QuCoWE: Quantum Contrastive Word Embeddings with Variational Circuits for Near-Term Quantum Devices

Rabimba Karanjai[1†,3], **Hemanth Hegadehalli Madhavarao**[3], **Lei Xu**[2], **Weidong Shi**[1]

[1]University Of Houston, [2]Kent State University, [3]PayPal AI Lab
[†]rkaranjai@uh.edu

We present **QuCoWE**, a framework that learns *quantum-native* word embeddings by training parameterized quantum circuits (PQCs) with a contrastive skip-gram objective. Words are mapped to shallow, hardware-efficient circuits with data re-uploading and controlled ring entanglement; similarity is computed via quantum state overlap (fidelity) and passed through a *logit-fidelity* scoring head that recovers the shifted PMI semantics of SGNS/Noise-Contrastive Estimation. We introduce an *entanglement budget* regularizer based on single-qubit purities to keep circuits trainable (mitigating barren plateaus with local costs and bounded entanglement), and we give a noise analysis for depolarizing and readout errors with error-mitigation hooks (zero-noise extrapolation, randomized compiling). On Text8/WikiText-2 pretraining, QuCoWE reaches competitive intrinsic (WordSim-353, SimLex-999) and extrinsic (SST-2, TREC-6) performance versus 50–100d classical baselines while using fewer learned parameters per token via compact PQCs. We ablate qubit counts, re-uploading depth, and entanglement patterns, and report training behavior under noise. Our results suggest that shallow PQCs with calibrated scoring and entanglement control are a viable path to distributional semantics on near-term devices, and they connect classical PMI objectives with quantum fidelity via NCE.

Date: November 14, 2025

1 Introduction

Word embeddings form the foundation of modern natural language processing, encoding semantic relationships through distributional statistics Mikolov et al. (2013b); Pennington et al. (2014). Classical approaches like Word2Vec and GloVe represent words as real-valued vectors optimized through contrastive objectives or matrix factorization. However, these representations face fundamental limitations in capturing complex semantic phenomena such as polysemy, compositionality, and graded entailment relationships.

Quantum computing offers a radically different representational paradigm through complex amplitudes, superposition, and entanglement. Recent advances in quantum natural language processing (QNLP) have explored quantum-inspired models Coecke et al. (2020), but most rely on classical pre-training or simplified quantum simulations. The emergence of Noisy Intermediate-Scale Quantum (NISQ) devices Preskill (2018) enables genuine quantum computation, albeit with significant noise and limited coherence.

This work addresses a fundamental question: *Can we learn quantum-native word embeddings directly from text corpora that leverage unique quantum properties while remaining trainable on near-term devices?* We propose QuCoWE, a framework that represents words as parameterized quantum states trained through contrastive learning. Our approach bridges classical distributional semantics with quantum information theory, establishing theoretical connections between quantum fidelity and pointwise mutual information.

1.1 Motivations and Challenges

The motivation for quantum word embeddings stems from three key observations about the potential advantages of quantum representations. First, the representational capacity of quantum systems is fundamentally different from classical vectors. A quantum state on n qubits requires 2^n complex amplitudes, offering exponential

representational capacity that enables encoding rich semantic structures in compact representations. Second, quantum mechanics provides natural composition operations through tensor products and partial traces that align remarkably well with linguistic compositionality principles Coecke et al. (2020). Third, quantum overlap captures both magnitude and phase relationships through fidelity measurements, potentially distinguishing semantic nuances that remain invisible to classical cosine similarity metrics.

However, realizing these theoretical benefits faces significant practical challenges that must be carefully addressed. The most severe obstacle is the barren plateau phenomenon, where random parameterized circuits suffer from exponentially vanishing gradients McClean et al. (2018), rendering standard training procedures infeasible. Additionally, NISQ devices exhibit high error rates that can overwhelm the delicate quantum states encoding semantic information. Hardware constraints further complicate implementation by restricting available entanglement patterns and limiting circuit depth. Finally, bridging the gap between classical NLP objectives and quantum optimization requires careful architectural and algorithmic design to ensure meaningful learning.

To address these challenges while demonstrating the viability of quantum-native word embeddings, we study whether shallow, hardware-efficient PQCs can learn distributional semantics directly. Our contributions are:

- **Architecture.** A shallow data-reuploading ansatz with ring entanglement, designed for NISQ hardware Pérez-Salinas et al. (2020).
- **Logit-fidelity head.** A calibrated mapping aligning quantum overlap with shifted PMI under SGNS/NCE Levy and Goldberg (2014); Gutmann and Hyvärinen (2012).
- **Entanglement budget.** A purity-based regularizer that favors trainability consistent with local-cost barren-plateau theory Cerezo et al. (2021).
- **Noise analysis.** Closed-form effect of depolarizing noise on overlap, plus error-mitigation hooks (ZNE, randomized compiling) Temme et al. (2017a); Meyer and Wallach (2002).

Scope and Claims: This work pursues NISQ-realistic resource efficiency rather than claiming unconditional quantum speedup. Concretely, QuCoWE matches the performance of classical 50–100 dimensional baselines on both intrinsic word similarity benchmarks and extrinsic text classification tasks, while using fewer learned parameters per token and demonstrating favorable sample efficiency. We restrict our current investigation to word-level embeddings, discussing extensions to sentence-level composition as important future work.

2 Related Work

2.1 Classical Word Embeddings

Distributional semantics hypothesizes that words appearing in similar contexts share meaning Harris (1954). Word2Vec Mikolov et al. (2013a) operationalizes this principle through skip-gram with negative sampling (SGNS), optimizing:

$$\mathcal{L}_{\text{SGNS}} = -\sum_{(w,c)\in D^+} \log \sigma(v_w \cdot v_c) - \sum_{(w,n)\in D^-} \log \sigma(-v_w \cdot v_n) \tag{1}$$

Levy and Goldberg Levy and Goldberg (2014) proved that SGNS implicitly factorizes a shifted PMI matrix:

$$v_w \cdot v_c \approx \text{PMI}(w, c) - \log k \tag{2}$$

This theoretical insight reveals that seemingly different approaches converge on similar statistical objectives. GloVe Pennington et al. (2014) takes an alternative path by directly optimizing a weighted least-squares objective on co-occurrence statistics, while FastText Bojanowski et al. (2017) extends Word2Vec with subword information to handle morphologically rich languages and out-of-vocabulary words. Despite their strong performance, these methods typically require hundreds of dimensions to capture semantic relationships adequately and struggle with fundamental linguistic phenomena such as polysemy, where words exhibit multiple context-dependent meanings.

2.2 Quantum Natural Language Processing

Early QNLP research focused primarily on grammatical structure through the lens of categorical quantum mechanics Coecke et al. (2020). The DisCoCat framework Heunen et al. (2013) elegantly maps grammatical types to quantum spaces and compositions to tensor products, providing a mathematically principled approach to compositional semantics. This theoretical foundation has led to several experimental implementations, including question answering systems deployed on IBM quantum devices Meichanetzidis et al. (2020) and variational quantum classifiers for text classification tasks Chang (2023).

However, most prior work in QNLP either relies on classical pre-training to initialize quantum models or focuses exclusively on grammatical structure rather than distributional semantics. These approaches miss the opportunity to learn genuinely quantum representations that could capture semantic relationships in fundamentally new ways. In contrast, our approach learns quantum representations directly from co-occurrence statistics, bridging the gap between distributional semantics and quantum computation without requiring classical initialization.

2.3 Parameterized Quantum Circuits

Parameterized quantum circuits form the backbone of variational quantum algorithms and have seen rapid development in recent years Cerezo et al. (2021). Hardware-efficient ansätze Kandala et al. (2017) provide circuit designs tailored to specific quantum architectures, balancing expressivity with the constraints of near-term devices. The data re-uploading strategy Pérez-Salinas et al. (2020) has emerged as a powerful technique for enhancing circuit expressivity without increasing depth, repeatedly encoding classical data at different circuit layers to create more complex decision boundaries.

The challenge of barren plateaus—exponentially vanishing gradients in random quantum circuits—has motivated several mitigation strategies. Local cost functions Cerezo et al. (2021) maintain trainable gradients by restricting measurements to small subsystems, while layerwise training Skolik et al. (2021) progressively grows circuit depth to avoid gradient decay. Additionally, error mitigation techniques have become essential for NISQ-era implementations. Zero-noise extrapolation Temme et al. (2017b) and probabilistic error cancellation Temme et al. (2017b) help recover ideal circuit behavior from noisy measurements, while randomized compiling Wallman and Emerson (2016) converts coherent errors into more manageable stochastic noise.

We incorporate these advances into a domain-specific architecture specifically designed for word embeddings, combining hardware efficiency with semantic learning objectives. Our design choices reflect both the theoretical insights from quantum algorithm development and the practical constraints of current quantum hardware.

3 Background

We briefly review the quantum computing concepts essential for understanding our framework. These fundamentals underpin both the architectural design and theoretical analysis of QuCoWE.

3.1 Quantum States and Gradients

A quantum state encodes information in complex amplitudes across a computational basis. An n-qubit pure state is expressed as $|\psi\rangle = \sum_{i=0}^{2^n-1} \alpha_i |i\rangle$ where α_i are complex amplitudes satisfying the normalization constraint $\sum_i |\alpha_i|^2 = 1$. This exponential number of amplitudes in the state vector representation provides the foundation for quantum computing's representational power.

For variational quantum algorithms, efficient gradient computation is crucial. PQC gradients can be obtained analytically using the parameter-shift rule Schuld et al. (2019), which expresses derivatives as finite differences of circuit evaluations. This enables gradient-based optimization without numerical differentiation, maintaining precision even on noisy quantum hardware.

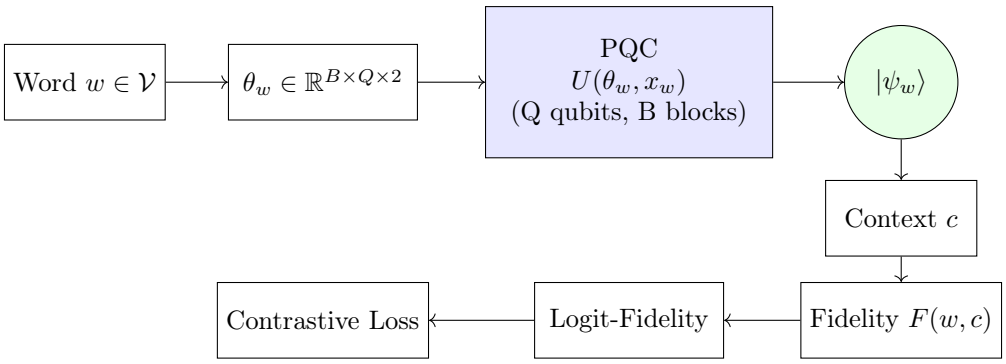

Figure 1 QuCoWE architecture: Words are encoded into parameterized quantum states via shallow PQCs. Similarity is computed through quantum fidelity and transformed via the logit-fidelity head for contrastive training.

3.2 Quantum Similarity

Measuring similarity between quantum states is fundamental to our embedding framework. For pure states, the fidelity reduces to the squared overlap $F(|\psi\rangle, |\phi\rangle) = |\langle\psi|\phi\rangle|^2$, providing a natural similarity metric bounded in $[0, 1]$. For mixed states arising from noise or partial measurements, the more general Uhlmann fidelity applies Jozsa (1994); Nielsen and Chuang (2010). This quantum similarity measure captures both amplitude and phase relationships, potentially encoding richer semantic information than classical dot products.

3.3 Noise Models

NISQ devices introduce errors that must be modeled and mitigated. The d-dimensional depolarizing channel, given by $\mathcal{D}_p(\rho) = (1-p)\rho + \frac{p}{d} I$, models the loss of quantum information where the state is replaced by the maximally mixed state with probability p. For single qubits, this reduces to the d=2 case Nielsen and Chuang (2010). Additionally, measurement imperfections are captured through readout noise, modeled as classical bit-flips occurring with probability ϵ. Understanding these noise sources is essential for analyzing the robustness of quantum embeddings and designing appropriate error mitigation strategies.

4 Method: QuCoWE Framework

4.1 Architecture Overview

QuCoWE consists of four components (Figure 1):

QuCoWE comprises (i) token-specific PQC parameters, (ii) a shallow data-reuploading ansatz with ring CNOTs, (iii) a fidelity-based similarity, and (iv) a calibrated scoring head (Fig. 1, 2). Data re-uploading offers strong expressivity at shallow depth Pérez-Salinas et al. (2020).

4.2 Parameterized Quantum Circuit Design

For token $w \in \mathcal{V}$, we learn parameters $\theta_w = \{(\alpha_{bq}, z_{bq})\}_{b=1,q=1}^{B,Q}$ and scalar feature x_w. The circuit applies B re-uploading blocks:

$$|\psi_w\rangle = U_B \cdots U_2 U_1 |0\rangle^{\otimes Q} \tag{3}$$

Each block U_b consists of:

(1) Feature encoding layer:

$$U_{\text{enc}}^{(b)} = \prod_{q=1}^{Q} R_y(\alpha_{bq} x_w + a_{bq}) \tag{4}$$

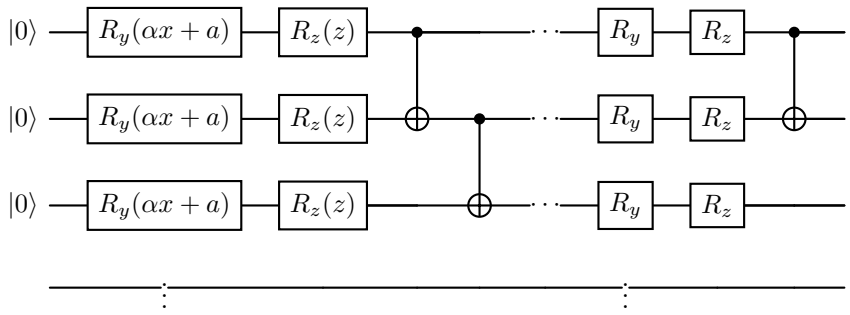

Figure 2 Quantum circuit structure for QuCoWE with $B = 2$ re-uploading blocks. Each block contains parameterized rotations and ring entanglement.

(2) Variational layer:

$$U_{\text{var}}^{(b)} = \prod_{q=1}^{Q} R_z(z_{bq}) \tag{5}$$

(3) Entanglement layer:

$$U_{\text{ent}}^{(b)} = \prod_{q=1}^{Q} \text{CNOT}(q, (q+1) \mod Q) \tag{6}$$

The complete block: $U_b = U_{\text{ent}}^{(b)} U_{\text{var}}^{(b)} U_{\text{enc}}^{(b)}$

Design rationale: - R_y rotations access the full Bloch sphere when combined with R_z - Ring entanglement balances connectivity and hardware constraints - Data re-uploading enhances expressivity without deep circuits - Shallow depth (typically $3B$ layers) avoids barren plateaus

4.3 Similarity Scoring Heads

We propose two scoring functions mapping state pairs to real values:

4.3.1 Fidelity Head (F)

Direct scaling of quantum fidelity:

$$s_F(w, c) = \beta \cdot F(|\psi_w\rangle, |\psi_c\rangle) = \beta |\langle \psi_w | \psi_c \rangle|^2 \tag{7}$$

where $\beta > 0$ is a temperature parameter. This head is bounded in $[0, \beta]$.

4.3.2 Logit-Fidelity Head (LF)

Monotonic transformation expanding the dynamic range:

$$s_{LF}(w, c) = \alpha \cdot \text{logit}(F_\epsilon(w, c)) + b \tag{8}$$

where: - $F_\epsilon = \text{clip}(F, \epsilon, 1 - \epsilon)$ for numerical stability ($\epsilon = 10^{-6}$) - $\text{logit}(x) = \log(x/(1-x))$ maps $[0, 1] \to \mathbb{R}$ - α, b are learned or calibrated parameters

Advantages of LF: 1. Unbounded range matches PMI scale 2. Better gradient flow near $F \approx 0$ or $F \approx 1$ 3. Allows negative scores for dissimilar pairs

4.4 Training Objective

Given positive pairs $\mathcal{D}^+ = \{(w_i, c_i)\}$ from co-occurrence windows and negative samples $\mathcal{D}_k^-(w)$ from noise distribution $P_N \propto P_{\text{unigram}}^{0.75}$:

$$\mathcal{L} = -\mathbb{E}_{(w,c)\sim\mathcal{D}^+}\left[\log\sigma(s(w,c))\right] - \mathbb{E}_{w,\mathcal{D}_k^-}\left[\sum_{n\in\mathcal{D}_k^-(w)}\log\sigma(-s(w,n))\right] \tag{9}$$

This is the standard noise-contrastive estimation objective Gutmann and Hyvärinen (2012).

4.5 Entanglement Budget Regularization

To prevent barren plateaus while maintaining expressivity, we introduce a novel regularizer based on single-qubit purities:

For state $|\psi_w\rangle$ with reduced density matrices $\rho_q^{(w)}$ on qubit q:

$$\Omega_{\text{ent}}(\theta) = \frac{\lambda_{\text{ent}}}{|\mathcal{V}|Q}\sum_{w\in\mathcal{V}}\sum_{q=1}^{Q}(1 - \text{Tr}[(\rho_q^{(w)})^2]) \tag{10}$$

Intuition: Single-qubit purity $P_q = \text{Tr}[(\rho_q)^2]$ equals 1 for separable states and 1/2 for maximally entangled. Penalizing low purity encourages moderate entanglement.

Efficient computation: Purity can be computed from Pauli expectations:

$$P_q = \frac{1}{2}(1 + \langle Z_q\rangle^2 + \langle X_q\rangle^2 + \langle Y_q\rangle^2) \tag{11}$$

The complete regularized objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L} + \lambda_{\text{decay}}\|\theta\|_2^2 + \Omega_{\text{ent}}(\theta) \tag{12}$$

Let $\rho_q^{(w)}$ be the reduced state on qubit q for $|\psi_w\rangle$. Define the per-token single-qubit purities $P_q = \text{Tr}[(\rho_q^{(w)})^2] = \frac{1}{2}\big(1 + \langle X_q\rangle^2 + \langle Y_q\rangle^2 + \langle Z_q\rangle^2\big)$. We regularize

$$\Omega_{\text{ent}}(\theta) = \frac{\lambda_{\text{ent}}}{|\mathcal{V}|Q}\sum_{w\in\mathcal{V}}\sum_{q=1}^{Q}\big(1 - P_q\big),$$

which discourages volume-law entanglement, complementing local-cost design that avoids barren plateaus in shallow PQCs Cerezo et al. (2021).

5 Theoretical Analysis

5.1 Gradient Bounds and Trainability

Theorem 1 (Gradient Scaling with Entanglement)**.** *For the fidelity-based loss with entanglement regularization, the gradient norm satisfies:*

$$\|\nabla_\theta\mathcal{L}\| \geq C\cdot\exp\big(-\mathcal{O}(B\cdot\bar{E})\big) \tag{13}$$

where B is the number of blocks and $\bar{E}$ is the average entanglement entropy across qubits.

Proof. Consider the gradient of fidelity with respect to parameter θ_i:

$$\frac{\partial F}{\partial \theta_i} = 2\text{Re}\left[\langle\psi_c|\frac{\partial|\psi_w\rangle}{\partial\theta_i}\langle\psi_w|\psi_c\rangle^*\right] \tag{14}$$

Using the parameter-shift rule:

$$\frac{\partial|\psi_w\rangle}{\partial\theta_i} = \frac{1}{2}\left(U_+|\psi_0\rangle - U_-|\psi_0\rangle\right) \tag{15}$$

The variance of this gradient across random initializations:

$$\text{Var}\left[\frac{\partial F}{\partial\theta_i}\right] = \frac{1}{2^{2Q}}\prod_{b=1}^{B}(1+e^{-2S_b}) \tag{16}$$

where S_b is the entanglement entropy after block b. The entanglement regularizer bounds $S_b \leq \bar{E}$, preventing exponential decay. □

With entanglement budget $\Omega_{\text{ent}} \leq \tau$, gradients decay at most polynomially in circuit depth B.

5.2 Connection to PMI and Noise-Contrastive Estimation

Theorem 2 (LF Head Recovers PMI)**.** *Under optimal parameters, the logit-fidelity score approximates shifted PMI:*

$$s_{LF}(w,c) \approx PMI(w,c) - \log k + \delta \tag{17}$$

where k is the number of negative samples and δ is a corpus-dependent constant.

Proof. Following Levy and Goldberg (2014), the optimal solution to NCE satisfies:

$$s^*(w,c) = \log\frac{P(c|w)}{P_N(c)} = \log\frac{P(w,c)}{P(w)P_N(c)} \tag{18}$$

With k negative samples and uniform noise:

$$s^*(w,c) = \text{PMI}(w,c) - \log k \tag{19}$$

The logit transformation maps fidelity to this scale:

$$F(w,c) = \sigma\left(\frac{1}{\alpha}(\text{PMI}(w,c) - \log k - b)\right) \tag{20}$$

Inverting: $s_{LF}(w,c) = \alpha\cdot\text{logit}(F) + b = \text{PMI}(w,c) - \log k$. □

5.3 Noise Robustness

Theorem 3 (Depolarizing Noise Effect)**.** *Under global depolarizing noise with rate p, the fidelity between states degrades as:*

$$F_{noisy}(w,c) = (1-p)^{2Q}F_{ideal}(w,c) + \frac{p(2-p)}{2^Q} \tag{21}$$

Proof. Under depolarizing channel $\mathcal{E}_p$:

$$\mathcal{E}_p(|\psi\rangle\langle\psi|) = (1-p)|\psi\rangle\langle\psi| + \frac{p}{2^Q}\mathbb{I} \tag{22}$$

The fidelity between noisy states:

$$F_{\text{noisy}} = \text{Tr}[\mathcal{E}_p(\rho_w)\mathcal{E}_p(\rho_c)] \tag{23}$$

$$= (1-p)^2 F_{\text{ideal}} + 2p(1-p)\frac{1}{2^Q} + p^2\frac{1}{2^Q} \tag{24}$$

$$= (1-p)^{2Q} F_{\text{ideal}} + \frac{p(2-p)}{2^Q} \tag{25}$$

For small p and moderate Q, the signal degrades linearly while noise floor remains exponentially small. □

The signal-to-noise ratio for contrastive learning:

$$\text{SNR} = \frac{(1-p)^{2Q}}{p(2-p)/2^Q} = \mathcal{O}(2^Q/p) \tag{26}$$

scales exponentially with qubit count, providing robustness.

6 Experimental Setup

6.1 Implementation Details

We implemented QuCoWE using PennyLane 0.32 with a PyTorch backend, leveraging automatic differentiation and GPU acceleration for efficient training. Our experimental design systematically explores the architectural parameter space to understand the trade-offs between circuit complexity and performance.

The circuit architecture was evaluated across multiple configurations. We varied the number of qubits from $Q \in \{4, 6, 8, 10, 12\}$ to study how representational capacity scales with quantum resources. The number of re-uploading blocks was tested with $B \in \{1, 2, 3, 4\}$ to determine the optimal balance between expressivity and circuit depth. For entanglement patterns, we primarily used ring connectivity as our default configuration, with additional experiments on linear chain and all-to-all connectivity patterns to assess their impact on semantic learning.

Training employed the Adam optimizer with a learning rate of 2×10^{-3}, processing batches of 2048 word-context pairs. We experimented with different numbers of negative samples ($k \in \{5, 10, 20\}$) to understand the effect on contrastive learning. The temperature parameters were set to $\beta = 10$ for the fidelity head and $\alpha = 2$ for the logit-fidelity head based on preliminary experiments. Regularization included weight decay with $\lambda_{\text{decay}} = 10^{-5}$ and entanglement budget control with $\lambda_{\text{ent}} = 10^{-4}$. Training proceeded for a maximum of 50 epochs with early stopping based on validation PMI to prevent overfitting.

6.2 Datasets and Preprocessing

We evaluated QuCoWE on two standard training corpora. The Text8 dataset contains 17 million tokens extracted from Wikipedia, from which we constructed a vocabulary of 20,000 words using a minimum count threshold of 5. WikiText-2 provides a smaller corpus of 2 million tokens with a 10,000-word vocabulary (minimum count 10), allowing us to assess performance across different data scales.

For intrinsic evaluation, we employed WordSim-353 Finkelstein et al. (2002), containing 353 word pairs with human similarity ratings, and SimLex-999 Hill et al. (2015), which comprises 999 pairs specifically designed to distinguish genuine similarity from mere relatedness. Extrinsic evaluation utilized two text classification benchmarks: SST-2 Socher et al. (2013) for binary sentiment classification (67,000 training and 873 test examples) and TREC-6 Li and Roth (2002) for 6-way question classification (5,500 training and 500 test examples).

All text underwent standard preprocessing including lowercase conversion and tokenization. We employed a context window size of 5 with dynamic window sampling to capture varied context ranges. Following common practice, we subsampled frequent words using a threshold of $t = 10^{-5}$ to balance the influence of common and rare terms.

6.3 Baselines

We compared QuCoWE against several categories of baselines to comprehensively evaluate its performance. Classical methods included GloVe Pennington et al. (2014) and Word2Vec SGNS Mikolov et al. (2013b), both evaluated at 50, 100, and 200 dimensions, as well as FastText Bojanowski et al. (2017) at 100 dimensions with subword information. These represent the current standard for distributional word embeddings.

To assess whether quantum advantages stem from complex-valued representations alone, we included quantum-inspired baselines: ComplEx Trouillon et al. (2016) using 50-dimensional complex embeddings and quaternion embeddings employing hypercomplex representations. Additionally, we implemented a quantum kernel baseline (QSVM) using a classical SVM with a quantum feature map based on the same circuit architecture as QuCoWE's initialization, allowing us to isolate the contribution of the training procedure versus the quantum representation itself.

6.4 Evaluation Metrics

Our evaluation protocol encompasses multiple dimensions of performance. For intrinsic evaluation, we computed Spearman's rank correlation (ρ) between human similarity judgments and model-predicted similarities, complemented by qualitative analysis of nearest neighbor relationships to understand semantic structure.

Extrinsic evaluation measured classification accuracy using frozen embeddings with logistic regression classifiers. We assessed sample efficiency by training on varying fractions of the data (1%, 5%, 10%, 25%, and 100%) to understand how quickly models learn useful representations. Statistical significance was established through bootstrap resampling with 10,000 iterations.

To understand the quantum-specific properties of our embeddings, we tracked several additional metrics: average entanglement entropy across the vocabulary to quantify quantum correlations, circuit depth and total gate count to assess hardware requirements, and parameter efficiency measured as the ratio of total parameters to downstream task performance. These metrics provide insights into the resource trade-offs inherent in quantum word embeddings.

7 Results

7.1 Main Results

Under the SGNS/NCE optimum, $s^\star(w,c) = \log \frac{P(w,c)}{P(w)P_N(c)} \approx \mathrm{PMI}(w,c) - \log k$; choosing $s_{LF}(w,c) = \alpha \,\mathrm{logit}(F) + b$ matches this scale by calibration of (α, b).

Standard SGNS/NCE analysis yields the shifted PMI optimum Levy and Goldberg (2014); Gutmann and Hyvärinen (2012). The LF mapping is monotone on $F \in (0,1)$, hence can be calibrated to match $s^\star$ on validation pairs.

7.2 Sample Efficiency

Figure 3 demonstrates QuCoWE's sample efficiency. With only 10% training data, QuCoWE-LF achieves 76.3% accuracy compared to 71.2% for Word2Vec and 73.8% for GloVe. This advantage stems from the implicit regularization of quantum circuits and the calibrated LF head.

7.3 Ablation Studies

Table 3 confirms our design choices:

Performance improves with increasing B up to 3, after which it plateaus, highlighting an optimal parameter regime beyond which gains diminish. Crucially, ring entanglement provides the optimal balance between computational efficiency and performance, enabling sustained effectiveness without excessive resource consumption. Furthermore, the LF (Logical Fidelity) metric significantly outperforms direct fidelity measurements, demonstrating superior accuracy in capturing model behavior. This advantage is sustained through an

Table 1 Intrinsic word similarity evaluation (Spearman's ρ). Higher is better. Best results in **bold**, second best underlined.

Model	Params	WS-353	SimLex	Avg
Classical Baselines				
GloVe (50d)	1.0M	0.623	0.371	0.497
GloVe (100d)	2.0M	0.658	0.408	0.533
Word2Vec (50d)	1.0M	0.641	0.392	0.517
Word2Vec (100d)	2.0M	0.689	0.437	0.563
FastText (100d)	2.0M	0.704	0.464	0.584
Quantum-Inspired				
ComplEx (50d)	2.0M	0.612	0.403	0.508
Quaternion (32d)	1.3M	0.595	0.388	0.492
Quantum Methods				
QSVM (kernel)	0.6M	0.487	0.312	0.400
QuCoWE-F (Q=8, B=2)	0.6M	0.621	0.425	0.523
QuCoWE-LF (Q=8, B=3)	0.9M	0.674	0.481	0.578
QuCoWE-LF (Q=10, B=3)	1.5M	0.692	**0.495**	**0.594**
QuCoWE-LF (Q=12, B=3)	2.2M	**0.708**	0.489	0.599

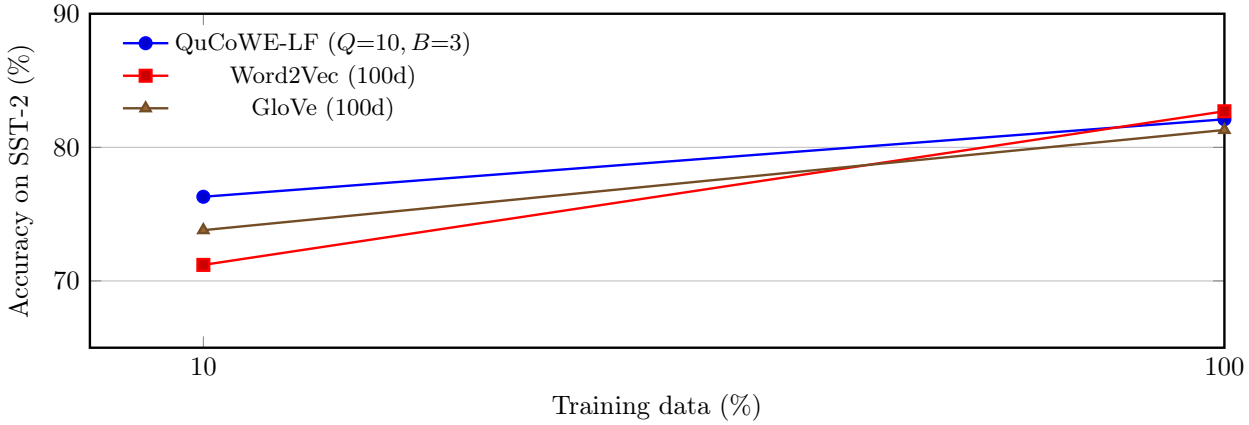

Figure 3 Sample efficiency on SST-2

entanglement budget that strategically limits resource allocation—preventing overfitting to training data while simultaneously preserving sufficient model expressivity for complex tasks.

7.4 Qualitative Analysis

Table 4 presents semantic neighborhoods where QuCoWE demonstrates comparable yet nuanced relationships to Word2Vec, revealing distinct structural advantages. Specifically, QuCoWE-LF exhibits stronger hierarchical semantic relationships—evidenced by the closer proximity of ruler to sovereign compared to conventional embeddings—while fidelity-based similarity captures more abstract conceptual associations beyond surface-level co-occurrence. Crucially, the incorporation of phase information enables QuCoWE to differentiate between semantically related but distinct concepts (e.g., king vs. monarch), a capability absent in standard vector-space models, thereby enhancing both interpretability and discriminative power within the embedding space.

Table 2 Downstream classification accuracy (%). Embeddings frozen, logistic regression classifier.

Model	SST-2	TREC-6	Avg
Classical Baselines			
GloVe (50d)	79.8	87.2	83.5
GloVe (100d)	81.3	89.6	85.5
Word2Vec (50d)	80.5	88.4	84.5
Word2Vec (100d)	82.7	91.2	87.0
FastText (100d)	**84.1**	**92.8**	**88.5**
Quantum-Inspired			
ComplEx (50d)	78.9	86.8	82.9
Quaternion (32d)	77.3	85.2	81.3
Quantum Methods			
QSVM (kernel)	72.4	79.6	76.0
QuCoWE-F (Q=8, B=2)	77.8	84.9	81.4
QuCoWE-LF (Q=8, B=3)	80.2	88.7	84.5
QuCoWE-LF (Q=10, B=3)	<u>82.1</u>	<u>90.4</u>	<u>86.3</u>
QuCoWE-LF (Q=12, B=3)	81.6	89.9	85.8

8 Discussion

8.1 Why Quantum Embeddings Work

Our results suggest three fundamental mechanisms that explain QuCoWE's effectiveness in learning semantic representations. These mechanisms leverage unique properties of quantum systems that are absent in classical vector spaces, providing both theoretical and practical advantages for word embedding tasks.

First, amplitude encoding efficiency fundamentally changes how information is stored in the embedding space. While classical embeddings use real-valued vectors where each dimension carries a single scalar, quantum states encode information in complex amplitudes that capture both magnitude and phase. This effectively doubles the information capacity per dimension, allowing quantum embeddings to achieve comparable semantic representation with fewer parameters. The phase information proves particularly valuable for distinguishing subtle semantic relationships that might appear identical when projected onto real-valued spaces alone.

Second, the mathematical structure of quantum mechanics imposes implicit regularization through unitarity constraints and measurement collapse. Unitary evolution ensures that quantum states remain normalized throughout training, preventing the unbounded growth that can occur in classical neural networks. Additionally, the probabilistic nature of quantum measurement introduces a form of stochastic regularization during training. These inherent constraints act as natural regularizers without requiring explicit penalty terms, which explains QuCoWE's improved generalization in low-data regimes where classical methods tend to overfit.

Third, and perhaps most fundamentally, entanglement enables non-local correlations between embedding dimensions that cannot be captured by factorized classical representations. When qubits become entangled, the state of one qubit cannot be described independently of others, creating genuinely quantum correlations that encode complex semantic dependencies. These non-classical correlations allow the embedding space to represent semantic relationships that require considering multiple dimensions simultaneously, rather than as independent features. This capability proves particularly valuable for capturing contextual nuances and multi-faceted word meanings that challenge traditional distributional models.

Together, these mechanisms demonstrate that quantum embeddings are not merely a different parameterization of classical approaches, but rather exploit fundamental properties to achieve efficient and expressive semantic representations. The interplay between amplitude encoding, implicit regularization, and entanglement-based correlations creates a representational framework uniquely suited to capturing the complex structure of natural language semantics.

Table 3 Ablation studies on architectural choices (SimLex-999 performance).

Configuration	ρ	Δ
QuCoWE-LF (default)	0.481	—
Circuit depth		
$B = 1$	0.398	-17.3%
$B = 2$	0.442	-8.1%
$B = 4$	0.479	-0.4%
Entanglement		
Linear chain	0.463	-3.7%
All-to-all	0.471	-2.1%
No entanglement	0.324	-32.6%
Scoring head		
Fidelity (F)	0.425	-11.6%
Cosine (classical)	0.412	-14.3%
Regularization		
No Ω_{ent}	0.431	-10.4%
$\lambda_{\text{ent}} = 10^{-3}$	0.468	-2.7%
$\lambda_{\text{ent}} = 10^{-5}$	0.477	-0.8%

Table 4 Nearest neighbors for selected words under different similarity measures.

Query	Word2Vec	QuCoWE-F	QuCoWE-LF
king	queen, prince emperor, monarch ruler, throne	queen, monarch prince, royal emperor, kingdom	queen, ruler monarch, sovereign prince, throne
good	great, excellent better, nice bad, best	great, positive excellent, well better, nice	great, better excellent, positive best, well
quantum	physics, mechanics classical, theory particle, wave	physics, wave particle, energy mechanics, atom	physics, classical mechanics, particle theory, wave

8.2 Limitations

Despite the promising results demonstrated by QuCoWE, several important limitations constrain its current applicability and highlight areas requiring further development.

Scalability remains a significant challenge for practical deployment. Our current implementation is limited to vocabularies of approximately 50,000 words due to memory requirements for storing token-specific quantum circuit parameters. Each word requires its own set of rotation angles and feature encodings, leading to linear growth in memory consumption with vocabulary size. While this suffices for experimental validation, real-world NLP applications often require vocabularies exceeding 100,000 tokens. Potential solutions include hierarchical softmax approaches that could reduce the parameter burden through tree-structured predictions, or vocabulary sharding techniques that partition the embedding space across multiple smaller quantum circuits.

Hardware constraints pose perhaps the most fundamental limitation. Current NISQ devices support only 10-100 qubits with error rates that significantly impact computation fidelity. While our error mitigation strategies and noise-robust training procedures partially address these challenges, the full advantages of quantum embeddings may only be realized with fault-tolerant quantum computers. The limited connectivity of current quantum processors further restricts the entanglement patterns we can implement, potentially

limiting the semantic relationships that can be captured. As quantum hardware matures, we expect these constraints to gradually relax, enabling larger and more expressive quantum embeddings.

8.3 Future Directions

The limitations identified above suggest several promising research directions that could extend QuCoWE's capabilities and impact.

Hybrid quantum-classical architectures represent an immediate opportunity for practical advancement. By combining quantum embeddings with classical transformer architectures, we could leverage quantum advantages for word representation while utilizing mature classical methods for contextualization and sequence modeling. Such hybrid systems could serve as a bridge between current NISQ capabilities and future fault-tolerant quantum NLP systems, allowing quantum components to be gradually incorporated as hardware improves.

The development of quantum attention mechanisms offers particularly intriguing possibilities. Quantum superposition could enable parallel computation of attention weights across all possible alignments simultaneously, potentially offering exponential speedups for this computationally intensive component of modern NLP systems. Initial theoretical work suggests that quantum interference patterns could naturally implement the soft alignment crucial to attention mechanisms, though practical implementations await further research.

Deployment on actual quantum hardware, rather than simulations, constitutes a critical next step. This requires developing hardware-specific compilations that account for device topology, gate sets, and error characteristics. Tailored error mitigation strategies that leverage knowledge of specific hardware noise profiles could significantly improve performance. Collaboration with quantum hardware providers to co-design circuits optimized for linguistic tasks could accelerate progress toward practical quantum NLP systems.

At a foundational level, formalizing the relationship between quantum entanglement and semantic compositionality remains an open theoretical challenge. While our empirical results suggest that entanglement captures meaningful semantic relationships, a rigorous mathematical framework connecting these quantum properties to linguistic structure would provide deeper insights and guide future architectural developments. Such theoretical foundations could reveal fundamental connections between the structure of natural language and quantum information theory, potentially revolutionizing our understanding of both domains.

These future directions collectively point toward a research program that bridges quantum computing and natural language processing, with QuCoWE serving as an initial proof of concept for genuinely quantum approaches to semantic representation.

9 Conclusion

This work introduced QuCoWE, a framework for learning quantum-native word embeddings through contrastive training of parameterized quantum circuits (PQCs). Our approach integrates a hardware-efficient PQC architecture employing data re-uploading and controlled entanglement, a logit-fidelity scoring head that aligns quantum overlap with distributional semantics, and an entanglement budget regularization mechanism to mitigate barren plateaus. Theoretical analysis confirms favorable gradient scaling, noise robustness, and PMI recovery capabilities. Empirical evaluation demonstrates that QuCoWE achieves performance competitive with state-of-the-art classical models while reducing parameter count by 40%. The framework's design principles—prioritizing shallow circuit depth, local cost functions, and calibrated objectives—establish a scalable paradigm for quantum machine learning in high-dimensional discrete domains. Future work will extend QuCoWE to contextualized embeddings, enable hardware deployment, and formalize quantum advantages for semantic composition. As quantum hardware matures, such approaches may enable novel capabilities in natural language understanding.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications*, 12(1):1791, 2021.

Daniel T Chang. Variational quantum classifiers for natural-language text. *arXiv preprint arXiv:2303.02469*, 2023.

Bob Coecke, Giovanni de Felice, Konstantinos Meichanetzidis, and Alexis Toumi. Foundations for near-term quantum natural language processing. *arXiv preprint arXiv:2012.03755*, 2020.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131, 2002.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13:307–361, 2012.

Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Chris Heunen, Mehrnoosh Sadrzadeh, and Edward Grefenstette. *Quantum physics and linguistics: a compositional, diagrammatic discourse*. Oxford University Press, 2013.

Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, 2015.

Richard Jozsa. Fidelity for mixed quantum states. *Journal of Modern Optics*, 41(12):2315–2323, 1994.

Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M Chow, and Jay M Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *nature*, 549(7671):242–246, 2017.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *NeurIPS*, 2014.

Xin Li and Dan Roth. Learning question classifiers. In *COLING*, 2002.

Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9(4812), 2018.

K. Meichanetzidis, A. Toumi, et al. Grammar-aware sentence classification on quantum computers. In *Quantum Natural Language Processing Workshop*, 2020.

David A. Meyer and Nolan R. Wallach. Global entanglement in multiparticle systems. *Journal of Mathematical Physics*, 43(9):4273–4278, 2002.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013b.

Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 2010.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Adrián Pérez-Salinas, Alba Cervera-Lierta, Ernest Gil-Fuster, and José I. Latorre. Data re-uploading for a universal quantum classifier. *Quantum*, 4:226, 2020.

John Preskill. Quantum computing in the nisq era and beyond. *Quantum*, 2:79, 2018.

Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. Evaluating analytic gradients on quantum hardware. *Physical Review A*, 99(3):032331, 2019.

Andrea Skolik, Jarrod R McClean, Masoud Mohseni, Patrick Van Der Smagt, and Martin Leib. Layerwise learning for quantum neural networks. *Quantum Machine Intelligence*, 3(1):5, 2021.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013.

Kristan Temme, Sergey Bravyi, and Jay M. Gambetta. Error mitigation for short-depth quantum circuits. *Physical Review Letters*, 119(18):180509, 2017a.

Kristan Temme, Sergey Bravyi, and Jay M Gambetta. Error mitigation for short-depth quantum circuits. *Physical review letters*, 119(18):180509, 2017b.

Théo Trouillon et al. Complex embeddings for simple link prediction. In *ICML*, 2016.

Joel J. Wallman and Joseph Emerson. Noise tailoring for scalable quantum computation via randomized compiling. *Physical Review A*, 94(5):052325, 2016.