langle\tilde{E}\rangle_\alpha$ with $\langle\tilde{E}\rangle_\alpha = \sum_j \alpha_j \tilde{E}_j$, so with $\bar{P} = Z'/m$,

$$\log \frac{1}{\bar{P}} = \mathcal{S} + 2\beta\langle\tilde{E}\rangle_\alpha. \quad (\text{B8})$$

The gauge $\tilde{E}_j \geq 0$ makes the second term non-negative, giving $\bar{P}^{-1} \geq e^{\mathcal{S}}$; amplitude amplification [28] reduces the required rounds to $O(\bar{P}^{-1/2})$, bounded below by $e^{\mathcal{S}/2}$. Uniform attention gives $\mathcal{S} = 0$ and $\bar{P} = 1$; maximal sharpness gives $\bar{P} \rightarrow 1/m$.

3. The Depth Bound

For evolution generated by a Hermitian $\hat{H}(t)$, the state overlap angle $\Theta(t) = \arccos|\langle\psi(0)|\psi(t)\rangle|$ satisfies [29]

$$\frac{d\Theta}{dt} \leq \frac{\Delta E(t)}{\hbar} \implies t \geq \frac{\hbar \Theta}{\overline{\Delta E}}, \quad (\text{B9})$$

with $\overline{\Delta E}$ the time-averaged energy spread. The bound involves the spread and not the mean, so an energy eigenstate does not evolve at all. Substituting $t = L\tau$ and the gauge $\hbar = \hbar_{\text{sem}}$ gives Eq. (19). Unlike a Robertson bound between two constructed observables, the right-hand side of Eq. (B9) is a c-number and cannot degenerate on a particular state. Equation (B9) is proved for linear Schrödinger evolution and is applied here to the linear segments of a layer, where the bulk of the accumulated rotation occurs.

4. Finite-Dimensional Restrictions

A trained model realizes a finite-dimensional restriction of the mode space of Appendix B1. The restriction is not innocuous, and we record the form the corrections take.

No finite-dimensional Hilbert space admits an exact representation of Eq. (13). If $\hat{X}, \hat{P}$ are operators on $\mathbb{C}^D$ with $[\hat{X}, \hat{P}] = i\hbar_{\text{sem}}\mathbb{1}_D$, taking the trace gives $0 = \text{Tr}[\hat{X}, \hat{P}] = i\hbar_{\text{sem}}D \neq 0$, a contradiction. Exact state-independence therefore requires the full mode space, and every finite realization replaces the c-number commutator by an operator whose expectation value depends on the state. The question is not whether a correction appears but how large it is.

Restricting each mode to occupations $n \leq N$ replaces $\hat{a}$ by $\hat{a}_N = \sum_{n=0}^{N-1} \sqrt{n+1} |n\rangle\langle n+1|$, for which

$$[\hat{a}_N, \hat{a}_N^\dagger] = \mathbb{1} - (N+1)|N\rangle\langle N|, \quad (\text{B10})$$

so that the quadratures of Eq. (B4) obey $[\hat{X}_k, \hat{P}_k] = i\hbar_{\text{sem}}(\mathbb{1} - (N+1)|N\rangle\langle N|)$ and

$$\Delta X_k \Delta P_k \geq \frac{\hbar_{\text{sem}}}{2} \zeta, \quad \zeta = |1 - (N+1)p_N|, \quad p_N = |\langle N | \psi \rangle|^2. \quad (\text{B11})$$

The correction term in Eq. (B10) is precisely what makes the trace of the commutator vanish, as the preceding argument requires. The departure from Eq. (13) is governed entirely by p_N , the population of the highest retained level: $\zeta = 1$ when that level is unoccupied, and the bound weakens only once p_N becomes comparable to $1/(N+1)$, vanishing at $p_N = 1/(N+1)$. This is the restriction relevant to Eq. (13), since it acts on the same operators, and the occupations recorded in Section 7.1 place trained representations far from its boundary.

A single-excitation encoding makes a different conjugate pair available, a coordinate on the basis together with its shift generator, and these are not the quadratures. There the correction factor measures the coherence between neighbouring basis amplitudes: it approaches unity for states varying slowly across the basis and vanishes for a state supported on a single basis direction, so the bound degenerates precisely in the regime of a confident single-token prediction. The two situations remain distinct in kind. At Level 2 the bound degenerates for every choice of conjugate observable, because the alignment operator is a rank-one projector and the commutator carries its state dependence identically; here the degeneracy is a property of the encoding and is removed by enlarging it. That construction, its zero-point energy, and the reporting of the relevant factor alongside any measured Level 3 uncertainty are given in [26].

Appendix C: Measurements on GPT-2

To empirically ground our theoretical analysis, we performed diagnostic measurements on GPT-2 Small ($L = 12$, $d = 768$, $H = 12$, V

= 50,257) across 50 evaluated query prompts. All angular and logit quantities are computed on the residual stream after the final layer normalization lnf, matching the exact state that produces the output logits. For each query, channels are evaluated against the contending competitor, defined as the incorrect token of highest logit.

An exhaustive sweep over all 1.26×10^9 token pairs in the GPT-2 embedding space yields a maximum pairwise similarity of 0.99774, corresponding to a semantic resolution limit of $\theta_{\min} = 0.0673$ radians (3.86°). The average pairwise cosine similarity is 0.268 ± 0.053 , reflecting anisotropic embedding cone collapse. The measured norm of the final readout state averages $\|\mathbf{h}\| = 239.3$. At this norm, assuming unit-normalized unembeddings, the maximum geometric logit gap $\delta_{\max} = 2\|\mathbf{h}\|\sin(\theta_{\min}/2) = 16.1$ implies a non-binding geometric error floor of $\sigma(-16.1) \approx 1 \times 10^{-7}$, confirming that fundamental error limits reside in readout capacity as opposed to embedding geometry.

Decomposing the readout logit gap into angular and norm channels reveals strong cancellation mechanics. Across the query set, the angular channel contributes +1.810 logits toward the correct candidate, while the norm channel opposes it by -1.447 logits, yielding a net mean logit gap of $\bar{\delta}_{\text{logit}} = +0.363$ logits. The angular channel serves as the primary driver of candidate selection, operating against an opposing norm baseline. Layer-wise trajectory analysis shows that angular alignment degrades monotonically through layers 1 to 8 before reversing and building positive discrimination in layers 9 to 11. Input embeddings initially favor the distractor by -0.397 logits, requiring intermediate transformer layers to generate +0.760 logits of net discrimination to achieve a positive output gap.

Evaluating the single-competitor lower bound $\sigma(-\delta_{\text{logit}})$ per query yields an expected hallucination probability mass prediction of $\mathbb{E}_q[\sigma(-\delta_q)] = 0.456$, compared against a measured conditional probability mass of 0.534 on incorrect candidates. (Evaluating the sigmoid at the mean gap yields $\sigma(-\bar{\delta}_{\text{logit}}) = 0.410$; the per-query expectation is 0.456 due to non-linear averaging across queries spanning positive and negative gaps).

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. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12), 1-38.
3. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM transactions on information systems*, 43(2), 1-55.
4. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 793, 9459-9474.
5. Yang, Z., Dai, Z., Salakhutdinov, R., & Cohen, W. W. (2018). Breaking the softmax bottleneck: A high-rank RNN language model, *6th International Conference on Learning Representations*.
6. Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.
7. Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.
8. Ethayarajh, K. (2019). How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 55-65. 2019.
9. Gao, J., He, D., Tan, X., Qin, T., Wang, L., & Liu, T. Y. (2019). Representation degeneration problem in training natural language generation models. *7th International Conference on Learning Representations*.
10. Aerts, D., & Czachor, M. (2004). Quantum aspects of semantic analysis and symbolic artificial intelligence. *Journal of Physics A: Mathematical and General*, 37(12), L123-L132.
11. Widdows, D., & Peters, S. (2003). Word vectors and quantum logic: Experiments with negation and disjunction. *Mathematics of language*, 8(141-154).
12. Widdows, D., & Widdows, D. (2004). *Geometry and meaning* (Vol. 773). Stanford: CSLI publications.
13. Laine, T. A. (2026). Structural Isomorphism Between LLM Embedding Spaces and Quantum Mechanical Systems. *OA J Applied Sci Technol*, 4(2), 01-18.
14. Laine, T. A. (2026). Quantum hierarchy for Understanding LLM Representations by Modeling Linear Projections and Nonlinear Dynamics. *OA J Applied Sci Technol*, 4(1), 01-43.
15. Laine, T. A. (2026). Quantum-Native Transformer Architecture: Interference, Partition Functions, and Nonlinear Schrödinger Dynamics, *OA J Applied Sci Technol*, 4(2), 01-28, 2026.
16. 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.

-
17. Hirose, A. (2006). *Complex-valued neural networks*. Berlin, Heidelberg: Springer Berlin Heidelberg.
 18. Bassey, J., Qian, L., & Li, X. (2021). A survey of complex-valued neural networks. *arXiv preprint arXiv:2101.12249*.
 19. Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016, June). Complex embeddings for simple link prediction. In *International conference on machine learning* (pp. 2071-2080). PMLR.
 20. Sun, Z., Deng, Z. H., Nie, J. Y., & Tang, J. (2019). Rotate: Knowledge graph embedding by relational rotation in complex space, *7th International Conference on Learning Representations*.
 21. Yang, M., Ma, M. Q., Li, D., Tsai, Y. H. H., & Salakhutdinov, R. (2020, May). Complex transformer: A framework for modeling complex-valued sequence. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4232-4236). IEEE.
 22. Eilers, F., & Jiang, X. (2023, June). Building blocks for a complex-valued transformer architecture. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.
 23. Arjovsky, M., Shah, A., & Bengio, Y. (2016, June). Unitary evolution recurrent neural networks. In *International conference on machine learning* (pp. 1120-1128). PMLR.
 24. Toutanova, K., & Chen, D. (2015, July). Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd workshop on continuous vector space models and their compositionality* (pp. 57-66).
 25. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.
 26. Laine, T. A. The Quantum Semantic Circuit: Hamiltonian Generators and Path-Integral Formulation, companion paper, in preparation.
 27. Maassen, H., & Uffink, J. B. (1988). Generalized entropic uncertainty relations. *Physical review letters*, 60(12), 1103- 1106.
 28. Brassard, G., Hoyer, P., Mosca, M., & Tapp, A. (2002). Quantum amplitude amplification and estimation. *Contemporary Mathematics*, 305, 53–74.
 29. Mandelstam, L., & Tamm, I. G. (1945). The uncertainty relation between energy and time in non-relativistic quantum mechanics. *Journal of Physics (USSR)*, 9, 249–254.

Copyright: ©2026 Timo Aukusti Laine. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.