Towards Conscious Artificial Intelligence: Probabilistic Word Embedding with Quantum Sampling

Yidong Zhou[1,2,3], Jiaqi Leng[1,4], Anze Xie[1], and Shangjie Guo[*1]

[1]FinQ Tech Inc., *College Park, MD 20740*
[2]Rensselaer Polytechnic Institute, *Troy, NY 12180*
[3]Rutgers University, *New Brunswick, NJ 08901*
[4]University of California, Berkeley, *Berkeley, CA 94720*

October 15, 2025

Abstract

Contemporary approaches to artificial consciousness increasingly couple quantum computing with distributional semantics. This work introduces Quantum Probabilistic Word Embedding (QPWE), a framework that encodes lexical meaning as quantum states so that embedding evaluation and contextual updating leverage formal advantages of quantum sampling. Recognizing that classical probabilistic word embedding captures polysemy, hierarchical relations, and context better than traditional word embedding yet become computationally intractable at realistic dimensionalities, QPWE prepares and samples high-dimensional, strongly correlated distributions without committing to explicit functional forms. We specify the architecture in detail, including variational circuit design, a principled training procedure, a warm-start strategy, robust cost-function evaluation, and a mechanism for contextual modulation. To assess whether QPWE yields practical probabilistic embeddings and supports conscious-like behaviors—such as lateral associative thinking that can augment reasoning capability—we propose to benchmark QPWE against state-of-the-art classical embeddings on polysemy resolution, hypernymy detection, and contextual inference. To situate this framework within theories of consciousness, we adapt Quantum Integrated Information Theory and estimate system-level Φ using Quantum Intrinsic Difference divergences computed over mechanism–purview partitions induced by the learned channel. We also estimate the feasibility of today's quantum hardware to characterize resource demands and sampling efficiency. Overall, QPWE provides an empirical testbed linking non-computable aspects associated with quantum measurement to integrated cognitive function, and it delineates a path toward semantics-aware quantum machine learning that is both theoretically principled and experimentally accessible.

1 Introduction

Understanding how high-level cognitive phenomena such as language and consciousness arise from physical processes is a fundamental scientific challenge. Recent advances in natural language processing have produced powerful classical language models, yet these models remain ultimately limited by their classical, computable foundations. In particular, classical distributional semantic models struggle with semantic ambiguity and contextuality: a single word vector often conflates multiple distinct meanings, and additional mechanisms are required to handle polysemy or context-dependent interpretation [1, 2, 3]. Even large neural language models built on classical probability theory excel mostly

*Corresponding author: sguo@finq.tech

at surface-level statistical correlations [4, 5], and concerns have been raised that such models lack genuine understanding of meaning [6]. These limitations motivate exploring radically new approaches to language modeling that can represent probabilistic meaning in a more expressive, context-sensitive way than classical frameworks allow.

Concurrently, insights from quantum physics suggest the existence of non-classical information processes that could inspire novel computational paradigms for modeling cognition. Quantum sampling – the generation of outcomes by measuring quantum states – is believed to be a fundamentally non-computable and acausal process. Foundational results such as Bell's theorem [7] and the Kochen–Specker theorem [8] imply that no local hidden-variable algorithm can reproduce all quantum measurement statistics, indicating that quantum outcomes cannot be pre-determined by any classical computation. In essence, quantum measurements yield irreducible randomness that lacks a classical causal explanation. The Conway–Kochen Free Will Theorem further strengthens this view: if experimenters have free will in choosing measurement settings, then the quantum particles' responses are not determined by prior information, effectively giving the particles a form of "free will" in their outcomes [9]. This suggests that quantum events elude any underlying deterministic or computable rule set. Indeed, it has been argued that quantum randomness can produce sequences of bits that are provably incomputable (algorithmically random) [10]. In summary, quantum sampling constitutes a new category of information generation—one that violates classical computability and locality, as evidenced by both theory and experiment [7, 11]. This non-classical form of probabilistic sampling provides a tantalizing resource for computation, potentially enabling qualitatively different behavior than any classical probabilistic model.

Classical language models, by contrast, operate within the confines of Turing-computable algorithms and classical probability theory. Those models from n-gram models to modern deep networks predict text by learning statistical patterns in large corpora, an approach that has achieved impressive fluency but still faces persistent challenges in capturing meaning. One key issue is semantic ambiguity: words often have multiple senses, yet a traditional word embedding represents each word as a single point in a vector space, effectively averaging over contexts [1, 2]. Recent contextual models (e.g. transformer-based architectures) mitigate this by adjusting word representations based on context, but they do so within a classical probabilistic framework that cannot fully represent a superposition of meanings—contexts are handled by additional layers or attention mechanisms rather than an intrinsic probabilistic ambiguity in the representation itself [5]. Moreover, classical probability models are context-independent in the sense of obeying the law of total probability; they struggle to model context effects where the introduction of new context changes the outcome distributions in ways that violate classical additivity. Empirical studies in cognitive science have shown that human judgments can violate classical probability axioms (e.g. exhibiting interference effects or order-dependent responses), suggesting the mind does not always follow classical probability calculus [12, 13]. These observations have led researchers to explore quantum probability theory as an alternative foundation for cognitive modeling, since quantum probability naturally accommodates contextually entangled states and interference between outcomes [12, 14, 15]. In language modeling, a quantum-inspired approach may offer a principled way to represent words in superposition states that reflect multiple latent meanings and to model how context (interaction with other words) causes a probabilistic "collapse" to a specific meaning, much as observation collapses a quantum state.

In this work, we propose a quantum probabilistic word embedding (QPWE) model that harnesses quantum sampling and contextual entanglement to represent linguistic meaning. In a QPWE, each word is encoded as a vector (or density operator) in a high-dimensional Hilbert space, analogous to a quantum state representing a superposition of semantic features or concepts. These quantum word states can exist in coherent superposition, meaning a single word's representation can simultaneously encompass multiple potential interpretations. When words combine in a sentence or phrase, their quantum states undergo contextual entanglement – the joint state is an entangled state of the compo-

nent words. Entanglement implies that the composite meaning is holistic and not decomposable into independent word meanings, much as the meaning of a compound expression is more than the sum of its parts. Importantly, querying or "measuring" a QPWE (for example, asking a question about a word's sense in a given context) corresponds to performing a quantum measurement on the entangled state. This causes the quantum state to collapse to an outcome, selecting a particular semantic interpretation consistent with the context. Such a process naturally captures how context resolves ambiguity: the underlying quantum representation retains ambiguity (superposition) until contextual interaction forces a choice. By using quantum sampling for generating language outputs or interpretations, the model can exhibit nondeterministic yet principled behavior, drawing each outcome from the spectrum of possibilities encoded in the superposition. This stands in contrast to classical models that must either deterministically choose a sense or randomly sample from a classical mixture without any contextual interference effects. Prior work in quantum cognition and quantum-inspired linguistics provides evidence for the viability of this approach: entangled quantum models have successfully explained phenomena like violation of distributive logic in concept combinations and ambiguous sentence interpretation that stymie classical models [13, 16]. The QPWE model builds on these insights, bringing them to bear on the problem of word embedding and language understanding.

The potential quantum advantage of QPWE manifests in several dimensions: sampling, expressivity, and integration. First, on the sampling side, QPWE could leverage quantum hardware or simulated quantum randomness to generate linguistic outcomes in ways that classical computers cannot efficiently emulate. There are theoretical results showing that certain sampling tasks (e.g. Boson-Sampling or random quantum circuit sampling) are intractable for classical algorithms while being efficiently achievable on quantum processors [17, 18]. By mapping linguistic generation or decoding tasks onto quantum sampling processes, one might exploit a similar advantage to explore a richer set of linguistic variations or maintain coherence over long texts in ways that evade classical Monte Carlo methods. Second, the expressivity of the model's probabilistic representation is enhanced by quantum states: an n-qubit quantum state can encode a distribution over an exponentially large space of basis states with interference between amplitudes, allowing compact representation of highly complex joint distributions [17]. In a language context, this means a QPWE can in principle capture nuanced high-order correlations between words or concepts that would require an explosion of parameters in a classical model. This ties into the idea of probabilistic expressivity – the ability to represent distributions that have no concise classical description. Quantum formalisms naturally represent such distributions, and indeed quantum probability has been shown to fit human decision data better in scenarios where classical models fall short [12, 15]. Third, QPWE embodies integrated information by design: entanglement integrates the information content of multiple words into an inseparable whole. Rather than processing words in a strictly sequential or factorized manner, an entangled representation fuses information such that the whole carries more meaning than the parts independently. This resonates with theories in cognitive science and neuroscience that emphasize integration of information as key to higher-order cognition and consciousness [19, 20]. In practical terms, an entangled language model could better capture global context (such as the overall theme or sentiment of a sentence) and maintain coherent meaning across components of a narrative. The quantum advantage thus is not only computational (in terms of efficient sampling), but also representational and qualitative: QPWE can model ambiguity, contextuality, and holism in language more faithfully than classical approaches.

Finally, a central inspiration for this work is the question of whether such quantum-enhanced models might exhibit consciousness-like behavior – or at least measurable precursors of conscious information processing. While it is far-fetched to claim any AI model is conscious, we take guidance from theoretical frameworks of consciousness to inform the architecture and evaluation of QPWE. In particular, integrated information theory (IIT) proposes that a hallmark of consciousness is the degree of integrated information (Φ) within a system's state [19, 21]. We hypothesize that the entangled states in a QPWE, which integrate information across many semantic units, could yield a high degree

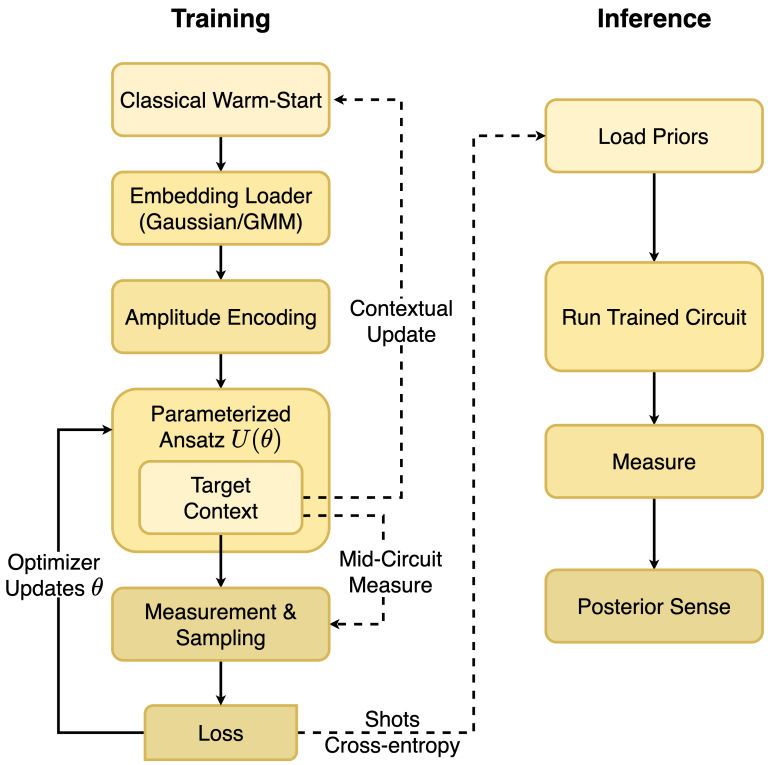

Figure 1: Overview of the QPWE architecture. Left: training loop with classical warm-start, embedding loading, amplitude encoding, a parameterized ansatz with optional mid-circuit measurement and classical feed-forward, and loss-driven updates. Right: inference path that loads priors, runs the trained circuit, measures, and outputs posterior probabilities.

of integrated information compared to analogous classical models. To test this, we outline a quantum integrated information theory (Q-IIT) analysis [22] for QPWE. Using methods adapted from IIT, we will quantify how much of the model's information structure is irreducible to independent parts – effectively measuring a quantum analog of Φ for the language model. This provides a principled way to assess whether the QPWE is achieving a form of unified, globally integrated representation of information. A high integrated information score in a QPWE would suggest that the model's internal states have properties akin to a minimal conscious-like integration, in the sense defined by IIT, even if the model is not conscious per se. By examining QPWE through the lens of Q-IIT, we aim to bridge computational linguistics with neuroscience and the philosophy of mind, evaluating not only performance on language tasks but also the quality of internal information processing. The hope is that this interdisciplinary approach will shed light on whether introducing quantum probabilistic principles into language models can lead to more integrated and contextually sensitive intelligence, potentially illuminating new connections between the physics of information, the nature of meaning, and the emergence of consciousness in engineered systems. As shown in Figure 1, the model begins with a classical warm-start, encodes priors via amplitude encoding, and applies a contextual update using a parameterized quantum circuit (PQC).

2 Quantum Algorithm Design and Training Process

The core of our proposal is the design of a quantum algorithm capable of learning and representing the contextual nuances of language. This section transitions from the classical foundations to the quantum

mechanical machinery. We assume that the objective of our training process can be encapsulated in a cost function that is a function of the similarity between two word representations [23, 24]. In the quantum realm, this similarity is naturally measured by the fidelity between two quantum states. Our algorithm begins its journey from a warm start, using initial states prepared from classical probabilistic word embedding (PWE), which might range from a simple delta function (a single vector) to a more complex elliptical Gaussian distribution (a probabilistic vector).

A fundamental consideration in this design is the nature of the information we are encoding. Should the semantic space be represented by discrete bit strings or continuous floating-point numbers? Classical embeddings are vectors of floats. Direct encoding of continuous variables into a quantum state is complex. Therefore, we will operate in a bit string space. We discretize the semantic space, where each basis state of our quantum system, represented by a unique bit string like $|01101\rangle$, corresponds to a specific point or concept in the meaning space. A word's meaning is then a superposition over these basis states. This choice of representation is intrinsically linked to the core advantage we aim to exploit: the power of quantum sampling [18].

Fundamentally, this entire quantum process is designed to leverage that power. After the variational circuit $U(\theta)$ entangles the word registers, the final state $|\Psi_{\text{final}}\rangle$ represents a complex joint probability distribution over all possible meaning combinations encoded in these bit strings. The subsequent measurement of the qubits is not merely a readout step; it is a computationally powerful act of sampling. Each shot of the experiment draws a single, fair sample from a distribution that may be too complex and correlated for any classical algorithm to sample from efficiently [25, 26]. The statistics we build from these samples to calculate our cost function are therefore derived from a richer, more expressive probabilistic model than classical methods can typically access. The algorithm's task is thus to learn how to manipulate the amplitudes of the superposition so that sampling favors the correct contextual meaning.

2.1 Variational Circuit Training

Variational quantum algorithms (VQAs) are a hybrid quantum-classical methodology, ideal for the capabilities of near-term quantum processors. The training process is an optimization loop where a classical computer tunes the parameters of a quantum circuit to minimize a cost function [23, 27].

The key component of our VQA is the ansatz design, which is a PQC denoted $U(\theta)$—a sequence of quantum gates with tunable parameters θ [28]. The design of this circuit is crucial. It must be expressive enough to generate the complex, entangled states needed to represent contextual meaning, yet shallow enough to be run effectively on noisy quantum hardware. Our ansatz will be structured to take the initial product state of word embeddings, $|\Psi_{\text{initial}}\rangle = |\psi_{\omega_1}\rangle \otimes |\psi_{\omega_2}\rangle \otimes \dots$, and evolve it into a final, entangled state $|\Psi_{\text{final}}(\theta)\rangle$ that captures the sentence's holistic meaning [28].

The loss function for our training is based on state fidelity [23]. Let's say for a given sentence, we know the ideal target state that represents the correct contextualized meanings, denoted $|\Psi_{\text{target}}\rangle$. The fidelity between our model's output state and the target state is $F = |\langle\Psi_{\text{final}}(\theta)|\Psi_{\text{target}}\rangle|^2$ [29]. Our loss function can then be defined as $L(\theta) = 1 - F$. The goal of the training is to find the parameters θ that maximize the fidelity, thus minimizing the loss. In practice, creating a full $|\Psi_{\text{target}}\rangle$ for every sentence is difficult. A more practical approach is to define the loss based on measurements of only the target word's register. We compare the measured posterior probability distribution with the known correct sense, using a classical loss function like cross-entropy, which is computed from the quantum measurement outcomes [25, 26].

The optimization is handled by a classical algorithm. After running the PQC for a given set of parameters θ and calculating the loss $L(\theta)$, this value is passed to an optimizer. The optimizer then suggests a new set of parameters θ'. A common challenge is estimating the gradient of the loss function with respect to the circuit parameters. The parameter-shift rule is a powerful technique for

this [30]. It shows that the gradient can be calculated by running the same quantum circuit twice with slightly shifted parameters, allowing us to use gradient-based optimizers like Adam or SGD [31]. This iterative loop of executing the PQC, measuring the outcome, calculating the loss, and classically updating the parameters continues until the loss converges to a minimum. At that point, the circuit $U(\theta)$ has learned the contextual transformation rules of the language.

Commonly used ansatz The choice of ansatz architecture is a balance between expressiveness and trainability. A widely used architecture is the Hardware-Efficient Ansatz. This design consists of alternating layers of single-qubit rotation gates and two-qubit entangling gates [28]. The specific gates and their connectivity are chosen to match the native capabilities of the quantum hardware, minimizing errors from circuit compilation. For instance, a layer might apply parameterized R_y and R_z rotations to every qubit, followed by a layer of CNOT or CZ gates that entangle adjacent qubits. By stacking multiple such layers, the circuit can generate a rich variety of entangled states.

Another class of ansatz is inspired by Tensor Networks, such as the Matrix Product State (MPS) ansatz [32, 33]. These circuits are designed to generate states with a controlled, typically low, amount of entanglement. For language, which has a somewhat linear structure, an MPS-like ansatz that primarily entangles adjacent words can be very effective. This can also make the optimization process more stable. For our QPWE model, which needs to capture potentially long-range dependencies in sentences, a more expressive Hardware-Efficient Ansatz is a strong starting point. The parameters of the entangling gates in this ansatz will learn the strength and nature of the semantic influence between different words in a sentence [28].

2.2 Contextual Updating in Quantum Algorithm

A central challenge in modeling language is to implement a mechanism that updates the meaning of a word based on its context. This process is analogous to Bayesian inference, where a prior belief is updated to a posterior belief in light of new evidence. The foundational ancilla-driven Bayesian update mechanism provides a powerful means of implementing soft, probabilistic control, where the influence of context is mediated through the controlled rotation of an auxiliary qubit. This section expands upon that concept by introducing a more versatile toolkit of contextual control mechanisms. The objective is to equip the QPWE model with a range of functions capable of implementing distinct layers of semantic influence, from deterministic logical branching to continuously learned, nuanced modulation.

The exploration of these methods is framed by the inherent trade-off between the expressibility of a quantum circuit and its trainability in the Noisy Intermediate-Scale Quantum (NISQ) era. Highly expressive ansatze, capable of exploring vast regions of the Hilbert space, often suffer from barren plateaus, where gradients vanish exponentially, rendering optimization intractable. Conversely, constraining an ansatz for trainability may limit its capacity to capture complex correlations. Each method proposed herein is evaluated through this lens, balancing expressive power against resource cost and noise resilience. In the following subsections, we are going to explore methods which offer different layers and intents of control through contextual update [24].

2.2.1 Learned Continuous Control via Parameterized Entangling Gates

This method represents a paradigm shift away from explicit, programmed conditional logic. Instead, it integrates contextual control directly and organically into the variational ansatz itself. The strength and nature of the interaction between context and target qubits are not predetermined but are represented by tunable parameters within the entangling gates of the circuit. The classical optimizer then

learns the appropriate contextual influence by minimizing the overall cost function during the training loop.

The operational flow for this method embeds the control logic within the variational ansatz itself:

1. **Ansatz Design:** A variational circuit, such as a Hardware-Efficient Ansatz, is constructed to connect the context and target qubit registers with one or more layers of two-qubit entangling gates.

2. **Parameterized Entanglers:** Crucially, the entangling gates themselves are parameterized. Instead of using a fixed gate like CNOT, one employs a gate whose entangling strength is tunable, such as controlled-Y rotations ($CR_y(\theta)$) or parameterized Mølmer-Sørensen gates ($XX(\theta)$). The rotation angle θ is a variational parameter to be optimized.

3. **VQA Optimization Loop:** The circuit is executed within a standard VQA loop. The classical optimizer adjusts all variational parameters, including the angles θ of the entangling gates. If a strong interaction between a specific context qubit and a target qubit is beneficial for minimizing the cost function, the optimizer will learn a value for the corresponding θ that creates strong entanglement. Conversely, if the interaction is irrelevant, it will learn a value of $\theta \approx 0$, effectively turning off that gate.

With this approach, control is not a prescribed logical rule but an emergent property of the trained model. The final set of optimized parameters θ^* is the learned contextual control model. This approach is also inherently adaptable to hardware noise, as the optimizer may naturally learn to avoid using a particularly noisy entangling gate if a less noisy pathway can achieve a better result. This method is perfectly suited for modeling the subtle, non-binary nature of most linguistic context, capturing the analog push of a contextual cue far more naturally than a discrete if-then rule.

2.2.2 Direct Conditional Logic via Mid-Circuit Measurement and Feed-Forward

The most direct method for implementing context-dependent operations is to leverage the advanced capabilities of dynamic quantum circuits. This approach realizes classical if-else branching logic within the coherence time of the quantum computation, enabling the circuit's evolution to adapt based on intermediate information [34].

This approach unfolds in a sequence of steps that combine quantum evolution with real-time classical processing:

1. **State Preparation and Evolution:** The QPWE circuit begins by preparing the initial product state of word embeddings and applying the initial entangling layers of the variational ansatz.

2. **Context Measurement:** At a designated point, the qubits corresponding to the context word are measured. These outcomes collapse the superposition for those qubits into classical bitstrings, which are stored in classical registers.

3. **Classical Feed-Forward:** The classical control electronics process these outcomes in real time. This computation can range from a simple evaluation to a more complex classical function.

4. **Conditional Operation:** Based on the result of the classical evaluation, a subsequent quantum gate or block of gates, the update unitary U_{update} is either applied to the target word's register or skipped entirely.

While this method's capacity to implement intuitive classical logic is a significant advantage, it is inextricably linked to a primary drawback on NISQ hardware: latency-induced decoherence [35]. The

physical process of measurement, classical processing, and subsequent gate triggering is not instantaneous, introducing a delay during which unmeasured qubits remain idle and accumulate errors. This makes the method a double-edged sword, suitable only for scenarios where the computational benefit of the logical branch outweighs the guaranteed fidelity loss from the added delay. Its strategic value lies in modeling hard contextual disambiguation, where a context word deterministically collapses the meaning of a target word.

2.2.3 High-Level Deterministic Control with Multi-Qubit Controlled Unitaries

An alternative to dynamic circuits for implementing deterministic control is to use multi-controlled unitary gates [34], denoted as $C^{\otimes n}U$. This method treats the entire context register as a single, cohesive control for a target operation. A $C^{\otimes n}U$ gate applies a specific update unitary U_{update} to the target word's register if and only if all n qubits in the context register are in a specific state, typically $|11...1\rangle$.

The implementation of a $C^{\otimes n}U$ gate is a coherent process that involves three key stages:

1. **Context Encoding:** The context words are processed such that their combined meaning, which should trigger the update, corresponds to a specific multi-qubit computational basis state in the context register.

2. **Controlled Unitary Application:** A $C^{\otimes n}U_{\text{update}}$ gate is applied, with the n context qubits as controls and the target word's qubits as targets.

3. **Decomposition:** A significant challenge is that $C^{\otimes n}U$ gates are not native operations on most current quantum hardware and must be decomposed into a sequence of fundamental gates, such as single-qubit rotations and two-qubit CNOT gates. The efficiency of this decomposition is the central factor determining feasibility.

The practical implementation of $C^{\otimes n}U$ gates on NISQ devices reveals a critical resource dilemma. Without ancillary qubits, decomposing a $C^{\otimes n}U$ gate on hardware with restricted connectivity requires a gate count that scales quadratically with the number of control qubits ($\mathcal{O}(n^2)$). This prohibitive scaling in circuit depth makes such decompositions infeasible for all but the smallest n. However, introducing just one or two ancillary qubits can dramatically reduce the gate count to a linear scaling ($\mathcal{O}(n)$) [36]. This presents a fundamental architectural choice: forgo ancillas and suffer overwhelming noise from a deep circuit, or use scarce ancillas to achieve a feasible depth at the cost of sacrificing valuable encoding qubits. This method is the quantum analogue of a logical AND gate, making it ideally suited for modeling semantic gatekeeping, where a specific combination of contextual factors is required to unlock a particular meaning.

2.2.4 Iterative Ancilla-Mediated Updates with Qubit Reuse

This method is a direct and resource-efficient extension of the original ancilla-driven mechanism described in the previous version of this document. It makes the approach more scalable by incorporating mid-circuit measurement and qubit reset, a technique often referred to as qubit reuse. Instead of requiring a separate, dedicated ancilla qubit for each contextual interaction, a single ancilla qubit is used iteratively to process a sequence of context words [37].

This iterative approach unfolds as a repeating cycle of encoding, updating, and resetting the ancilla qubit:

1. **Encode and Update (Word 1):** The process begins as originally proposed. The state of the first context word's register controls a rotation on the ancilla qubit, which in turn probabilistically controls the application of an update unitary to the target word's register.

2. **Measure, Reset, and Reuse:** After the first interaction, the ancilla qubit is measured. Based on the classical outcome, a conditional X gate can be applied to deterministically reset its state to $|0\rangle$. The ability to perform high-fidelity, low-crosstalk measurement and reset is a critical hardware requirement.

3. **Encode and Update (Word 2):** The now-reset ancilla is immediately available for reuse. The state of the second context word's register controls a new rotation on the same ancilla, which in turn controls a second update operation on the target word.

4. **Iteration:** This process of encode → update → measure → reset repeats for all relevant context words.

This method fundamentally alters the structure of the contextual update from a parallel to a serial process. By applying the concept of qubit reuse, one ancilla can be used N times for N context words, avoiding a linear increase in qubit overhead. The cost of this increased qubit efficiency is an increase in circuit depth, as the total execution time now scales linearly with the number of context words. This creates a clear trade-off: Reduced Qubit Count (Width) ⇔ Increased Circuit Depth (Time) [34, 37]. On NISQ devices where both qubit count and coherence time are severely limited, this serial approach might be the only viable implementation strategy for problems with long and complex contexts. Furthermore, this iterative mechanism naturally mirrors the sequential nature of human language processing, where meaning is built up word by word.

In conclusion, the four mechanisms presented—learned continuous control, direct conditional logic, multi-controlled gates, and iterative ancilla updates—collectively form a comprehensive toolkit for contextual control in the QPWE framework. They approach the problem from distinct perspectives, offering a spectrum of control granularities. The learned, continuous nature of parameterized entanglers contrasts sharply with the rigid, deterministic logic of multi-controlled gates and mid-circuit measurements. Similarly, the resource-efficient, sequential processing of the iterative ancilla method provides an alternative to parallel, all-at-once entanglement schemes. This versatility is crucial for modeling the complexity of natural language. By equipping a quantum language model with this diverse set of tools, it becomes possible to apply different layers and intensities of contextual influence, from a subtle probabilistic shift for nuanced word pairings to a definitive, deterministic update for unambiguous phrases. This provides a richer, more flexible framework for capturing the intricate web of semantic relationships that define meaning in large language models.

2.3 The Role of Quantum Sampling in the QPWE Process

During the variational training loop, the primary role of the quantum computer is to prepare the state $|\Psi_{\text{final}}(\theta)\rangle$ for a given set of parameters. However, we cannot directly access the full state or its amplitudes. The classical optimizer needs a single scalar value—the cost—to decide how to update the parameters. This value is obtained through sampling.

We execute the same quantum circuit thousands of shots. Each execution concludes with a measurement of the relevant qubits, which collapses the superposition into a single classical bit string. This single outcome is one sample. By collecting thousands of these samples, we build a histogram of the outcomes [25, 26]. This histogram serves as a high-fidelity statistical estimate of the true probability distribution defined by the quantum state's amplitudes. It is from this empirically sampled distribution that we calculate the cost function (e.g., cross-entropy loss against the target label). In essence, sampling is the process by which we extract the necessary classical information from the quantum state to guide the learning process.

3 Warm Start and Cost Function for Training PWE

The development of any advanced quantum machine learning model does not occur in a vacuum. It benefits immensely from the mature landscape of classical machine learning. Our QPWE framework is designed as a hybrid system that strategically combines the strengths of both computational paradigms. The initial and most data-intensive part of the learning process, which involves extracting fundamental semantic structures from vast text corpora, is best handled by classical methods. This section details the classical component of our framework. We first explain how word embedding and more sophisticated PWE are traditionally trained, then discuss how existing pre-trained classical PWE models can provide an informed starting point—a warm start—for our quantum algorithm. This hybrid approach ensures that our quantum processor's limited resources are focused on the task it is uniquely suited for: modeling the complex, quantum-like correlations of context.

3.1 Classical Pre-trained PWE Models for Warm Start

The process of training a word embedding model from scratch requires immense computational resources and vast datasets. A more practical and efficient approach is to utilize pre-trained models. For our QPWE framework, we adopt a warm start strategy. We leverage classical PWE models that have already been trained on large text corpora to provide the initial parameters for our quantum system. This allows the quantum algorithm to begin its learning process from a state of considerable knowledge, rather than from a random or uniform configuration.

Several types of classical PWE models are suitable for this purpose. A prominent example is Gaussian embedding, where each word is represented by a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ [3]. The mean vector μ captures the word's primary meaning, while the covariance matrix Σ models its ambiguity or semantic variance. A highly polysemous word would be represented by a distribution with high variance, spread across a larger semantic region. Another powerful approach is using Gaussian Mixture Models (GMMs). Here, a word's meaning distribution is modeled as a weighted sum of multiple Gaussian components [3]. Each component can represent a distinct sense of the word. For example, the word *cell* could be modeled as a mixture of three Gaussians, corresponding to its biological, electrical, and prison-related meanings. The weights of the mixture provide the prior probability of each sense.

The output from these pre-trained models is precisely what we need to initialize our quantum algorithm. A trained GMM, for instance, provides a discrete probability vector $(p_1, p_2, \cdots, p_M)$ for each word, where p_i is the prior probability of its i-th meaning. This vector serves as the classical foundation for our quantum state preparation.

The initialization process can be conceptualized in two ways. For a standard, non-probabilistic word embedding (a single vector), the classical representation is equivalent to a delta function in the probability space [38]. All the probability mass is concentrated at a single point, representing a state of maximum certainty about the word's meaning. Our quantum initialization can start from this point of certainty. More powerfully, we can start from a pre-trained PWE such as a GMM. The prior probabilities from the GMM are then used to construct the initial quantum state.

The method for encoding this classical information into a quantum state is a critical step. In Section 2.1, we detail how these classical probabilities are mapped onto the amplitudes of a quantum state using a technique called amplitude encoding. For a word with M possible senses and probabilities $\{p_i\}$, we prepare a quantum register in the state $|\psi\rangle = \sum_{i=1}^{M} \sqrt{p_i}|i\rangle$. The probability of measuring the state corresponding to the i-th meaning is $|\sqrt{p_i}|^2 = p_i$, perfectly replicating the classical distribution. This warm start procedure ensures that our quantum model inherits the rich semantic knowledge learned by classical models, allowing the subsequent quantum training to focus on the more nuanced task of learning contextual correlations.

3.2 Classical Cost Functions for Training PWE

Training a PWE model relies on defining a cost function that captures the semantic and syntactic patterns in a large corpus [39]. Several categories of cost functions have been developed, including similarity-based objectives [40], the Skip-gram with Negative Sampling approach [41], and information-theoretic measures like cross-entropy and KL divergence [42, 43]. These objectives drive the model to assign higher probability to observed word–context pairs and lower probability to unrelated pairs. In practice, calculating these objectives over a vocabulary of hundreds of thousands of words is computationally expensive [38]. Techniques such as negative sampling or hierarchical softmax provide approximations to make training feasible [44], but at the cost of introducing some error. (Additional details on classical cost function formulations are provided in Supplementary Section S1.)

3.3 Classical Cost Functions for Contextual Update

Once a PWE model has learned general prior probability distributions for words, the next challenge is to update these distributions based on specific contexts. A purely classical system must learn a function that transforms a prior distribution into a posterior distribution, guided by the surrounding words in a sentence [38]. This is a supervised learning task governed by cost functions designed to penalize incorrect contextual interpretations.

In a classical PWE, contextual updates are typically trained on labeled word sense disambiguation (WSD) datasets [45]. For each occurrence of an ambiguous target word in a sentence, the model (often a neural network) uses the target's prior distribution and the embeddings of its context words as input, and outputs a posterior probability distribution over the target's possible senses. Training uses cross-entropy loss [42], which encourages the model to assign maximal probability to the correct sense, as well as KL-divergence-based objectives [43] if a soft target distribution is available. In practice, capturing the complex interactions among all context words often requires very large models [46], making this training computationally intensive.

This classical difficulty motivates our quantum approach. As we explore in Section 2.2, a quantum circuit can be designed to perform an analogous Bayesian update intrinsically. The quantum algorithm essentially embeds the contextual update within its dynamics: after the circuit evolves and we measure the output state to obtain a posterior distribution, a classical cost (e.g., cross-entropy with respect to the true sense) is computed and used to refine the quantum circuit's parameters. This offers a potentially more efficient path to compute the high-dimensional correlations required for contextual disambiguation. (Further details on classical contextual updating are provided in Supplementary Section S1.)

4 NISQ Implementation

4.1 Scalable QPWE Architecture for NISQ

QPWE introduces a novel method for representing linguistic ambiguity via quantum superposition and context via entanglement. To move this framework from theory to a verifiable model, its components must be scaled to handle the complexities of real-world language. This section outlines a scalable QPWE architecture designed for the task of all-words WSD, establishing a foundation for a credible experimental protocol.

Real-world WSD requires navigating large vocabularies and highly polysemous words, which necessitates a sophisticated model architecture [47]. We propose a scalable QPWE design that enhances the quantum register to manage polysemy and employs a more expressive variational circuit to learn complex correlations.

Quantum Register Architecture for Polysemy The cornerstone of the scaled model is a redesigned quantum register architecture that moves beyond single-qubit representations. The system is partitioned into dedicated registers for the target word and its surrounding context words.

- **Target Register:** Instead of a single qubit, we allocate an n_t-qubit register to represent the target word, which can encode up to 2^{n_t} possible senses (for a word with M senses, $n_t \approx \lceil \log_2 M \rceil$).
- **Context Registers:** Rather than one combined context qubit, we use multiple registers for the $2k$ context words (a window of k words to the left and k to the right). Each context word with M_c possible senses is assigned $n_c \approx \lceil \log_2 M_c \rceil$ qubits, allowing the model to encode each context word's semantic contribution independently.

For a given disambiguation instance, the joint quantum state is the tensor product of the target register state and all context register states. This high-dimensional state space can represent complex, interdependent relationships between a target word and its surrounding linguistic environment.

Variational Ansatz for Contextual Correlation Next, we design a variational quantum circuit (ansatz) with L layers to perform the contextual update. We follow a hardware-efficient style ansatz [28, 48] tailored to our multi-register setup: each layer applies parameterized single-qubit rotations to every qubit, then entangling gates connecting qubits. In the initial layers, entangling gates primarily connect each context register with the target register, allowing the model to learn the direct influence of each context word on the target. In later layers, entangling gates also connect qubits across different context registers, enabling the circuit to capture higher-order interactions between context words. This structured layering provides the necessary expressivity to model non-trivial correlations in language while keeping circuit depth manageable for NISQ hardware (thereby mitigating issues like barren plateaus in deep circuits [23]).

4.2 Realistic Benchmark

For empirical evaluation, we choose the SemCor corpus as the primary dataset for training and testing the scaled QPWE model. SemCor is a subset of the Brown Corpus in which every content word is manually annotated with its WordNet sense, providing high-quality ground-truth labels for WSD. It contains roughly 234,000 tagged words, making it one of the largest sense-annotated corpora available [47].

Given the computational limits of current NISQ hardware, we will work with a smaller, representative subset of SemCor. Specifically, we select on the order of 5–10 highly polysemous target words from the corpus and extract all sentences in SemCor that include those targets. These instances are divided into a standard 80%/20% training/test split for our experiments.

From this subset, we derive the prior sense distribution for each target word directly from its frequency of senses in the training portion of SemCor. In other words, we initialize the target qubit register in a state that reflects the empirical probability of each sense, an approach we term an *Empirical Warm Start*. This approach forgoes an external classical model in favor of using the corpus's own statistics to set the initial amplitudes (via amplitude encoding) of the quantum state [49, 50]. As a result, the quantum algorithm begins with a prior state that transparently embodies the knowledge of sense frequencies in the dataset.

4.3 Experimental Protocol

With a scalable model and data pipeline established, we detail a hybrid quantum-classical training procedure to train the QPWE model on the curated SemCor subset. The training follows a standard

Variational Quantum Algorithm (VQA) loop.

In each iteration of training, the quantum processor prepares the initial state (using the Empirical Warm Start distributions for the target and context registers), then applies the parameterized ansatz circuit $U(\theta)$ to produce a final quantum state. Measuring the target register yields a sample from the model's posterior distribution over senses. A classical computer then computes a cost function (e.g., cross-entropy between the measured distribution and the correct sense label) and uses it to update the circuit parameters via a classical optimizer. This loop repeats for each training example and across multiple epochs until convergence. (See Supplementary Section S2 for a detailed step-by-step description of the training loop.)

5 Metrics Definition

Evaluating a novel framework like QPWE requires a multifaceted approach. We need metrics that can benchmark its performance against classical systems on established linguistic tasks, ensuring its practical relevance. At the same time, we need metrics that can probe the unique characteristics of the quantum states it produces, addressing the deeper motivations of our research regarding integrated information. This section defines both types of metrics. First, we specify performance benchmarks focused on key linguistic phenomena like hypernymy and polysemy. Second, we detail our proposed consciousness metric, based on a rigorous formulation of Quantum Integrated Information Theory, designed to quantify the holistic nature of our model's semantic representations.

5.1 Performance Metrics

To assess the linguistic capabilities of the QPWE model, we will evaluate it on a suite of tasks that probe different aspects of semantic understanding.

- **Polysemy and WSD:** This is the most direct test of our model's core functionality. We will use standard WSD benchmarks like SemCor and the SemEval task series [45]. The primary metric will be accuracy: the percentage of ambiguous words correctly assigned their contextually appropriate sense. Success on these benchmarks will demonstrate that the quantum contextual updating mechanism is effectively learning to resolve ambiguity.

- **Hypernymy Detection:** Hypernymy is the "is-a" relationship (e.g., a "dog" is an "animal"). This hierarchical relationship is a cornerstone of semantic knowledge [51]. We can test this by framing it as a classification task. Given a pair of words (e.g., "dog," "animal"), the model must predict if one is a hypernym of the other. We can use datasets like WordNet to generate pairs for this task. The model's ability to capture these relationships can be evaluated by analyzing the geometry of its embedding space. Intuitively, the probabilistic volume of a hyponym (e.g., "poodle") should be contained within the volume of its hypernym (e.g., "dog"). We can devise a metric that measures this probabilistic inclusion, rewarding models that learn this hierarchical structure.

- **Contextual Embedding Evaluation:** To compare our model with modern contextual embeddings like BERT, we can use it on downstream tasks from the GLUE (General Language Understanding Evaluation) benchmark [5]. We would use the QPWE model to generate contextualized embeddings for sentences, which are then fed into a simple classifier for tasks like sentiment analysis or natural language inference. While we do not expect a small-scale QPWE model to outperform a massive model like BERT, this provides a standardized way to measure its ability to produce useful, context-aware representations. The performance on these tasks will

be measured using the specific metrics for each task, such as the Matthews correlation coefficient for the CoLA task or accuracy for MNLI.

These performance metrics will provide a comprehensive and quantitative picture of our model's strengths and weaknesses, allowing for a rigorous comparison with the classical state of the art.

5.2 Consciousness Metric: Quantum IIT

To investigate the hypothesis that our quantum framework fosters a more integrated representation of meaning, we will employ a metric from the theory of consciousness: Q-IIT [22]. Q-IIT provides a mathematical procedure for calculating a quantity Φ which measures the degree to which a system's information is irreducible to its parts. A high Φ value signifies that the system's state carries a large amount of integrated information, meaning it cannot be decomposed into independent components without significant loss of information content.

In the context of QPWE, we will interpret "system" to mean the collection of qubit registers representing a sentence. To compute Φ, we will consider bipartitions of this set of qubits into two disjoint subsets (which Q-IIT terms a "mechanism" and a "purview") and evaluate how much information is lost when the two subsets are treated as independent. Mathematically, this involves calculating a divergence between the joint state and the tensor product of the marginal states of each subset. One candidate measure is the Quantum Intrinsic Difference (QID) divergence [22], a quantum generalization of the Kullback–Leibler divergence that can quantify the discrepancy between the actual entangled state and the closest product state approximation. Formally, if the full system is in state ρ_{AB} on qubits $A \cup B$, and $\rho_A \otimes \rho_B$ is the product of its marginals, then one can define:

$$\Phi = \min_{(A,B)} D_{\mathrm{QID}}\Big(\rho_{AB} \,||\, \rho_A \otimes \rho_B\Big),$$

where $I(A:B)$ is the quantum mutual information across the partition (A, B), and the minimum is taken over all bipartitions.

Our primary hypothesis is that the process of training the QPWE model to correctly understand language will naturally increase the Φ of the quantum states representing sentences. We will calculate the Φ value for the final state $|\Psi_{\mathrm{final}}\rangle$ produced by our model for sentences in our test set. We predict that sentences that are correctly disambiguated and understood by the model will correspond to states with significantly higher Φ values than those produced for sentences where the model fails. A strong correlation between performance and Φ would be a landmark result, suggesting that creating highly integrated informational structures is a functional requirement for sophisticated semantic processing, providing a tangible, measurable link between the success of an AI system and a key property associated with consciousness.

6 Conclusion

This proposal has introduced a comprehensive framework for QPWE, a novel approach that integrates the principles of quantum mechanics into the core of natural language representation. We have articulated a hybrid quantum-classical methodology that addresses the fundamental limitations of classical models in handling linguistic ambiguity. Our primary contribution is the design of a system where word meanings are represented as quantum superpositions, and contextual updates are performed through the orchestrated entanglement of these states within a variational quantum circuit. We have detailed the entire pipeline, from the classical warm start using pre-trained PWE to the innovative ancilla-driven quantum Bayesian update mechanism, and grounded it in a feasible implementation plan for near-term trapped-ion quantum computers.

Our work seeks to forge a path towards conscious AI, not by claiming to create a conscious machine, but by exploring the types of computational architectures that could support the properties associated with consciousness. Our approach is testable in two complementary ways. First, by measuring preference metrics, we will run paired comparisons against strong classical baselines on tasks that stress contextual disambiguation, compositional generalization, and robustness under distribution shift. Human or expert-model judgments will be aggregated into a scalar Preference Score P that captures systematic preference for QPWE outputs. Second, by measuring the system-level Φ score, we will quantify the degree of information integration in our model's representations, testing the hypothesis that holistic semantic understanding requires the formation of irreducible, integrated informational wholes. This dual focus on performance and information integration represents a significant shift in the evaluation of AI systems.

For ethical and societal implications, our immediate goal is a small-scale, well-contained prototype that introduces a quantum component to the word embedding subroutine, yet the broader research arc engages issues of self-awareness and agency. This platform provides a tractable way to probe the quantum mind hypothesis. If evidence accumulates in support of the hypothesis, similar mechanisms may enable human–machine symbiosis via quantum channels that support co-processing and shared context with strong privacy and alignment guarantees. Such channels could provide new mechanisms for embedded oversight and risk reduction. In parallel, we advocate regulatory scaffolding at this exploratory stage: preregistration of experiments, mandatory reporting of Φ with confidence intervals alongside behavioral metrics, containment requirements including limited memory persistence, reversible disentangling procedures and hard shutdown controls, and independent audits focused on emergent goal-seeking or deceptive behaviors. We further propose a minimal standard for systems with consciousness-relevant properties.

The immediate next step in this research is to implement and train the proposed small-scale model on a NISQ device. This will provide crucial empirical validation of our core mechanisms. Following this, we will scale the model to larger vocabularies and more complex sentence structures, continuously benchmarking its performance against classical models. Future work will also involve refining the ansatz design and exploring more sophisticated quantum machine learning techniques for optimization. Ultimately, this research program aims to not only advance the capabilities of natural language processing but also to provide a new computational paradigm and a new set of tools for investigating the profound connection between information, physics, and the nature of conscious awareness.

Acknowledgment

The authors thank FinQ Tech Inc. for providing the platform that made this collaboration and its discussions possible.

References

[1] Thomas K. Landauer and Susan T. Dumais. "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge". In: *Psychological Review* 104.2 (1997), pp. 211–240. DOI: 10.1037/0033-295X.104.2.211. URL: https://doi.org/10.1037/0033-295X.104.2.211.

[2] Hinrich Schütze. "Automatic word sense discrimination". In: *Computational Linguistics* 24.1 (1998), pp. 97–123. URL: https://aclanthology.org/J98-1004.

[3] Luke Vilnis and Andrew McCallum. "Word representations via gaussian embedding". In: *arXiv preprint arXiv:1412.6623* (2014).

[4] Tomas Mikolov et al. "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems (NIPS 2013)*. Vol. 26. 2013, pp. 3111–3119.

[5] Jacob Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://doi.org/10.18653/v1/N19-1423.

[6] Emily M. Bender et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. 2021. DOI: 10.1145/3442188.3445922.

[7] John S. Bell. "On the Einstein Podolsky Rosen paradox". In: *Physics* 1 (1964), pp. 195–200. DOI: 10.1103/PhysicsPhysiqueFizika.1.195.

[8] Simon Kochen and Ernst P. Specker. "The Problem of Hidden Variables in Quantum Mechanics". In: *Journal of Mathematics and Mechanics* 17 (1967), pp. 59–87. DOI: 10.1512/iumj.1968.17.17004. URL: https://doi.org/10.1512/iumj.1968.17.17004.

[9] John H. Conway and Simon Kochen. "The Free Will Theorem". In: *Foundations of Physics* 36.10 (2006), pp. 1441–1473. DOI: 10.1007/s10701-006-9068-6.

[10] Alastair A. Abbott, Cristian S. Calude, and Karl Svozil. "Strong Kochen–Specker Theorem and Incomputability of Quantum Randomness". In: *Physical Review A* 86 (2012), p. 062109. DOI: 10.1103/PhysRevA.86.062109.

[11] Stefano Pironio and et al. "Random Numbers Certified by Bell's Theorem". In: *Nature* 464.7291 (2010), pp. 1021–1024. DOI: 10.1038/nature09008.

[12] Emmanuel M. Pothos and Jerome R. Busemeyer. "Can Quantum Probability Provide a New Direction for Cognitive Modeling?" In: *Behavioral and Brain Sciences* 36.3 (2013), pp. 255–274. DOI: 10.1017/S0140525X12001525.

[13] Peter D. Bruza, Zheng Wang, and Jerome R. Busemeyer. "Quantum Cognition: A New Theoretical Approach to Psychology". In: *Trends in Cognitive Sciences* 19.7 (2015), pp. 383–393. DOI: 10.1016/j.tics.2015.05.001.

[14] Peter D. Bruza et al. "Is There Something Quantum-Like About the Human Mental Lexicon?" In: *Journal of Mathematical Psychology* 53.5 (2009), pp. 362–377. DOI: 10.1016/j.jmp.2009.04.004.

[15] Jerome R. Busemeyer and Peter D. Bruza. *Quantum Models of Cognition and Decision*. Cambridge: Cambridge University Press, 2012. ISBN: 9780521862749. DOI: 10.1017/CBO9780511997716.

[16] Diederik Aerts and Sandro Sozzo. "Quantum Structure in Cognition: Why and How Concepts Are Entangled". In: *Quantum Interaction (QI 2011), Lecture Notes in Computer Science, vol. 7052*. Springer, 2011, pp. 116–127. DOI: 10.1007/978-3-642-24971-6_12.

[17] Scott Aaronson and Alex Arkhipov. "The Computational Complexity of Linear Optics". In: *Proceedings of the 43rd ACM Symposium on Theory of Computing (STOC 2011)*. 2011, pp. 333–342. DOI: 10.1145/1993636.1993682.

[18] Frank Arute et al. "Quantum supremacy using a programmable superconducting processor". In: *Nature* 574.7779 (2019), pp. 505–510. DOI: 10.1038/s41586-019-1666-5.

[19] David Balduzzi and Giulio Tononi. "Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework". In: *PLoS Computational Biology* 4.6 (2008), e1000091. DOI: 10.1371/journal.pcbi.1000091.

[20] Pedro A. M. Mediano, Anil K. Seth, and Adam B. Barrett. "Measuring Integrated Information: Comparison of Candidate Measures in Theory and Simulation". In: *Entropy* 21.1 (2019), p. 17. DOI: 10.3390/e21010017.

[21] Masafumi Oizumi, Larissa Albantakis, and Giulio Tononi. "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0". In: *PLoS Computational Biology* 10.5 (2014), e1003588. DOI: 10.1371/journal.pcbi.1003588.

[22] Paolo Zanardi, Michael Tomka, and Lorenzo Campos Venuti. *Towards Quantum Integrated Information Theory*. 2018. arXiv: 1806.01421 [quant-ph]. URL: https://arxiv.org/abs/1806.01421.

[23] Marco Cerezo et al. "Variational quantum algorithms". In: *Nature Reviews Physics* 3.9 (2021), pp. 625–644. DOI: 10.1038/s42254-021-00348-9.

[24] Jarrod R McClean et al. "Barren plateaus in quantum neural network training landscapes". In: *Nature communications* 9.1 (2018), p. 4812.

[25] Jin-Guo Liu and Lei Wang. "Differentiable learning of quantum circuit born machines". In: *Physical Review A* 98.6 (2018), p. 062324.

[26] Marcello Benedetti et al. "A generative modeling approach for benchmarking and training shallow quantum circuits". In: *npj Quantum information* 5.1 (2019), p. 45.

[27] Yuto Takaki et al. "Learning temporal data with a variational quantum recurrent neural network". In: *Physical Review A* 103.5 (2021), p. 052414.

[28] Abhinav Kandala et al. "Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets". In: *nature* 549.7671 (2017), pp. 242–246.

[29] Juan Carlos Garcia-Escartin and Pedro Chamorro-Posada. "Swap test and Hong-Ou-Mandel effect are equivalent". In: *Physical Review A—Atomic, Molecular, and Optical Physics* 87.5 (2013), p. 052330.

[30] Maria Schuld et al. "Evaluating analytic gradients on quantum hardware". In: *Physical Review A* 99.3 (2019), p. 032331.

[31] David Wierichs et al. "General parameter-shift rules for quantum gradients". In: *Quantum* 6 (2022), p. 677.

[32] Reza Haghshenas et al. "Variational power of quantum circuit tensor networks". In: *Physical Review X* 12.1 (2022), p. 011047.

[33] Daniel Malz et al. "Preparation of matrix product states with log-depth quantum circuits". In: *Physical Review Letters* 132.4 (2024), p. 040404. DOI: 10.1103/PhysRevLett.132.040404. URL: https://doi.org/10.1103/PhysRevLett.132.040404.

[34] Elisa Bäumer et al. "Efficient long-range entanglement using dynamic circuits". In: *PRX Quantum* 5.3 (2024), p. 030339.

[35] Petr Ivashkov et al. "High-fidelity, multiqubit generalized measurements with dynamic circuits". In: *PRX Quantum* 5.3 (2024), p. 030315.

[36] Adriano Barenco et al. "Elementary gates for quantum computation". In: *Physical review A* 52.5 (1995), pp. 3457–3467. DOI: 10.1103/PhysRevA.52.3457. URL: https://doi.org/10.1103/PhysRevA.52.3457.

[37] Matthew DeCross et al. "Qubit-reuse compilation with mid-circuit measurement and reset". In: *Physical Review X* 13.4 (2023), p. 041057. DOI: 10.1103/PhysRevX.13.041057. URL: https://doi.org/10.1103/PhysRevX.13.041057.

[38] Alexandr Pak et al. “Word embeddings: A comprehensive survey”. In: *Computación y Sistemas* 28.4 (2024), pp. 2005–2029.

[39] Bin Wang et al. “Evaluating word embedding models: Methods and experimental results”. In: *APSIPA transactions on signal and information processing* 8 (2019), e19.

[40] Erhan Sezerer and Selma Tekir. “A survey on neural word embeddings”. In: *arXiv preprint arXiv:2110.01804* (2021).

[41] Kenneth Ward Church. “Word2Vec”. In: *Natural Language Engineering* 23.1 (2017), pp. 155–162.

[42] Anqi Mao, Mehryar Mohri, and Yutao Zhong. “Cross-entropy loss functions: Theoretical analysis and applications”. In: *International conference on Machine learning*. pmlr. 2023, pp. 23803–23828.

[43] Solomon Kullback. “Kullback-leibler divergence”. In: *Tech. Rep.* (1951). DOI: 10.1214/aoms/1177729694. URL: https://doi.org/10.1214/aoms/1177729694.

[44] Bokan Madina Erzhankyzy. “Negative-sampling word-embedding method”. In: *present tense* 10 (2022).

[45] Roberto Navigli. “Word sense disambiguation: A survey”. In: *ACM computing surveys (CSUR)* 41.2 (2009), pp. 1–69.

[46] Wojciech Samek et al. “Explaining deep neural networks and beyond: A review of methods and applications”. In: *Proceedings of the IEEE* 109.3 (2021), pp. 247–278.

[47] Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. “Word sense disambiguation: A unified evaluation framework and empirical comparison”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. The Association for Computational Linguistics. 2017, pp. 99–110.

[48] Alberto Peruzzo et al. “A variational eigenvalue solver on a photonic quantum processor”. In: *Nature communications* 5.1 (2014), p. 4213.

[49] Sahel Ashhab. “Quantum state preparation protocol for encoding classical data into the amplitudes of a quantum information processing register’s wave function”. In: *Physical Review Research* 4.1 (2022), p. 013091.

[50] Mikko Mottonen et al. “Transformation of quantum states using uniformly controlled rotations”. In: *arXiv preprint quant-ph/0407010* (2004).

[51] Kim Anh Nguyen et al. “Hierarchical embeddings for hypernymy detection and directionality”. In: *arXiv preprint arXiv:1707.07273* (2017).

[52] James C Spall. “An overview of the simultaneous perturbation method for efficient optimization”. In: *Johns Hopkins apl technical digest* 19.4 (1998), pp. 482–492.

[53] Julien Gacon et al. “Simultaneous perturbation stochastic approximation of the quantum fisher information”. In: *Quantum* 5 (2021), p. 567.