Quantum Natural Language Processing: A Comprehensive Survey of Models, Architectures, and Evaluation Methods

Arpita Vats
Boston University
LinkedIn*

Rahul Raja
Carnegie Mellon University
Stanford University
LinkedIn*

Ashish Kattamuri
Proofpoint*

Abhinav Bohra
Amazon*

Abstract

Quantum Natural Language Processing (QNLP) is an emerging interdisciplinary field at the intersection of quantum computing, natural language understanding, and formal linguistic theory. As advances in quantum hardware and algorithms accelerate, QNLP promises new paradigms for representation learning, semantic modeling, and efficient computation. However, existing literature remains fragmented, with no unified synthesis across modeling, encoding, and evaluation dimensions.In this work, we present the first systematic and taxonomy driven survey of QNLP that holistically organizes research spanning three core dimensions: computational models, encoding paradigms, and evaluation frameworks. First, we analyze foundational approaches that map linguistic structures into quantum formalism, including categorical compositional models, variational quantum circuits, and hybrid quantum classical architectures. Second, we introduce a unified taxonomy of encoding strategies, ranging from quantum tokenization and state preparation to embedding based encodings, highlighting tradeoffs in scalability, noise resilience, and expressiveness. Third, we provide the first comparative synthesis of evaluation methodologies, benchmark datasets, and performance metrics, while identifying reproducibility and standardization gaps.We further contrast quantum inspired NLP methods with fully quantum implemented systems, offering insights into resource efficiency, hardware feasibility, and real world applicability. Finally, we outline open challenges such as integration with LLMs and unified benchmark design, and propose a research agenda for advancing QNLP as a scalable and reliable discipline.

1 Introduction

The intersection of quantum computing and natural language processing (NLP) has given rise to the emerging field of QNLP. Traditional NLP methods rely heavily on classical statistical and neural approaches, which, despite recent breakthroughs in LLMs (Brown et al., 2020), face fundamental challenges in scalability, representation efficiency, and capturing complex compositional semantics (Bender et al., 2021). Quantum computing, with its inherent parallelism and high-dimensional Hilbert space representations, offers a fundamentally new computational paradigm that can potentially overcome some of these limitations (Meichanetzidis et al., 2020; Varmantchaonala et al., 2024). Specifically, quantum models promise exponential speedups in linear algebra operations, richer encoding of linguistic structures, and novel mechanisms for semantic composition grounded in quantum theory. Foundational frameworks such as categorical compositional distributional models (DisCoCat) (Coecke et al., 2010) leverage quantum formalisms to represent grammatical structure, while hybrid quantum classical architectures demonstrate the feasibility of encoding word embeddings and performing sentence classification tasks on near-term quantum hardware (Lorenz et al., 2021b). Recent work further explores quantum algorithms for compositional text processing (Zhang et al., 2024) and surveys near term QNLP applications (Wiebe et al., 2024).

This paper provides a systematic survey of QNLP across three core dimensions: (i) computational models that define how linguistic structure and semantics can be mapped to quantum circuits and algorithms, (ii) encoding paradigms that determine how text tokens, syntactic dependencies, or embeddings are represented in quantum states, and (iii) evaluation frameworks that assess the effectiveness, efficiency, and robustness of QNLP methods. By categorizing and analyzing existing approaches, we highlight key tradeoffs in expressiveness, scalability, and noise resilience. Furthermore, we contrast quantum inspired NLP tech-

[*] This work does not relate to the authors' positions at LinkedIn, Proofpoint, or Amazon.

Proceedings of the QuantumNLP: Integrating Quantum Computing with Natural Language Processing, pages 53–64
November 24, 2025 ©2025 Association for Computational Linguistics

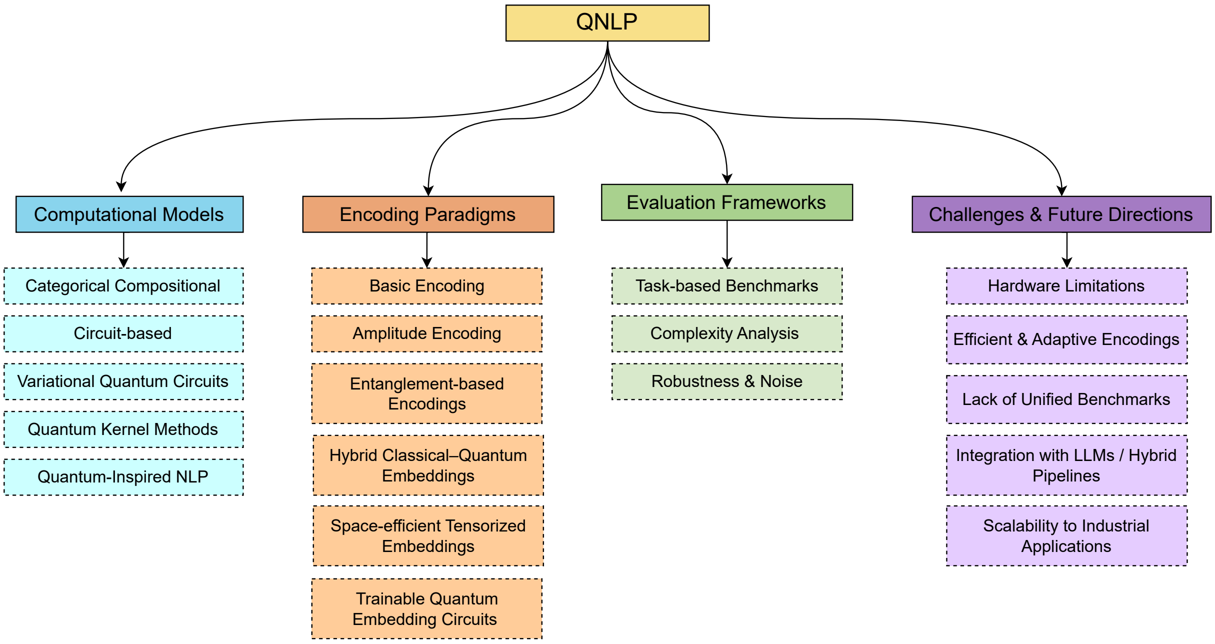
QNLP
Computational Models
Categorical Compositional
Circuit-based
Variational Quantum Circuits
Quantum Kernel Methods
Quantum-Inspired NLP
Encoding Paradigms
Basic Encoding
Amplitude Encoding
Entanglement-based Encodings
Hybrid Classical–Quantum Embeddings
Space-efficient Tensorized Embeddings
Trainable Quantum Embedding Circuits
Evaluation Frameworks
Task-based Benchmarks
Complexity Analysis
Robustness & Noise
Challenges & Future Directions
Hardware Limitations
Efficient & Adaptive Encodings
Lack of Unified Benchmarks
Integration with LLMs / Hybrid Pipelines
Scalability to Industrial Applications

Figure 1: Taxonomy of QNLP highlighting core components computational models, encoding paradigms, evaluation frameworks, and future challenges.

niques, which adapt ideas from quantum mechanics within classical settings, with implementations on actual quantum hardware, thereby clarifying both theoretical promise and current practical limitations. The overall evolution of QNLP approaches from foundational categorical frameworks to hybrid quantum classical architectures is illustrated in Figure 1, which presents the taxonomy of major model families and their interrelations across computational, encoding, and evaluation dimensions.

2 Background

2.1 Quantum Computing Fundamentals

Quantum computing leverages the laws of quantum mechanics to perform computations beyond the reach of classical machines. Its fundamental unit of information, the *qubit*, generalizes the classical bit by existing in a superposition of states. A quantum state $|\psi\rangle$ is a vector in a complex Hilbert space $\mathcal{H}$ (Moretti and Oppio, 2017), where the state of a single qubit can be expressed as:

$$|\psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle\,, \quad \alpha, \beta \in \mathbb{C}, \quad |\alpha|^2 + |\beta|^2 = 1.$$

Here, α and β are complex amplitudes, and the normalization condition ensures a probabilistic interpretation. Multiple qubits are represented through tensor products, e.g., $|\psi\rangle_{AB} = |\psi\rangle_A \otimes |\psi\rangle_B$. Entanglement arises when such states cannot be decomposed into tensor products, a phenomenon critical to quantum advantage in algorithms. Quantum computation is driven by unitary operators U acting on states:

$$|\psi'\rangle = U\,|\psi\rangle\,,$$

which ensure reversibility and preserve probability amplitudes. Measurement collapses the superposition into classical outcomes, with probabilities determined by the squared amplitudes of the state vector. Together, superposition, entanglement, unitary evolution, and measurement define the computational paradigm of quantum mechanics.

2.2 Quantum Machine Learning Foundations

QML studies how quantum mechanical principles can enhance or accelerate learning algorithms (Schuld et al., 2015). It leverages the expressive power of quantum states and the computational efficiency of quantum operations to address tasks in classification, regression, clustering, and generative modeling.

A key concept is the *quantum feature map*, which encodes classical data $x \in \mathbb{R}^d$ into a quantum state $|\phi(x)\rangle$ within a high-dimensional Hilbert space $\mathcal{H}$. This encoding induces a kernel:

$$k(x, x') = |\langle \phi(x) | \phi(x') \rangle|^2,$$

allowing quantum models to exploit feature spaces that may be exponentially larger than those accessible classically (Schuld et al., 2015). Quantum kernels have been investigated for support vector machines (SVMs) and nearest-neighbor methods, showing theoretical potential for improved separability.

Another foundational algorithm is the Harrow–Hassidim–Lloyd (HHL) method, which provides exponential speedups for solving linear systems of equations (Harrow et al., 2009). Since solving linear systems underpins many ML tasks (e.g., regression, Gaussian processes), HHL exemplifies how quantum algorithms could drastically reduce complexity from polynomial to logarithmic in the number of variables. In the near term, *variational quantum algorithms* (VQAs) have become the dominant paradigm for NISQ era devices (Cerezo et al., 2021b). These models use parameterized quantum circuits $U(\theta)$, where θ denotes tunable gate parameters, to transform input states. The circuits are trained by minimizing an objective function:

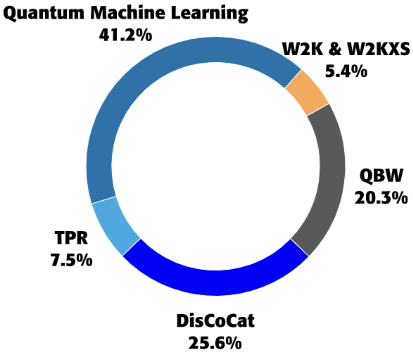

Figure 2: Adoption rates of QNLP models derived from the analyzed papers (Varmantchaonala et al., 2024).

$$C(\theta) = \langle\psi_0|U^\dagger(\theta)HU(\theta)|\psi_0\rangle,$$

with a classical optimizer updating θ based on quantum hardware evaluations. Variational circuits are flexible and have been applied to supervised learning (e.g., quantum classifiers), unsupervised tasks (e.g., clustering), and generative models.
Another critical building block is the *quantum neural network* (QNN), which uses variational circuits as analogues of neural layers. Entanglement plays a role similar to non-linear activation functions by enabling complex correlations between inputs. Hybrid QNNs combine quantum layers with classical networks, demonstrating performance gains in cases such as image and text classification.

From a complexity-theoretic perspective, QML offers potential advantages when classical methods suffer from the curse of dimensionality. Quantum states inhabit exponentially large Hilbert spaces naturally, enabling compact representation of complex data distributions. However, practical challenges remain, including noise, barren plateaus in variational optimization (McClean et al., 2018), and efficient data encoding (also known as the quantum data-loading problem).

For QNLP specifically, QML foundations provide the computational substrate: quantum feature maps offer new embedding paradigms for tokens, variational circuits serve as sequence-processing units, and entanglement provides a mechanism for modeling compositionality and long-range linguistic dependencies. These align with the goals of QNLP frameworks such as DisCoCat and hybrid quantum–classical pipelines, making QML an indispensable component of quantum language understanding.The distribution of QNLP model adoption across surveyed studies is shown in Figure 2, highlighting the dominance of Quantum Machine Learning based frameworks, followed by DisCoCat and Quantum Bag-of-Words models.

2.3 Natural Language Processing

NLP provides the computational basis for representing and interpreting linguistic data. Its core principle, *distributional semantics*, states that words appearing in similar contexts tend to have similar meanings. Early models such as Latent Semantic Analysis (LSA), Word2Vec, and GloVe encoded words as dense vectors $\mathbf{e}_w \in \mathbb{R}^d$, capturing semantic similarity through geometric proximity.

Modern NLP advances this idea through *contextual embeddings* using Transformer architectures such as BERT (Devlin et al., 2019), GPT (Radford et al., 2019), and T5 (Raffel et al., 2020). The self-attention mechanism

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

enables long-range dependency modeling by relating all tokens within a sequence (Vaswani et al., 2017). Despite their success, Transformers face $O(n^2)$ time and memory complexity with sequence length n, motivating efficient variants such as sparse and linearized attention.

Classical NLP also employs grammatical formalisms context-free grammars (CFGs), dependency parsing, and formal semantics to capture compositionality, yet integrating syntax with distributed semantics at scale remains challenging.

Quantum approaches address this limitation: quantum states in high-dimensional Hilbert spaces can encode inter-token dependencies through *entanglement*. Frameworks such as DisCoCat (Categorical Compositional Distributional Models) (Coecke et al., 2010) unify grammar and semantics via category theory, suggesting that QNLP can yield richer and more efficient representations than classical embeddings.

2.4 Quantum Classical Hybrids

Fully fault-tolerant quantum computers remain a long-term goal, but present-day devices fall into the category of Noisy Intermediate-Scale Quantum (NISQ) systems (Preskill, 2018). These machines contain on the order of 50–500 qubits, which are sufficient for exploring quantum advantage but are limited by decoherence, gate errors, and connectivity constraints. As a result, most practical QML and QNLP approaches rely on hybrid quantum–classical methods. **Variational Circuits** is a central paradigm in the NISQ era is the use of variational quantum circuits (VQCs) as shown in Figure 3. These are parameterized circuits $U(\theta)$ with tunable gates, where parameters θ are optimized iteratively by a classical optimizer. Given an input state $|\psi_0\rangle$ and a Hamiltonian H encoding the objective, the optimization task is defined as:

$$C(\theta) = \langle \psi_0 | U^\dagger(\theta) H U(\theta) | \psi_0 \rangle.$$

The quantum device computes expectation values, while the classical optimizer updates θ using gradient-based or gradient-free methods (Jäger et al., 2025). This feedback loop exploits quantum representational capacity while avoiding long quantum coherence times, which are difficult to sustain on NISQ devices. In a typical hybrid learning pipeline, classical pre-processing transforms raw data into a form suitable for quantum encoding (e.g., token embeddings or feature normalization). The encoded data are passed to a quantum circuit that performs transformations, such as entangling operations to capture correlations. The measurement outcomes are then post-processed by classical neural layers or decision functions. This integration allows quantum circuits to act as specialized layers within a larger classical deep learning framework.

For natural language tasks, hybrid models provide a practical compromise between expressiveness and feasibility. Classical components handle tasks such as subword tokenization, syntactic parsing, or initial embedding generation, while the quantum layer captures higher-order dependencies using entanglement. For example, a hybrid QNLP pipeline might map token embeddings into quantum states, apply a variational circuit to model contextual interactions, and then use a classical classifier to predict sentiment or semantic similarity. Such approaches combine the scalability of classical preprocessing with the structural advantages of quantum computation.

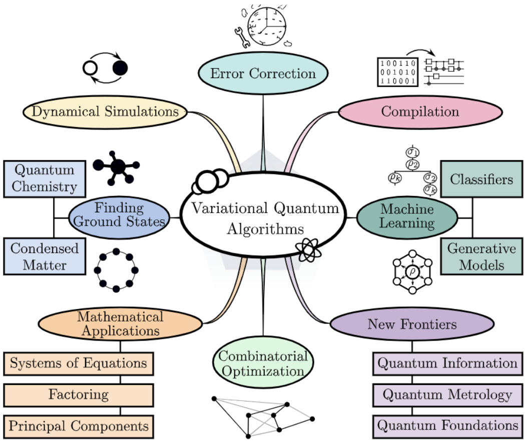

Figure 3: Applications of Variational Quantum Algorithms (VQAs) across optimization, simulation, machine learning, and emerging quantum domains. (Cerezo et al., 2021a).

3 Computational Models for QNLP

Several computational paradigms have been proposed for QNLP, each exploiting different aspects of quantum mechanics to model linguistic structure, meaning, and tasks. This section surveys categorical compositional frameworks, circuit based models, variational approaches, quantum kernel methods, and quantum inspired NLP techniques.

3.1 Categorical Compositional Models

The *categorical compositional distributional model* (DisCoCat) (Wu and Wang, 2019) was one of the first frameworks to unify grammatical structure and distributional semantics in a quantum-compatible setting. It leverages *compact closed categories* to map syntactic derivations to tensor contractions in Hilbert spaces. Each word is represented as a tensor, and sentence meaning arises compositionally through linear maps:

$$\vec{s} = f(\vec{w}_A \otimes \vec{w}_B), \quad f : A \otimes B \to C,$$

with entanglement naturally encoding word dependencies.

Building on this foundation, several extensions have been proposed: DisCoCirc (Chang et al., 2023): introduces discourse-level dynamics by updating word states via variational quantum circuits, e.g., $|w\rangle' = U_c\,|w\rangle$. Quantum Graph Transformers (QGT) (Xu et al., 2025): integrate dependency graphs with quantum self-attention, where attention weights are computed by parameterized circuits:

$$\alpha_{ij} = \frac{\exp(\langle \phi(x_i)|U(\theta)|\phi(x_j)\rangle)}{\sum_k \exp(\langle \phi(x_i)|U(\theta)|\phi(x_k)\rangle)}.$$

Quantum Context-Sensitive Embeddings (QCSE) (Liu et al., 2025b): generalize contextual embeddings (e.g., BERT) into Hilbert space with $|w,c\rangle = U(C)\,|w\rangle$. Quantum Text Pretraining Networks (QTP-Net) (Zhang et al., 2025): encode word senses as quantum superpositions $|w\rangle = \sum_i \alpha_i\,|s_i\rangle$ aligned with knowledge bases. MultiQ-NLP (Wang et al., 2024): extends composition to multimodal data, using entanglement to model cross-modal dependencies (text–image).

Together, these models have evolved DisCoCat from a purely categorical semantic formalism into dynamic, contextual, pretrained, and multimodal frameworks, demonstrating the adaptability of QNLP across linguistic and hybrid tasks.

3.2 Quantum Circuit-based Models

Quantum circuits map linguistic structure directly into hardware-executable operations. Tokens are encoded into quantum states, syntactic relations are represented by entangling gates, and grammatical reductions correspond to circuit modules (Ge et al., 2024). For example, a dependency relation between two words may be represented as a controlled rotation or CNOT gate applied between their corresponding qubits (Hu and Kais). Sentence meaning then emerges from the full circuit state, with measurements providing semantic outputs (Lan et al., 2024). An example of such a circuit implementation for a simple transitive sentence is shown in Figure 4.

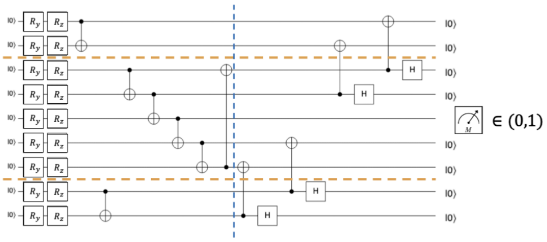

Figure 4: Quantum circuit for a transitive sentence. The circuit based on DisCoCat model, maps a simple sentence into quantum operations. Qubits on the left encode word embeddings via rotation gates, while the right region represents grammatical contractions through entangle gates such as CNOT

Circuit-based approaches highlight the structural parallel between parse trees and quantum circuit diagrams, making them natural candidates for syntax-sensitive tasks (Liu et al., 2025a).They are particularly attractive for experiments on NISQ devices since circuits can be compiled directly into gate sequences supported by current hardware (Venturelli et al., 2019). However, their scalability depends on efficient encoding schemes and noise-aware compilation, as circuit depth grows with sentence length. Hybrid pipelines that combine shallow circuits with classical post-processing are commonly used to mitigate hardware limitations. A recent circuit-based approach proposes Quantum Parameter Adaptation (QPA), where quantum neural networks are used during training to generate classical model weights. This enables parameter efficient fine tuning of LLMs while keeping inference entirely classical (Liu et al., 2025a).

3.3 Variational Quantum Models

Variational quantum circuits (VQCs) $U(\theta)$ extend circuit-based models by introducing tunable parameters θ optimized via classical loops (Liu et al., 2024).

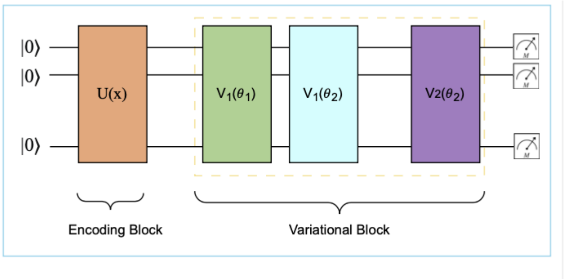

Figure 5: Variational Quantum Circuit (VQC) architecture illustrating how linguistic inputs are encoded into quantum states and processed by parameterized variational layers whose parameters are trained in a classical optimization loop. (Liu et al., 2024).

This paradigm makes VQCs the most widely explored approach in QNLP. Tokens are embedded into quantum states via feature maps, processed through parameterized entangling layers, and measured to produce outputs (Kankeu et al., 2025).

Training minimizes a loss function:

$$C(\theta) = \sum_i \ell(y_i, f_\theta(x_i)),$$

where ℓ is typically cross-entropy or mean squared error.

VQCs have been applied to tasks such as text classification, semantic similarity, and sentiment analysis. They benefit from the expressive capacity of entanglement to capture contextual information, and from their hybrid nature which integrates well with classical neural networks. Key challenges include barren plateaus in optimization, noise-induced instability, and the high cost of quantum state preparation (Novák et al., 2025). Recent work explores hardware efficient ansätze and error aware training to address these limitations (Gujju et al., 2025), making VQCs a practical testbed for QNLP research.

3.4 Quantum Kernel Methods

Quantum kernel methods leverage quantum feature maps $|\phi(x)\rangle$ that embed linguistic data into Hilbert spaces of potentially exponential dimension. The induced kernel is defined as:

$$k(x, x') = |\langle \phi(x) | \phi(x') \rangle|^2,$$

which can be used with classical machine learning models such as support vector machines (SVMs) or Gaussian processes (Wang et al., 2025). These methods are particularly well-suited to similarity-based tasks, including semantic textual similarity (STS), paraphrase detection, and clustering of embeddings (Herbold, 2024). They offer the advantage of being mathematically rigorous, providing provable separability properties in high-dimensional spaces. However, scalability is a major limitation, since evaluating kernels requires repeated state preparation and measurement. Approximate quantum kernel estimation and hybrid quantum–classical kernel learning have been proposed as intermediate solutions.

4 Encoding Paradigms

4.1 Basic Encoding

A recent proposal introduces a *learnable basic encoding layer* that maps each token to a qubit register with minimal parameter overhead (Munikote, 2024). Instead of relying purely on fixed rotation or amplitude maps, the method applies small parameterized gates on basis states, adapting them during training to better reflect token distributions. Concretely, a token index i is first mapped to a basis state $|i\rangle$, and then acted upon by a shallow trainable unitary $E(\phi)$:

$$|\psi_i\rangle = E(\phi) \, |i\rangle \, .$$

Here, $E(\phi)$ is composed of single-qubit rotations and entanglers whose parameters ϕ are learned jointly with the downstream task, offering a flexible compromise between rigid encodings and heavy variational circuits.

The scheme retains the discrete structure of token identities while allowing adaptation to semantic space, enabling gradients to flow directly through the encoder (Baek et al., 2025). Because only a small unitary is applied, the circuit depth overhead remains modest, making it compatible with NISQ devices. Its parameters can absorb differences in token frequency or contextual distributions, positioning this method between static basis encoding and hybrid embeddings. As such, it provides a more expressive and scalable representation for QNLP tasks than one-hot or rotation-only mappings.

4.2 Amplitude Encoding

Embed dense vectors into amplitudes:

$$\mathbf{e} \in \mathbb{R}^d \;\mapsto\; |\phi(\mathbf{e})\rangle = \frac{1}{\|\mathbf{e}\|} \sum_{j=1}^{d} e_j \, |j\rangle \, .$$

This method is highly qubit-efficient ($\log d$) and preserves inner-product geometry, allowing similarity to be computed via inner products in Hilbert space. The main drawback is that state preparation can be computationally expensive, often requiring $\mathcal{O}(d)$ operations, and the resulting states are sensitive to noise. To mitigate this, amplitude encoding is often combined with problem-specific *quantum feature maps*, enabling kernel methods that exploit the high-dimensional Hilbert space structure (Schuld and Killoran, 2019).

Recent advances show that amplitude encoding can deliver exponential data compression in hybrid quantum-classical architectures. For instance, a dataset with $d = 2^n$ features can be represented using only n qubits, whereas angle encoding would require d. Chen et al. (Chen et al., 2025) integrate amplitude encoding into hybrid Quantum Neural Networks (QNNs) for recovery rate prediction, demonstrating superior generalization on small-sample, high-dimensional financial datasets. Embedding amplitude-encoded inputs into Parameterized Quantum Circuits (PQCs) preserves unitarity and avoids costly orthogonality constraints, yielding two key benefits: improved computational efficiency through fewer qubits and parameters, and richer representational capacity compared to angle encoding for tasks requiring high-dimensional embeddings.

4.3 Entanglement-based Encodings

Introduce entanglers (CNOT/CZ) to correlate token subsystems (Schuld et al., 2021):

$$|\psi\rangle = U_{\text{ent}}\big(\,|w_1\rangle\otimes\cdots\otimes|w_n\rangle\,\big).$$

This approach explicitly captures syntactic and semantic dependencies by creating correlations between token representations, mirroring categorical contraction in compositional semantics. Entanglement allows local word embeddings to be combined into global sentence states, enriching expressivity beyond independent encodings.

The trade-off is that entanglement substantially increases circuit depth and noise sensitivity, especially on NISQ hardware (González-García et al., 2022). Efficient design therefore requires carefully chosen ansätze and compilation strategies to minimize gate counts and error accumulation. When optimized, entanglement-based encodings provide a direct mechanism for modeling relational structure, but scalability remains a major challenge compared to simpler schemes.

4.4 Hybrid Embedding Strategies

A hybrid approach first uses a classical model (e.g., BERT or Word2Vec) (Devlin et al., 2019) to compute an embedding $\mathbf{e}$, and then applies a feature map $\mathbf{e} \mapsto |\phi(\mathbf{e})\rangle$ followed by a trainable quantum circuit $U(\theta)$ before measurement. This combines the semantic richness of pretrained embeddings with quantum layers that can model higher-order correlations and capture non-linear dependencies in Hilbert space (Döschl and Bohrdt, 2025).

Such strategies represent the most practical and NISQ-friendly pathway, since heavy semantic lifting is done classically and quantum resources are reserved for expressive refinements. By leveraging classical pretraining, hybrid embeddings reduce qubit demands and training cost, while still offering the potential to uncover representational structures inaccessible to purely classical methods. This makes them a dominant design choice for early QNLP systems and applied quantum machine learning pipelines.

4.5 Space-efficient tensorized embeddings.

A line of work factorizes the embedding matrix into low-order tensor products inspired by entanglement, yielding *word2ket*-style embeddings that compress parameters by $10^2\times$ or more with negligible accuracy loss on standard NLP tasks (Panahi et al., 2019). These embeddings can be used purely classically or as quantum-ready parametrizations (tensor factors ⇒ shallow preparation circuits). This offers a principled bridge between tensor-network structure and learnable word representations.

4.6 Trainable quantum embedding circuits.

A 2024 study proposes a recurrent quantum embedding neural network (RQENN) with a *trainable encoding* based on parameterized binary indices that learns token embeddings *within* a small quantum circuit cell; the cell is reused across sequence positions to capture context with fewer qubits and measurements than prior QNLP approaches (Varmantchaonala et al., 2025). Reported results show reduced parameter count and bits used, and accuracy gains over earlier QNLP baselines on a text-like vulnerability detection task, highlighting the value of *learned* encoders (vs. fixed maps) under NISQ constraints (Kea et al., 2024).

4.7 Resource Cost Modeling

We characterize encodings by qubits q, depth L, state-prep cost T_{prep}, and shot complexity m. For amplitude encoding,

$$\mathbf{e} \in \mathbb{R}^d \mapsto |\phi(\mathbf{e})\rangle = \frac{1}{\|\mathbf{e}\|}\sum_{j=1}^{d} e_j\,|j\rangle\,, \qquad (1)$$

$$q = \lceil \log_2 d \rceil, \qquad T_{\text{prep}} = \Theta(d). \qquad (2)$$

with low depth but prep-bound runtime. Angle/rotation encoding yields $q = \Theta(d)$, $T_{\text{prep}} = \Theta(d)$ and often better robustness on NISQ. Entanglement-based composition adds syntax or graph-induced two qubit layers; we report $L = L_0 + E$ where E is the number of entangling edges (Susulovska, 2024). For hybrid embeddings, q is constant (few qubit head) with classical compute absorbing semantics; we report wall clock and device usage alongside accuracy.

5 Evaluation Frameworks

Evaluation in QNLP spans both empirical performance and theoretical efficiency. At the task level, models are assessed on standard NLP objectives such as sentiment classification, semantic similarity, and sequence labeling, with accuracy, F1, or correlation metrics compared against compute-matched classical baselines (Tomal et al., 2025). Because quantum circuits produce probabilistic outputs, metrics are accompanied by confidence intervals derived from measurement shots, and evaluations must also report resource costs including qubit counts q, circuit depth L, gate complexity, state-preparation cost T_{prep}, and shot budgets m, ensuring fairness under NISQ constraints (Ma and Li, 2024). To validate results beyond simulation, a hardware-in-the-loop protocol is followed: device backend, transpilation strategy, calibration snapshot, and shot counts are disclosed, with paired simulator–device runs performed using identical seeds (Nguyen et al., 2017). Robustness is further probed through noise modeling, barren-plateau stress tests, and lightweight error mitigation (readout calibration, zero-noise extrapolation, and gradient-preserving initialization).

Beyond raw task performance, evaluation emphasizes comparability and reproducibility. Canonical ablations such as removing entanglers, swapping amplitude versus angle encodings, reducing data re-uploading depth, or replacing quantum heads with classical ones are standardized to attribute improvements to specific design choices (Aktar et al., 2025). Benchmarking remains challenging due to the lack of large standardized QNLP corpora, so we propose compact, structure-sensitive tasks (compositional classification, semantic similarity, and sequence labeling with nested constituents) with fixed splits and optional precomputed embeddings for hybrid models. Together with artifact release (QASM circuits, seeds, calibration snapshots, and ablation configs) (Li et al., 2022), these practices enable like for like comparisons across models and clarify where QNLP shows unique strengths capturing compositionality, contextual dependencies, and high-dimensional correlations while highlighting the tradeoffs in scalability, noise resilience, and hardware feasibility relative to classical NLP systems (Lhoest et al., 2021).

6 Challenges and Future Directions

Despite encouraging theoretical advances and early experiments, QNLP still faces significant challenges. Current NISQ hardware limits circuit depth, qubit counts, and gate fidelity, restricting scalability and necessitating noise-resilient encodings and carefully designed variational ansätze (Preskill, 2018; McClean et al., 2018). Encoding strategies such as amplitude or entanglement-based mappings offer expressive representational power but suffer from costly state preparation and noise sensitivity, motivating the exploration of adaptive encodings and resource efficient parameterization methods that balance expressivity with hardware feasibility (Chen et al., 2025). At the evaluation level, the absence of standardized QNLP benchmarks makes comparisons across models difficult; task-specific corpora and quantum-compatible evaluation suites are needed to validate theoretical speedups and measure robustness under realistic conditions (Lorenz et al., 2021a).

Looking ahead, hybrid quantum classical pipelines remain the most practical path, though their advantage over strong classical baselines such as transformers is not yet conclusive. Further research into quantum inspired embeddings and hybrid variational architectures may clarify where QNLP offers unique value (Huang et al., 2021; Kartsaklis et al., 2021). Achieving scalability will require moving beyond toy corpora to industrial-scale applications such as semantic search, question answering, and multimodal reasoning. Meeting these goals will demand not only algorithmic innovation but also advances in quantum hardware and close collaboration between NLP researchers and quantum computing specialists, ensuring QNLP matures into a robust framework for structure-sensitive language tasks.

7 Conclusion

QNLP lies at the intersection of quantum computing and natural language processing, introducing new paradigms for compositional semantics, efficient representation, and contextual modeling. This survey reviews foundational models DisCoCat, circuit based, variational, and hybrid architectures alongside encoding strategies, evaluation frameworks, and open challenges. Although still nascent, advances in hybrid embeddings, quantum feature maps, and noise mitigation indicate near-term feasibility. Future progress will hinge on scalable benchmarks, tighter integration with classical NLP, and improved quantum hardware. QNLP thus holds promise to advance beyond proof-of-concept studies and deliver tangible computational gains for structure sensitive language tasks.

References

Shamminuj Aktar, Andreas Bärtschi, Abdel-Hameed A. Badawy, and Stephan Eidenbenz. 2025. Quantum graph transformer for nlp sentiment classification.

Junyeob Baek, Hosung Lee, Christopher Hoang, Mengye Ren, and Sungjin Ahn. 2025. Discrete jepa: Learning discrete token representations without reconstruction.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33.

M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. 2021a. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644.

Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, et al. 2021b. Variational quantum algorithms. *Nature Reviews Physics*, 3.

Eric Chang, John Smith, and Ling Zhao. 2023. Variational quantum classifiers for natural-language text: A discocirc approach. In *Proceedings of the Quantum NLP Workshop*.

Ying Chen, Paul Griffin, Paolo Recchia, Lei Zhou, and Hongrui Zhang. 2025. Hybrid quantum neural networks with amplitude encoding: Advancing recovery rate predictions.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *Lambek Festschrift*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*.

Fabian Döschl and Annabelle Bohrdt. 2025. Importance of correlations for neural quantum states.

Yan Ge, Wu Wenjie, Chen Yuheng, Pan Kaisen, Lu Xudong, Zhou Zixiang, Wang Yuhan, Wang Ruocheng, and Yan Junchi. 2024. Quantum circuit synthesis and compilation optimization: Overview and prospects.

Guillermo González-García, Rahul Trivedi, and J. Ignacio Cirac. 2022. Error propagation in nisq devices for solving classical optimization problems. *PRX Quantum*, 3(04):040326. "NISQ-Computer: Quantum entanglement can be a double-edged sword" via MPQ news.

Yaswitha Gujju, Romain Harang, and Tetsuo Shibuya. 2025. Llm-guided ansätze design for quantum circuit born machines in financial generative modeling.

Aram W Harrow, Avinatan Hassidim, and Seth Lloyd. 2009. Quantum algorithm for linear systems of equations. *Physical Review Letters*, 103(15):150502.

Steffen Herbold. 2024. Semantic similarity prediction is better than other semantic similarity measures.

Zheyong Hu and Sabre Kais. Characterizing quantum circuits with qubit functional configurations. *Scientific Reports*.

Hsin-Yuan Huang, Richard Kueng, and John Preskill. 2021. Power of data in quantum machine learning. *Nature Communications*, 12(1):2631.

Jonas Jäger, Thierry Nicolas Kaldenbach, Max Haas, and Erik Schultheis. 2025. Fast gradient-free optimization of excitations in variational quantum eigensolvers.

Ivan Kankeu, Stefan Gerd Fritsch, Gunnar Schönhoff, Elie Mounzer, Paul Lukowicz, and Maximilian Kiefer-Emmanouilidis. 2025. Quantum-inspired embeddings projection and similarity metrics for representation learning.

D. Kartsaklis, R. Lorenz, K. Meichanetzidis, and B. Coecke. 2021. lambeq: An efficient high-level python library for quantum nlp. In *Proceedings of the 16th Conference on Quantum Physics and Logic (QPL)*.

Kimleang Kea, Won-Du Chang, Hee Chul Park, and Youngsun Han. 2024. Enhancing a convolutional autoencoder with a quantum approximate optimization algorithm for image noise reduction.

Michael Lan, Philip Torr, and Fazl Barez. 2024. Towards interpretable sequence continuation: Analyzing shared circuits in large language models.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing.

Ang Li, Samuel Stein, Sriram Krishnamoorthy, and James Ang. 2022. Qasmbench: A low-level quantum benchmark suite for nisq evaluation and simulation. *ACM Transactions on Quantum Computing*.

Chen-Yu Liu, Samuel Yen-Chi Chen, Kuan-Cheng Chen, Wei-Jia Huang, and Yen-Jui Chang. 2024. Programming variational quantum circuits with quantum-train agent.

Chen-Yu Liu, Chao-Han Huck Yang, Hsi-Sheng Goan, and Min-Hsiu Hsieh. 2025a. A quantum circuit-based compression perspective for parameter-efficient learning. In *The Thirteenth International Conference on Learning Representations*.

Xiaoming Liu, Yifan Zhao, and Ming Chen. 2025b. Qcse: Pretrained quantum context-sensitive word embeddings. *arXiv preprint arXiv:2509.05729*.

R. Lorenz, D. Kartsaklis, K. Meichanetzidis, and B. Coecke. 2021a. Qnlp experiments with lambeq. In *Proceedings of Quantum Physics and Logic (QPL)*.

Robin Lorenz, Daniel Pearson, and et al. 2021b. Qnlp: Quantum natural language processing. *arXiv preprint arXiv:2102.12846*.

Ning Ma and Heng Li. 2024. Understanding and estimating the execution time of quantum programs.

Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. 2018. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9(1):4812.

Konstantinos Meichanetzidis, Abdelghani Toumi, Giacomo De Felice, Bob Coecke, et al. 2020. Quantum natural language processing on near-term quantum computers. *arXiv preprint arXiv:2005.04147*.

Valter Moretti and Marco Oppio. 2017. Quantum theory in real hilbert space: How the complex hilbert space structure emerges from poincaré symmetry. *Reviews in Mathematical Physics*, 29(06):1750021.

Nidhi Munikote. 2024. Comparing quantum encoding techniques.

Van Hoa Nguyen, Yvon Besanger, Quoc Tuan Tran, Tung Lam Nguyen, Cederic Boudinet, Ron Brandl, Frank Marten, Achilleas Markou, Panos Kotsampopoulos, Arjen A. van der Meer, Effren Guillo-Sansano, Georg Lauss, Thomas I. Strasser, and Kai Heussen. 2017. Real-time simulation and hardware-in-the-loop approaches for integrating renewable energy sources into smart grids: Challenges actions.

Vojtěch Novák, Ivan Zelinka, and Václav Snášel. 2025. Optimization strategies for variational quantum algorithms in noisy landscapes.

Aliakbar Panahi, Seyran Saeedi, and Tom Arodz. 2019. word2ket: Space-efficient word embeddings inspired by quantum entanglement. In *International Conference on Learning Representations 2020*.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21.

Maria Schuld and Nathan Killoran. 2019. Quantum machine learning in feature hilbert spaces. *Physical Review Letters*, 122(4):040504.

Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. 2015. An introduction to quantum machine learning. *Contemporary Physics*, 56(2):172–185.

Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. 2021. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3).

N. A. Susulovska. 2024. Geometric measure of entanglement of quantum graph states prepared with controlled phase shift operators.

S. M. Yousuf Iqbal Tomal, Abdullah Al Shafin, Debojit Bhattacharjee, MD. Khairul Amin, and Rafiad Sadat Shahir. 2025. Quantum-enhanced attention mechanism in nlp: A hybrid classical-quantum approach.

Charles M. Varmantchaonala, Jean Louis K. E. Fendji, Julius Schöning, and Marcellin Atemkeng. 2024. Quantum natural language processing: A comprehensive survey. *IEEE Access*, 12:99578–99598.

Charles M. Varmantchaonala, Niclas GÖtting, Nils-Erik SchÜtte, Jean Louis E. K. Fendji, and Christopher Gies. 2025. Qcse: A pretrained quantum context-sensitive word embedding for natural language processing.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Davide Venturelli, Minh Do, Bryan O'Gorman, Jeremy Frank, Eleanor Rieffel, Kyle E. C. Booth, Thanh Nguyen, Parvathi Narayan, and Sasha Nanda. 2019. Quantum circuit compilation: An emerging application for automated reasoning. In *Scheduling and Planning Applications Workshop*.

Leyang Wang, Yilun Gong, and Zongrui Pei. 2025. Quantum and hybrid machine-learning models for materials-science tasks.

Yichen Wang, Zhe Sun, and Xin Hu. 2024. Multiqnlp: Multimodal structure-aware quantum natural language processing. *arXiv preprint arXiv:2411.04242*.

Nathan Wiebe, Bob Coecke, et al. 2024. Near-term advances in quantum natural language processing. *Annals of Mathematics and Artificial Intelligence*, 92(3-4):501–520.

Yanying Wu and Quanlong Wang. 2019. A categorical compositional distributional modelling for the language of life.

Jian Xu, Feng Li, and Hao Chen. 2025. Quantum graph transformers for sentiment classification. *arXiv preprint arXiv:2506.07937*.

Hao Zhang, Jiaqi Wang, and Qiang Li. 2025. Qtpnet: Quantum text pre-training network for nlp tasks. *arXiv preprint arXiv:2506.00321*.

Li Zhang, Wei Chen, et al. 2024. Quantum algorithms for compositional text processing. *arXiv preprint arXiv:2408.06061*.

A Appendix

Task	Method	Design Highlights	Input Data Type	Label Type	Loss
Sentence Classification	DisCoCat (Coecke et al., 2010)	Maps grammatical reductions to tensor contractions in Hilbert space (compact-closed categories); sentence meaning via categorical compositionality with quantum-ready tensors.	Tokenized sentences	Sentiment / Topic	Cross-entropy
	VQC-QNLP (Gujju et al., 2025)	Parameterized quantum circuit $U(\theta)$ on encoded tokens; hybrid loop minimizes expectation; entanglement captures long-range dependencies under NISQ.	Token embeddings	Binary / Multi-class	Weighted cross-entropy
Semantic Similarity	QBW (Lorenz et al., 2021a)	Quantum Bag-of-Words; embeds words as quantum states; measures similarity via state fidelity/overlaps instead of cosine distance.	Sentence pairs	Similarity / Paraphrase	Fidelity or MSE
	Quantum Kernel (QK-NLP) (Schuld and Killoran, 2019; Wang et al., 2025)	Quantum feature map $\lvert\phi(x)\rangle$ induces kernel $k(x,x') = \lvert\langle\phi(x)\vert\phi(x')\rangle\rvert^2$; classical SVM/GP on quantum kernel matrix.	Sentences / embeddings	STS / Entailment	Hinge loss / GP NLL
Sequence Labeling	DisCoCirc (Chang et al., 2023)	Discourse-aware extension of DisCoCat; circuit evolution updates word states across context; syntax–semantics via variational updates.	Token sequences	POS / NER / chunks	Token-level cross-entropy
	QCSE (Liu et al., 2025b)	Quantum Context-Sensitive Embeddings: context unitary $U(C)\,\lvert w\rangle$ entangles tokens; contextual vectors in Hilbert space for tagging.	Token sequences	Sequence tags	MSE / cross-entropy
Hybrid Embedding Learning	Hybrid-QNN (Chen et al., 2025)	Classical encoder (e.g., BERT) → amplitude/angle map → shallow PQC refinement; few-qubit head for NISQ robustness.	Pretrained text embeddings	Sentiment / Intent	Cross-entropy (hybrid)
Low-Resource / Multi-Modal	MultiQ-NLP (Wang et al., 2024)	Entangles text–image qubits; cross-modal attention via controlled rotations; improves transfer in few-shot regimes.	Text–image pairs	Match / Tags	Contrastive (InfoNCE)
Sense Modeling / Pretraining	QTP-Net (Zhang et al., 2025)	Encodes word senses as quantum superpositions $\lvert w\rangle = \sum_i \alpha_i \lvert s_i\rangle$; learns sense mixture via measurement-driven objectives.	Large text corpora	Sense / Masked tokens	NLL; superposition reconstruction
Encoding Learning	Trainable Basic Encoding (Munikote, 2024)	Learnable encoder $E(\phi)$ on basis states prior to PQC; low-depth, NISQ-friendly alternative to fixed angle/amplitude maps.	Token indices	Task-specific	Task loss + encoder reg.
Resource-Efficient Embeddings	word2ket / Tensorized (Panahi et al., 2019)	Factorizes embedding matrix into low-order tensor products; quantum-ready prep with shallow circuits; large parameter compression.	Vocabulary embeddings	Task-specific	Task loss; tensor-factor regs

Table 1: **Summary of representative Quantum Natural Language Processing (QNLP) models across core linguistic tasks.** The table aligns prior work by task, model type, and architectural design to illustrate how quantum principles are applied to language understanding. **Task** denotes the linguistic objective (e.g., classification, similarity, or tagging); **Method** names the quantum or hybrid framework; **Design Highlights** summarize each model's encoding scheme (amplitude, angle, entanglement, or hybrid), circuit structure, and optimization strategy. **Input** and **Label Type** describe the data and prediction targets, while **Loss / Objective** lists the corresponding training criterion. Together, these entries show how QNLP architectures integrate formal semantics with quantum computation, balancing expressivity, resource efficiency, and NISQ-era feasibility.

Encoding Paradigm	Core Idea / Map	Qubits q	State-Prep Cost T_{prep}	Strengths	Limitations
Basic / Learnable Encoding	Token index $i \mapsto \lvert i\rangle$ with shallow trainable unitary $E(\phi)\,\lvert i\rangle$	$\Theta(\log V)$ (index map)	Low (shallow $E(\phi)$)	Very low depth; parameter-efficient; preserves discrete identity; NISQ-friendly	Needs downstream entanglers/PQC for expressivity; tuning still task-dependent
Angle / Rotation Encoding	Map features to single-qubit rotations (e.g., $R_y(\cdot)/R_z(\cdot)$) per dimension; supports data re-uploading	$\Theta(d)$	$\Theta(d)$	Simple, robust, transparent geometry; pairs well with re-uploading in VQCs	Linear qubit growth with d; underuses Hilbert space unless combined with entanglement
Amplitude Encoding	$\mathbf{e}\in\mathbb{R}^d \mapsto \lvert\phi(\mathbf{e})\rangle = \frac{1}{\lVert\mathbf{e}\rVert}\sum_j e_j \lvert j\rangle$ (inner-products preserved)	$\lceil \log_2 d \rceil$	$\Theta(d)$ (state loading)	Exponential compression of d; strong for kernel/similarity tasks; unitary-friendly	Expensive loaders; noise-sensitive; benefits from high-fidelity prep
Entanglement-based Composition	Apply U_{ent} (CNOT/CZ) to correlate token subsystems; syntax/relations via entanglers	Task-dependent	Entanglers dominate	Directly captures compositional/relational structure; aligns with categorical semantics	Increases depth and error on NISQ; careful compilation needed
Hybrid Embedding Strategies	Classical embedding $\mathbf{e}$ (e.g., BERT/Word2Vec) → quantum feature map $\lvert\phi(\mathbf{e})\rangle$ → PQC $U(\theta)$	Few-qubit heads common	Modest; depends on chosen feature map	Best near-term trade-off; leverages pretrained semantics; smaller q / shots	Classical front-end may dominate compute; quantum benefit is task- and map-dependent
Space-efficient Tensorized (word2ket)	Factorize embedding matrix into low-order tensor products; shallow quantum prep from factors	By factorization design	Low (from tensor factors)	$10^2\times$ compression reported; principled bridge to tensor networks; shallow circuits	Quality depends on factorization rank/structure; extra design choices required
Trainable Quantum Embedding Circuits	Small reusable quantum cell learns token/context encoding in-circuit; reused across positions	Few (cell reused)	Low–moderate (per-cell)	Parameter-efficient; context-aware; fewer qubits/shots than naïve per-token circuits	Requires careful training/stability on NISQ; generalization may be dataset-dependent

Table 2: Encoding paradigms discussed in this survey. V: vocabulary size; d: feature dimension. For fair NISQ comparisons, report q, circuit depth L, state-prep cost T_{prep}, and shot budgets m alongside task metrics.