Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Quantum Destructive Self-Attention for NISQ-EraTransformers

Yousra BOUAKBA

yousra.bouakba@univ-constantine2.dz

LISIA Laboratory

Hacene Belhadef

Université Constantine 2

Research Article

Keywords: Quantum Machine Learning, Quantum Enhanced-Transformer, Quantum Word Similarity, Text Classification

Posted Date: February 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8946631/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Quantum Destructive Self-Attention for NISQ-Era Transformers

Yousra Bouakba[1,2*] and Hacene Belhadef [2]

[1*]LISIA Laboratory , NTIC Faculty, Constantine, 25016, Algeria.
[2*]NTIC Faculty , University of Constantine 2 - Abdelhamid Mehri, Constantine, 25016, Algeria.

*Corresponding author(s). E-mail(s): yousra.bouakba@univ-constantine2.dz;
Contributing authors: hacene.belhadef@univ-constantine2.dz;

Abstract

In the noisy intermediate-scale quantum (NISQ) era, developing attention mechanisms suitable for quantum hardware is essential for advancing quantum-enhanced Transformers applications. This work introduces a novel quantum Transformer model that integrates a low-depth, ancilla-free attention mechanism based on the Destructive swap-test variant, enabling efficient computation of attention scores without auxiliary qubits. Designed specifically for NISQ constraints, the model emphasizes noise resilience and reduced circuit depth, making it practical for current quantum devices. We evaluate the model using transparent and well-documented Natural Language Processing datasets designed for Quantum Computing (QC). Our results demonstrate competitive performance on language tasks, supported by both quantitative metrics that offer interpretable insights into the model's behavior. This work provides a viable pathway toward practical, interpretable quantum language models in the near term.

Keywords: Quantum Machine Learning, Quantum Enhanced-Transformer, Quantum Word Similarity, Text Classification

1 Introduction

Transformer architectures have revolutionized artificial intelligence, achieving state-of-the-art performance in natural language processing, vision, and multi-modal tasks.

Their success stems from the self-attention mechanism [1], which dynamically models token interactions across sequences without recurrence. However, this power comes at a cost: computing all pairwise interactions results in quadratic time and memory complexity, limiting scalability to longer sequences and increasing energy demands. As datasets and model sizes continue to expand, these bottlenecks raise critical concerns regarding the efficiency and sustainability of Transformer-based learning [2].

To address these challenges, research has increasingly turned toward quantum-enhanced Transformers, aiming to exploit the high-dimensional representational capacity of quantum states for more efficient sequence modeling. Quantum architectures promise compact data embeddings and novel mechanisms for similarity computation that could mitigate the quadratic scaling of classical attention . According to the recent survey by Zhang and Zhao [3], these approaches can be broadly divided into two paradigms: quantum linear algebra (QLA)-based methods, which exploit block-encoding and singular value transformations, and parameterized quantum circuit (PQC)-based methods, which embed data and learnable parameters into variational quantum circuits. This classification highlights the trade-off between the theoretical scalability of QLA-based models and the near-term feasibility of PQC-based designs under NISQ constraints.

While QLA-based models [4]-[5] offer theoretical exponential speedups, they typically demand fault-tolerant hardware. Consequently, PQC-based designs have become the dominant line of research for the NISQ era. Despite their promise, PQC-based quantum Transformers remain hampered by computational inefficiencies. Most existing designs [6]-[7] continue to inherit the $O(N^2)$ complexity of classical attention, as quantum similarity tests such as the swap test or Hadamard test must be repeated for every token pair. Moreover, many of these circuits require ancilla qubits and deep entangling operations, which increase resource demands and degrade performance under noise. These challenges underscore the need for lightweight, ancilla-free, and depth-efficient quantum attention mechanisms that align with NISQ-era hardware constraints while maintaining expressive power. In this work, we address this gap by proposing an ancilla-free, low-depth quantum similarity estimation circuit, originally designed for state overlap estimation, as a novel and efficient method for PQC-based attention computation.

In this work, we make the following contributions to the development of quantum Transformer architectures:

1. **Purpose of a destructive swap-test variant:** We introduce the novel application of an existing ancilla-free, low-depth swap-test variant to the computation of attention scores in quantum Transformers. While this circuit has been studied in other contexts, we demonstrate for the first time its suitability as a practical, noise-resilient alternative to similarity estimation methods in attention mechanisms.
2. **Design of a new encoder-only Transformer block:** Building on this primitive, we develop a hybrid encoder-only Transformer architecture tailored for NISQ devices. The model leverages the destructive swap-test variant for efficient attention while maintaining compatibility with shallow, hardware-executable PQCs.

3. **Fair and reproducible evaluation:** To address inconsistencies in prior work caused by heavy or undocumented preprocessing on popular datasets (e.g., IMDb, Yelp), we propose evaluation on alternative NLP datasets designed for quantum computation. This enables reproducibility and fair benchmarking under NISQ constraints.
4. **Demonstration of efficiency and robustness:** Through theoretical analysis and simulation, we show that the proposed mechanism achieves advantages in circuit depth and noise resilience compared to conventional PQC-based approaches, underscoring its potential for near-term quantum deployment.

The remainder of this paper is organized as follows. Section 2 reviews preliminaries on PQCs and the Transformer architecture, followed by Section 3, which surveys related work and categorizes existing approaches into QLA-based and PQC-based paradigms. Section 4 provides a theoretical analysis of existing pair-wise similarity-based attention mechanisms, highlighting their computational and practical limitations. In Section 5, we introduce the destructive swap test for word similarity, describe its ancilla-free construction, and present a noise analysis. Section 6 details our methodology, including the design of a hybrid encoder-only Transformer and the integration of the destructive attention module. Section 7 outlines the experimental setup, datasets, and evaluation protocols, while Section 8 presents results and discussion, including comparisons with prior quantum NLP models. Finally, Section 9 concludes the paper with a summary of contributions and directions for future work.

2 Preliminaries

2.1 Classical Transformer Architecture and the Self-Attention Mechanism

The Transformer has become a cornerstone model in modern machine learning due to its ability to model long-range dependencies across sequences without recurrence or convolution. Its effectiveness arises from the self-attention mechanism, which computes pairwise token interactions in parallel, making it highly expressive for tasks such as natural language understanding and sequence modeling. Formally, the self-attention mechanism constructs context-dependent representations by computing weighted combinations of input tokens through the use of query (Q), key (K), and value (V) matrices. As originally introduced in **Attention Is All You Need** [1], the output of self-attention is defined as:

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

Where d_k is the dimensionality of the key vectors.

While the original formulation of self-attention established the foundation for modern sequence modeling, subsequent research has proposed numerous variants to address its quadratic complexity. The Sparse Transformer introduced fixed sparsity patterns

to restrict attention to selected token pairs, reducing the overall number of computations [8]. Building on this idea, Long-former extended sparsity by combining local sliding-window attention with global tokens, making it suitable for processing long documents efficiently [9]. Similarly, Big-Bird generalized sparse attention with a mix of local, random, and global connections, offering theoretical guarantees for universal approximation while maintaining linear complexity [10]. In contrast, Performer replaced the exact softmax attention with a kernel-based approximation using random feature methods, enabling attention to be computed in linear time [11]. More recently, Mamba departed from purely attention-based mechanisms by incorporating selective state space models (SSMs) into a Transformer-like architecture, achieving linear-time sequence modeling while retaining expressive power [12]. Collectively, these milestones illustrate how self-attention has evolved along multiple dimensions: sparsity, approximation, and architectural redesign, while underscoring the persistent tension between efficiency and representational capacity, motivating the exploration of quantum-enhanced alternatives.

2.2 Parameterized Quantum Circuits (PQCs)

Parameterized quantum circuits (PQCs) represent a fundamental paradigm in quantum computation, where unitary operations are expressed as families of gates dependent on continuous parameters. Formally, a PQC defines a unitary operator $U(\theta)$ constructed from a sequence of quantum gates whose rotation angles or phases are parameterized by a vector $\theta \in \mathbb{R}^d$. These circuits are central to the variational method in quantum computing, where optimization over θ is used to approximate target states, minimize energy functions, or encode structured transformations.

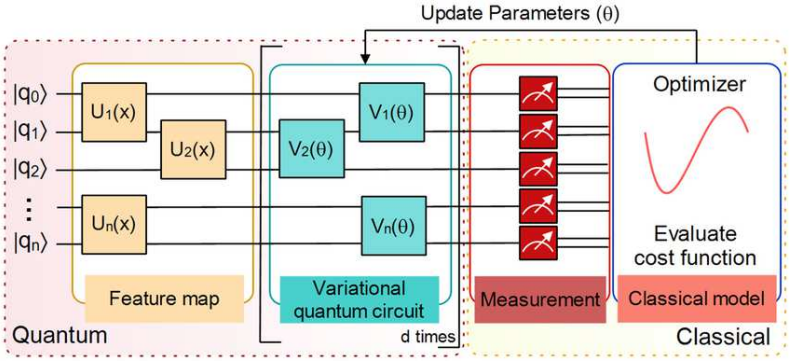


Fig. 1: Structure of a Variational Quantum Algorithm Using Parameterized Quantum Circuits (PQCs)

A PQC typically consists of alternating layers of parameterized single-qubit rotations and entangling gates, producing an ansatz capable of exploring rich subspaces of the Hilbert space. The figure 1 illustrates the workflow of a variational quantum algorithm built on parameterized quantum circuits (PQCs). The process begins with

the preparation of an initial quantum state $|q_0, q_1, \ldots, q_n\rangle$, which is encoded through a **feature map** $U(x)$ that embeds classical data into the quantum Hilbert space. This is followed by a **variational quantum circuit** $V(\theta)$, composed of parameterized unitary gates whose rotation angles θ are adjusted iteratively. After several layers of such parameterized operations, the quantum state is measured, and the outcomes are passed to a **classical optimizer**, which evaluates a cost function and updates the circuit parameters. This hybrid loop between quantum state evolution and classical optimization forms the basis of variational methods such as the Variational Quantum Eigensolver (VQE) and the Quantum Approximate Optimization Algorithm (QAOA) [13].

The expressive power of PQCs has been investigated in terms of circuit depth, entanglement generation, and universality, establishing them not only as central tools in variational algorithms such as VQE and QAOA, but also as a cornerstone of quantum machine learning (QML). In the QML setting, PQCs are employed as trainable models that map classical data into high-dimensional Hilbert spaces, enabling the extraction of complex correlations that may be intractable for classical architectures. Their tunable nature allows them to function analogously to parameterized layers in neural networks, while their quantum properties provide access to richer representational capacities. As such, PQCs serve as a general framework for building adaptable quantum learning models, bridging theoretical notions of expressivity with practical implementations on NISQ devices [14].

2.3 Quantum State Similarity Estimation

A central task in quantum information processing is the estimation of similarity, or **overlap**, between two quantum states. Given two pure states $|\psi\rangle$ and $|\phi\rangle$, their similarity is quantified by the squared inner product:

$$S(\psi, \phi) = |\langle\psi|\phi\rangle|^2,$$

which measures the probability that $|\psi\rangle$ is projected onto $|\phi\rangle$. Several quantum circuit-based techniques have been proposed to estimate this quantity, differing in resource requirements and noise sensitivity. Among the most widely studied are the swap, Hadamard tests, and quantum kernel methods. Other approaches include techniques based on interference experiments (e.g., Hong–Ou–Mandel interference) or shadow tomography for overlap estimation in more general mixed-state settings.

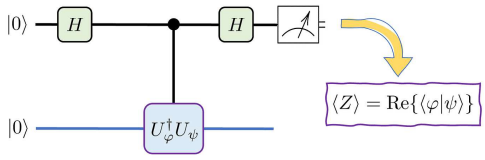


Fig. 2: The Hadamard Test Circuit for Inner Product Estimation

1. **Hadamard Test** is a broader primitive for estimating expectation values of the form $\langle\psi|U|\psi\rangle$, where U is a unitary operator [15]. As illustrate in Figure 2 the hadamard test circuit begins with an **ancilla qubit** (top wire) initialized in the state $|0\rangle$, which undergoes a **Hadamard gate (H)**. The ancilla then acts as a control for a unitary operation $U_\varphi^\dagger U_\psi$ applied to the second register (bottom wire), which is initialized in a reference state. A final Hadamard gate is applied to the ancilla before measurement, that statistics of the ancilla yield the expectation value

$$\langle Z\rangle = \mathrm{Re}\{\langle\varphi|\psi\rangle\},$$

which corresponds to the **real part of the overlap** between the states $|\varphi\rangle$ and $|\psi\rangle$ (See Figure 2). This construction generalizes to the estimation of expectation values of the form $\langle\psi|U|\psi\rangle$ for arbitrary unitaries U, making the Hadamard test a versatile primitive in quantum algorithms.

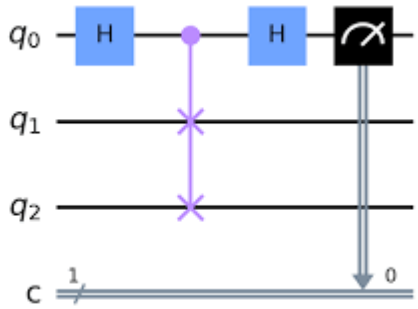


Fig. 3: Swap Test Circuit for Quantum State Overlap Estimation.

2. **Swap Test** is the canonical method for estimating state overlap [16]. As presented in Figure 3 it introduces an **ancilla qubit** prepared in $|0\rangle$, applies a Hadamard gate, followed by a controlled-SWAP between the two registers encoding $|\psi\rangle$ and $|\phi\rangle$. After a final Hadamard on the ancilla, measurement yields

$$P(0) = \frac{1 + |\langle\psi|\phi\rangle|^2}{2},$$

from which $|\langle\psi|\phi\rangle|^2$ can be inferred. While elegant, this approach requires both an ancilla and controlled-SWAP gates, which are costly on NISQ devices.
3. **Quantum Kernels** a different approach where classical data samples are mapped to quantum states $|\psi(x)\rangle$ and similarity is defined as their overlap kernel:

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

This strategy bypasses explicit overlap circuits by evaluating kernel matrices directly through repeated state preparation and measurement [17]. Quantum kernels thus provide a higher-level similarity framework, integrating naturally with classical kernel methods like SVMs.

3 Related Work

Early efforts to design quantum Transformer models have drawn upon Quantum linear algebra (QLA) methods, which exploit primitives such as block encoding and Quantum Singular Value Transformation (QSVT) to replicate matrix multiplications and nonlinearities within Transformers, with additional primitives such as Linear Combination of Unitaries (LCU) and amplitude transformations used for residual connections and normalization. In [4] authors introduced a full QLA-based Transformer framework, demonstrating theoretical polynomial-to-exponential speedups but relying on pre-trained parameters and costly block-encoding updates, which limit its practicality on NISQ hardware. Subsequent works combined QLA with variational circuits to reduce complexity [18], while others explored direct LCU–QSVT encoding [19]or qRAM-based residual connections [5]. Although elegant in theory, these designs remain infeasible for near-term devices due to their depth, ancilla requirements, and reliance on fault-tolerant resources.

In response to these limitations, the research focus has shifted toward **parameterized quantum circuit (PQC)-based approaches**, which are more compatible with today's noisy hardware. This line of work can be divided into four directions, each with distinct benefits and shortcomings. **(i) QKV-only quantum mapping** methods, such as QSANN [6], replace the projection matrices W_Q, W_K, W_V with PQCs to generate richer embeddings. While hardware-friendly, they do not address the attention computation bottleneck and rely heavily on repeated quantum–classical conversions. **(ii) Pairwise attention mechanisms** retain Q, K, and V as quantum states and directly compute similarities using swap [20], Hadamard tests [21], or quantum kernels [22]. These approaches best capture the essence of self-attention but still face quadratic scaling in sequence length and often require deep circuits or additional ancilla qubits. **(iii) Global attention methods** bypass pairwise comparisons by performing global token mixing in a single circuit using encoding strategies such as quantum orthogonal layers[23] or Fourier transforms [24]. Although they reduce complexity, they sacrifice the adaptive, context-dependent weighting that defines self-attention. **(iv) Quantum-assisted acceleration** applies quantum algorithms, such as Grover's search [7], to speed up classical attention computations. While offering polynomial improvements under sparsity assumptions, they do not replace classical architectures and lack experimental validation.

Among these categories, **pairwise attention methods** have attracted particular interest, as they provide the most faithful quantum analogy of classical self-attention. The study [20] employed the swap test on mixed quantum states to compute token similarities, where [25] extended this approach to global image embeddings, and [21] adapted the Hadamard test for molecular generation tasks, and [22] introduced quantum kernel methods to reduce resource requirements. Despite these advances, pairwise

approaches continue to struggle with **quadratic complexity, circuit depth, and reliance on ancilla qubits**, limiting their scalability on NISQ devices. Inner-product estimators such as swap and Hadamard tests achieve elegant constant-time overlap computations, but scaling them to full attention matrices requires $O(N^2)$ circuit executions. These challenges highlight the pressing need for **more efficient, low-depth, and noise-resilient quantum similarity estimation methods**, which can preserve the fidelity of attention while remaining feasible on near-term quantum hardware.

4 Destructive Swap Test For Word Similarity

4.1 Theoretical analysis : Existing Pairwise-based Attention Mechanisms

As outlined in the preliminaries, several quantum techniques have been explored to estimate token similarities directly in Hilbert space, including the Swap, Hadamard tests and quantum Kernels. While these approaches represent the most direct analogy to classical self-attention, they also expose important limitations: swap and Hadamard tests require ancillary qubits and suffer from circuit depth and noise fragility, whereas kernel methods reduce ancilla overhead but still retain quadratic complexity due to repeated sampling.

A particularly relevant alternative is the **destructive swap test**, a variant of the conventional swap test that eliminates the need for ancillary qubits. Instead of encoding similarity into the state of an ancilla, this method directly applies a Hadamard transformation to one register followed by a measurement of both registers. By exploiting the interference patterns of measurement outcomes, the squared inner product between two quantum states $|\psi\rangle$ and $|\phi\rangle$ can be reconstructed without additional qubits. Mathematically, if both registers are measured after the Hadamard operation, the probability distributions of outcomes encode $|\langle\psi|\phi\rangle|^2$, which can be extracted by statistical averaging.

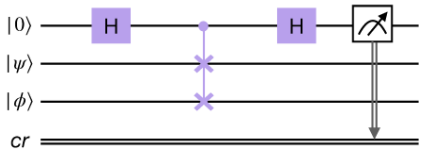
This adjustment not only reduces **qubit overhead** but also shortens **circuit depth**, as no controlled-SWAP operations are required. Moreover, the destructive nature of the measurement that discards the quantum registers after each run, makes the approach more **robust to hardware noise** since it minimizes entangling gates, which are typically the most error-prone on NISQ devices. Importantly, while the destructive swap test has been introduced in the quantum computing literature for state similarity estimation [26], it has not yet been exploited for **attention matrix computation**. This gap highlights a promising opportunity: adapting this method to compute pairwise token similarities in Transformers could combine the theoretical efficiency of quantum inner-product estimation with practical feasibility on near-term devices.

Table 1 provides a comparative overview of these methods, highlighting their qubit requirements, circuit depth, advantages, and drawbacks. Notably, while Swap and Hadamard tests offer straightforward formulations, their reliance on ancilla qubits and controlled operations constrains scalability on NISQ hardware. Quantum kernel approaches alleviate some of these issues but remain computationally demanding.

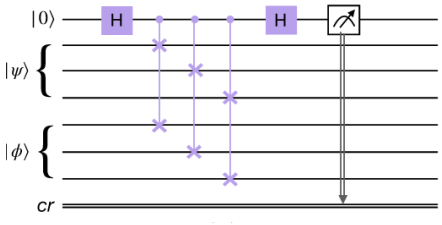
Table 1: Resource and complexity comparison of quantum similarity estimation methods for attention computation, expressed in terms of vector size n.

Method	Ancilla Qubits	Circuit Depth (as function of n)	Strengths	Limitations
Swap Test	1	$O(\log n)$	Direct overlap estimation; conceptually simple	Requires ancilla; sensitive to noise; $O(N^2)$ repetitions for full attention matrix
Hadamard Test	1	$O(n)$	Provides real/imag parts of inner product; useful for expectation values	Requires controlled unitaries; high depth; noise fragile
Quantum Kernels	0 (shared register)	$O(\log n)$	Ancilla-free; lower qubit count; efficient for small n	Sampling overhead; still $O(N^2)$ for full attention
Destructive Swap Test (variant)	0	$O(\log n)$	Ancilla-free; shallow circuits; more noise-resilient	Less explored; not yet applied to attention computation

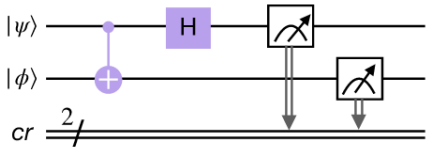
In contrast, destructive variant of the swap test offer a promising pathway, combining ancilla-free operation, low depth, and noise resilience properties that make it particularly attractive for NISQ-era attention computation.



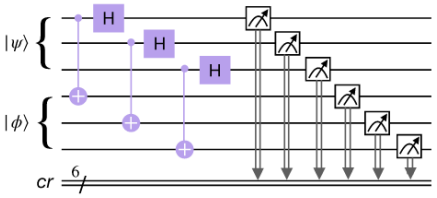
(a) Standard Swap Test Circuit for single qubit states



(b) Standard Swap Test Circuit for three-qubit states



(c) Destructive Swap Test circuit for single qubit states



(d) Destructive Swap Test ciruit for three-qubits states

Fig. 4: Comparison of quantum circuits of Standard Swap and Destructive Swap Test.

4.2 Destructive Swap vs Standard Swap Test For Word Similarity

Previous studies have already demonstrated the potential of destructive swap test–based protocols for similarity estimation in pattern recognition tasks. For instance, authors in [27] study applied both the standard swap test and its destructive variant (See Figure 4) to image data, showing that while the conventional swap test

could effectively detect similarities in low-dimensional cases, its performance degraded significantly under noise as the system size increased. In contrast, the destructive swap test exhibited enhanced robustness on real NISQ devices and enabled more reliable similarity evaluation across higher-dimensional image segments, including datasets such as MNIST and medical images.

Building on this foundation, our work extends the scope of destructive swap test applications beyond graphical data to the language domain, introducing it for the first time as a mechanism to compute word-level similarity within quantum attention. Unlike prior approaches in quantum pattern recognition, our focus lies in adapting this ancilla-free, low-depth, and noise-resilient circuit to the computation of attention matrices in Transformer architectures. This direction addresses a central limitation in existing PQC-based attention mechanisms, which remain constrained by the quadratic complexity of inner-product similarity computations and the fragility of circuits under realistic noise. By demonstrating that destructive swap test can effectively serve as a replacement for conventional similarity estimation within Transformer models, we contribute both a novel application context and a step toward practical, noise-aware quantum attention suitable for NISQ hardware.

4.3 Noise Analysis

To ensure that our evaluation realistically reflects the performance of quantum circuits on near-term devices, we complement ideal simulations with noisy simulations. In particular, we employ a depolarizing noise model, which is widely adopted in quantum algorithm benchmarking due to its ability to approximate the dominant noise processes in superconducting qubit platforms. The depolarizing channel acts by replacing the intended quantum state with the maximally mixed state with a certain probability p, thereby modeling the loss of coherence and gate imperfections in a hardware-agnostic manner.

In our implementation, a single-qubit gate is followed by a depolarizing error with probability $p = 0.001$, while two-qubit gates incur a higher error probability of $p = 0.01$, consistent with empirical observations that two-qubit operations are significantly more error-prone than single-qubit ones. These error rates capture the relative fragility of entangling operations, which are central to Pairwise-based similarity tests. By adopting this noise model, we are able to assess not only the theoretical advantage of destructive swap test over the standard swap test but also its practical robustness under realistic noise conditions.

Figure 5 presents a comparative analysis of quantum similarity estimation between the Standard Swap Test and the Destructive Swap Test under both ideal (noise-free) and noisy quantum simulation conditions. The evaluation encompasses three semantic similarity categories of word pairs: identical ("king–king"), semantically related ("king–queen"), and semantically distant ("king–ocean").

Under ideal (noiseless) conditions, both quantum similarity tests exhibit behavior consistent with theoretical predictions. They achieve a similarity value of 1.0 for identical pairs and show a gradual decrease as semantic distance increases, reflecting their sensitivity to relational structure. When noise is introduced, however, performance degradation becomes evident for both circuits. Despite this decline, the Destructive

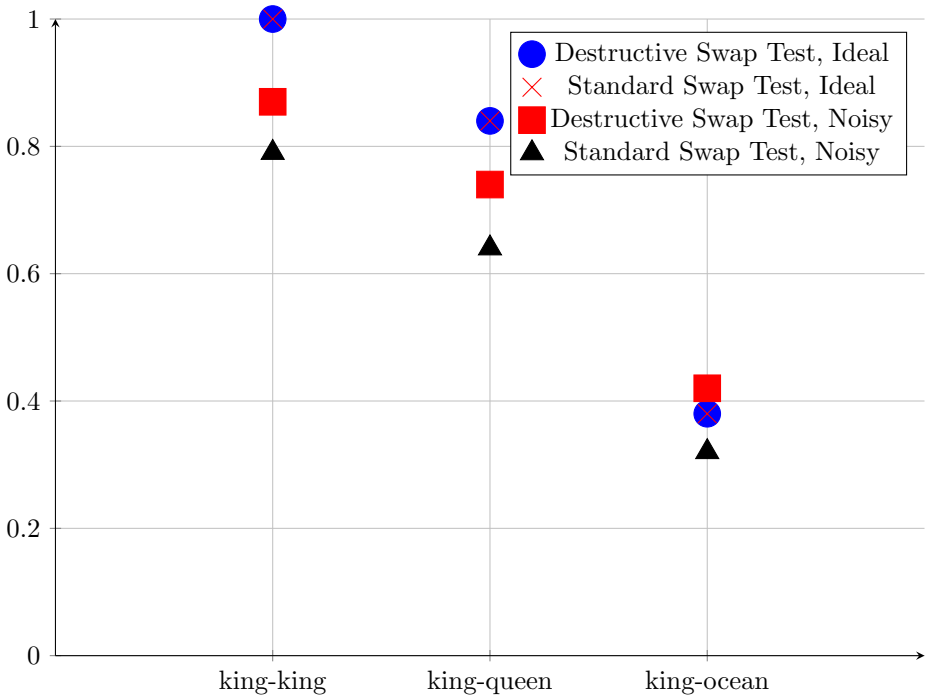


Fig. 5: Comparison of word similarity estimation using the Standard Swap Test and the Destructive Swap Test under both ideal and noisy simulation conditions. Results are shown for three semantic similarity levels between word pairs: identical (king, king), semantically near(king, queen), and semantically distant(king, ocean).

Swap Test consistently demonstrates greater robustness, preserving higher similarity values across all pair categories. The Standard Swap Test exhibits a similar qualitative pattern but with systematically lower fidelity values, suggesting a reduced capacity to maintain fine-grained semantic distinctions under noisy conditions. This contrast is particularly pronounced for the king–queen pair, where the Destructive Swap Test remains closer to the ideal noiseless result, further underscoring its enhanced resilience and stability. These observations collectively indicate that the destructive variant provides a more reliable similarity estimation mechanism in realistic, noise-affected quantum environments.

The figure 6 illustrates another quantitative comparison between the Standard Swap Test and the Destructive Swap Test under noisy simulation conditions. The x-axis represents the embedding vector dimensionality (8, 16, and 32 features), while the y-axis quantifies the similarity estimation accuracy (fidelity).

From the plotted data, both tests exhibit a clear performance degradation as the embedding dimensionality increases. This trend aligns with the intuition that higher-dimensional quantum encoding is more sensitive to cumulative gate noise and decoherence effects. However, the Destructive Swap Test consistently achieves higher similarity scores than the Standard Swap Test across all feature dimensions. The

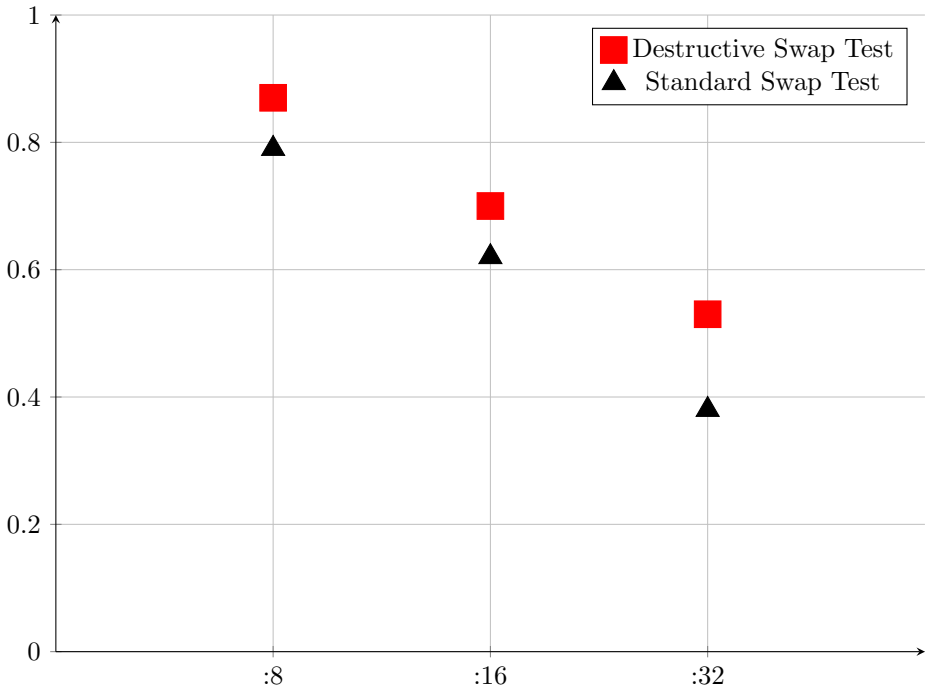


Fig. 6: Comparison of word similarity estimation using the Standard Swap Test and the Destructive Swap Test under noisy simulation conditions. Results are presented for identical word pairs (e.g., "king, king") across three embedding vector dimensions: 8, 16, and 32 features, respectively.

advantage of the destructive variant becomes more pronounced at larger dimensions (d=32), where the fidelity difference exceeds 10–15%, suggesting that it is more noise-resilient in complex encoding scenarios.

These findings underscore the superior robustness of the Destructive Swap Test under realistic noisy conditions. The destructive approach, which avoids ancilla-based interference and reduces circuit depth, mitigates cumulative error propagation, leading to more reliable similarity estimations in higher-dimensional quantum embeddings.

5 Methodology

This study employs the destructive swap test as the foundational mechanism for quantum similarity computation within a hybrid Transformer framework. Building upon prior demonstrations of its superior noise resilience and low circuit depth under NISQ constraints, the destructive swap test is leveraged as an efficient and scalable alternative for traditional Swap Test or inner-product–based similarity estimation. Extending its application from image-based to linguistic data, the proposed approach introduces word-level pairwise similarity computation directly within the quantum attention process. This methodological adaptation not only aligns with the practical

resource limitations of current quantum hardware but also advances the development of noise-tolerant quantum-enhanced attention mechanisms suitable for real-world NISQ environments.

5.1 Quantum Destructive Self-Attention Transformer Encoder-Only Architecture for Classification

The proposed model employs an encoder-only Transformer architecture augmented with a quantum-based similarity computation module integrated into the multi-head self-attention mechanism. This hybrid design enhances token-level relational modeling by incorporating quantum fidelity as an alternative attention scoring method, while maintaining compatibility with the classical Transformer framework.

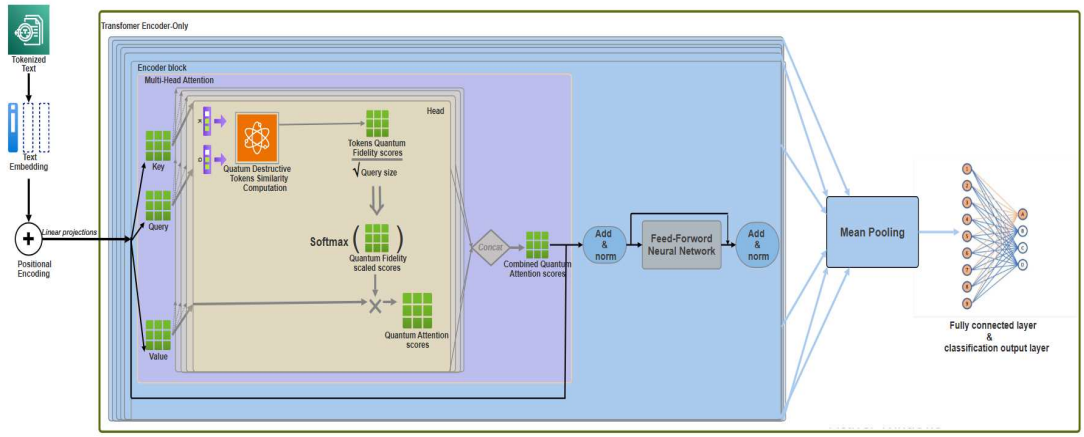


Fig. 7: Quantum Destructive Self-Attention Transformer Encoder-Only Architecture for Classification tasks

From Figure 7 the Input text is first tokenized and transformed into continuous vector embeddings through a trainable embedding layer. Positional encoding is then added to preserve the sequential structure. The resulting representations are linearly projected to form the query (Q), key (K), and value (V) matrices used throughout the attention mechanism. The key advancement of the pairwise quantum approach over classical Transformers lies in substituting the traditional dot-product similarity with a quantum fidelity–based computation. Each attention head performs the following operations:

1. **Quantum State Encoding:** Query and key vectors are encoded into parameterized quantum states using a suitable variational circuit.
2. **Destructive Swap Test:** A destructive swap test protocol is used to compute quantum fidelity between encoded tokens, producing a similarity matrix F. This method offers reduced circuit depth and improved robustness for NISQ hardware.

3. **Score Normalization:** Fidelity scores are scaled by the query dimension and passed through a softmax function to obtain quantum attention weights.
4. **Context Computation:** The value matrix is weighted by the quantum attention distribution to generate the head output.

Outputs from all heads are concatenated and processed through a residual connection followed by layer normalization, mirroring the standard Transformer design. Then the combined attention output is forwarded to a feed-forward neural network composed of two linear layers with a nonlinear activation in between. A residual connection and subsequent normalization layer ensure stable optimization and effective hierarchical representation learning. Since the architecture is encoder-only, token-level representations are aggregated using mean pooling across the sequence length. This pooled representation is then passed to a fully connected classification layer, producing either binary or multi-class predictions depending on the target task.

Quantum pairwise similarity methods provide an alternative to the classical dot-product attention by comparing token representations through the geometry of quantum states rather than through linear inner products. This enables the model to capture more expressive relationships between embeddings, especially when subtle variations in amplitude distributions are relevant. Among quantum similarity techniques, the Destructive Swap test offers key advantages that motivate its use in our design. Compared to the standard Swap and Hadamard-based fidelity estimators, it requires fewer controlled operations and achieves a shallower circuit structure, making it more robust to NISQ-level noise. In contrast to general quantum kernels, which often rely on repeated state preparation and measurement across large feature maps, the destructive swap test provides a direct and efficient fidelity estimate between pairs of encoded tokens. These properties make it a practical and noise-resilient quantum similarity module, strengthening the model's contextual reasoning while maintaining compatibility with classical Transformer pipelines.

6 Experimental Setup

To evaluate the effectiveness of the proposed Quantum Destructive Self-Attention Transformer architecture, we conducted a series of experiments on standard NLP classification tasks. Our experiments focus on assessing whether this quantum destructive swap test as a fidelity-based token similarity estimator yields measurable improvements in representing contextual dependencies, particularly in binary and multi-class text classification settings.

6.1 Datasets

In this work, our objective is to establish a comprehensive evaluation framework to assess the robustness and generalization capabilities of the proposed model across multiple classification tasks. To achieve this, we employ six benchmark datasets derived from the twenty introduced in [28], which were specifically designed for evaluating quantum NLP models under constrained qubit resources. We retain only the adapted

datasets, excluding the artificial ones that exhibit overly simplified structures and near-perfect performance in related studies. The selected datasets are balanced and derived from real-world corpora, subsequently adapted for quantum NLP tasks to ensure strict control over grammatical composition, label distribution, and vocabulary scope. This controlled configuration ensures compatibility with shallow quantum circuits and fixed grammatical categories. Such a design is crucial for aligning dataset complexity with qubit availability, thereby enabling a fair, scalable, and reproducible evaluation of quantum text classification architectures.

Dataset	Sub-data	Language	Labels	Seq.length	Train samples
KAGG	kagg-2	English	2	4	350
	kagg-3		3		420
	kagg-4		4		560
MILK	milk-2	Spanish	2	3	336
	milk-3		3		252
	milk-4		4		700

Table 2: Summary of benchmark datasets used in QNLP experiments, including dataset subdivision, languages,label sets, sequence length and the number of training samples. [28]

Table 2 summarizes the benchmark datasets utilized in the quantum natural language processing (QNLP) experiments. The datasets are categorized into two main groups: KAGG (English) and MILK (Spanish), each further subdivided into three variations based on the number of label classes. Specifically, each dataset includes binary (2-class), ternary (3-class), and quaternary (4-class) classification tasks.

The datasets are syntactically constrained to enable controlled linguistic analysis. The English KAGG dataset consists of sequences with a fixed length of 4 tokens, adhering to a uniform syntactic template of **Determiner + Noun + Verb + Adjective** . In contrast, the Spanish MILK dataset employs 3-token sequences with more variable internal structures, including patterns such as **Noun + (Brand + Adjective / Adjective + Brand)** and **Brand + (Noun + Adjective / Adjective + Noun)**. This design introduces controlled syntactic variability between the language groups.

Training sample sizes are strategically varied across subdivisions to test model robustness across different data scales. Sample counts range from the smallest subset, milk-3 (252 samples), to the largest, milk-4 (700 samples). This structured collection facilitates a multi-faceted comparative analysis, allowing for a systematic investigation of model performance across differing languages (English vs. Spanish), label granularity (2 to 4 classes), syntactic structures (fixed vs. variable templates), and dataset sizes (250 to 700 samples). Therefore it is well-suited for probing the interaction between linguistic complexity, task difficulty, and model generalization in the QNLP domain.

6.2 Evaluation Protocol

We implemented a multi-faceted evaluation protocol to rigorously assess model performance and, critically, inspect the internal behavior of our novel quantum attention mechanism. During training and validation, performance was primarily tracked using **accuracy** and the **F1-score**. While accuracy provides a straightforward measure of overall correctness, the F1-score is particularly valuable as it balances precision and recall, providing a more robust metric for datasets [29]:

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN},$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

For the final test set, we extended our analysis to include the **Area Under the Receiver Operating Characteristic Curve (AUC-ROC)**. This threshold-independent AUC-ROC score summarizes model's overall discriminative power across all classification thresholds [30].

6.3 Implementation Details

Our model is implemented using a hybrid quantum–classical pipeline that combines PennyLane, Qiskit, and PyTorch, executed on Google's backend environment with NVIDIA GPU acceleration. PennyLane serves as the core quantum–classical interface, enabling the construction and optimization of quantum neural networks while integrating smoothly with PyTorch's computation graph. This integration allows us to leverage automatic differentiation, efficient batching, and GPU-accelerated tensor operations, thereby streamlining the training process of the quantum-enhanced attention module.

For quantum simulation, we rely on the Qiskit Aer quantum simulator, chosen for its capability to support both noiseless and noisy quantum experiments. This dual-mode operation is essential for evaluating the practical behavior of our model on current NISQ hardware as well as its idealized performance under fault-tolerant conditions. In noiseless simulations, we adopt the statevector method, which provides exact quantum state evolution and is well-suited for analyzing the theoretical upper bound of model performance. For noisy simulations, we use the density-matrix method, the recommended approach within Qiskit for accurately modeling decoherence and gate-level noise while maintaining compatibility with GPU execution. These choices align with best practices for quantum simulation and ensure that both experimental regimes are handled with high fidelity.

The use of NVIDIA GPUs benefits not only the classical components of the Transformer but also enables high-performance quantum simulation through Qiskit-Aer-GPU package. This unified GPU-based workflow significantly reduces simulation time, supports larger batch sizes, and offers a scalable and rapid environment for experimenting with hybrid models. Overall, this implementation strategy enables efficient training, robust noise evaluation, and reliable benchmarking of the proposed Quantum Destructive Transformer across diverse experimental settings.

6.4 Results & Discussion

The results reported in Table 3 reveal several structured patterns across datasets, noise regimes, and classification complexity levels. First, in terms of training time, a substantial increase is observed when moving from noiseless to noisy simulations, particularly for multi-class settings. Binary tasks (KAGG-2, MILK-2) require only 13–14 minutes in the noiseless case, whereas noisy execution increases the cost to approximately 80 minutes. This computational gap becomes considerably more pronounced for higher-class problems, especially in four-class datasets, where training extends to 24 hours. This confirms that circuit execution under noise and increasing label cardinality significantly amplifies computational overhead, reflecting the cumulative effect of deeper circuit evaluations and repeated sampling.

For binary classification, the KAGG-2 dataset exhibits near-perfect training F1-scores (99.72%) in both noiseless and noisy cases, indicating a strong fitting capacity. Interestingly, noisy execution slightly improves validation, testing F1, accuracy, and AUC. In contrast, MILK-2 shows a marked degradation under noise: while noiseless performance is strong (testing F1 = 98%, AUC = 100%), noisy results drop substantially (testing F1= 68%, AUC = 71%). This indicates that dataset complexity and structure strongly influence noise sensitivity.

For three-class classification, KAGG-3 maintains excellent training performance ($\approx 100\%$) in both regimes, with relatively stable validation and testing metrics. Noise introduces only minor degradation, demonstrating robustness on this dataset. However, MILK-3 behaves differently: although training F1 remains moderate (=65%) in both regimes, noisy validation and testing scores improve slightly compared to the noiseless case, and AUC increases notably (from 70.20% to 78.79% for AUC). This suggests that, for this configuration, noise may act as an implicit regularizer, mitigating overfitting and improving generalization.

For four-class classification, divergence between datasets becomes more evident. In KAGG-4, noiseless performance is substantially stronger than noisy performance across all evaluation metrics, indicating higher vulnerability to noise as class complexity increases. Conversely, MILK-4 presents a different pattern: although overall performance remains moderate, the noisy configuration slightly outperforms the noiseless one in validation, testing F1, accuracy, and AUC. This again supports the hypothesis that, in certain complex regimes, controlled noise can stabilize learning dynamics.

Dataset	Noiseless						Noisy					
	Training		Validation	Testing			Training		Validation	Testing		
	Time	F1-score		Accuracy	AUC		Time	F1-score		Accuracy	AUC	
Binary Classification												
KAGG-2	**14 min**	**99.72%**	81.36%	78.74%	73.00%	79.36%	84 min	**99.72%**	**83.33%**	**80.33%**	**76.00%**	**89.00%**
MILK-2	**13 min**	**90.38%**	**100.00%**	**97.96%**	**97.92%**	**100.00%**	81 min	63.95%	66.67%	67.77%	59.38%	70.62%
Three-Class Classification												
KAGG-3	**15 min**	**100.00%**	71.67%	**80.83%**	**80.83%**	**93.13%**	-	99.52%	**73.33%**	79.17%	79.17%	91.29%
MILK-3	**339 min**	64.68%	44.44%	51.39%	51.39%	70.20%	608 min	**65.87%**	**66.67%**	**52.78%**	**52.78%**	**78.79%**
Four-Class Classification												
KAGG-4	-	**93.39%**	**62.86%**	**75.71%**	**75.71%**	**87.97%**	-	66.25%	46.25%	50.00%	50.00%	74.85%
MILK-4	1020 min	51.86%	61.00%	48.00%	48.00%	83.32%	-	**53.43%**	**64.00%**	**58.00%**	**58.00%**	**84.35%**

Table 3: Training Time and Best Performance Comparison using Destructive Quantum Attention Computation in Noiseless and Noisy quantum simulation environment for Binary, Three-Class, and Four-Class Classification Tasks

Overall, two principal trends emerge. First, classification complexity (binary → multi-class) significantly increases computational cost and generally reduces performance stability. Second, the impact of noise is dataset-dependent: while it degrades performance in some cases (notably MILK-2 and KAGG-4), it can also provide a regularization-like effect in others (MILK-3 and MILK-4). These findings highlight that the destructive quantum attention mechanism demonstrates competitive robustness under noisy simulation, but its sensitivity is closely tied to dataset structure and task complexity, underscoring the importance of noise-aware evaluation in NISQ-oriented quantum NLP architectures.

6.5 Benchmarking

This study establishes a comparative benchmarking framework to evaluate the performance of our proposed quantum destructive attention computation against established quantum pair-wise methods and the classical scaled dot-product attention baseline. All models are trained and evaluated under identical conditions within a noisy quantum simulation environment, using a consistent set of hyper-parameters across the six benchmark datasets. The results of this comprehensive comparison, presented in Table 4, detail the performance across training, validation, and testing phases for each method and dataset.

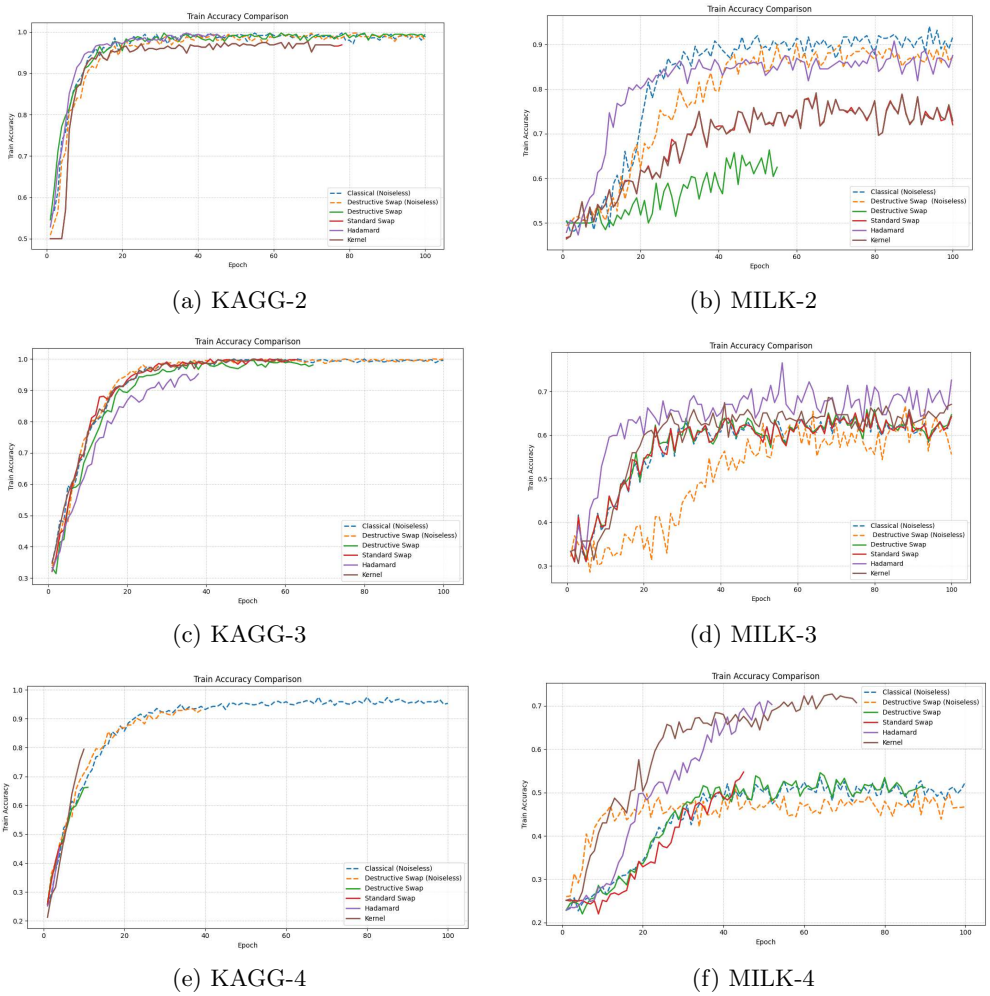
The comparative analysis highlights clear distinctions among the studied approaches in terms of efficiency, robustness, and adaptability under NISQ constraints. The destructive swap–based method emerges as the fastest quantum approach, second only to the classical baseline, while the Hadamard-based method records the longest training time, reflecting its higher circuit complexity and reduced efficiency when limited qubits and noise are considered.

On the KAGG dataset, both the destructive and kernel methods achieved performance levels comparable to the classical baseline, with the destructive model offering similar accuracy at a significantly reduced training time. Despite the presence of overfitting across all models (ranging between 4% and 45%), the destructive model maintains a balanced accuracy–efficiency trade-off, demonstrating that it can preserve generalization while remaining computationally economical.

In contrast, performance on the MILK dataset is more moderate across all approaches. The Hadamard and kernel methods outperform other models, while the classical baseline experiences a notable performance drop due to the dataset's higher complexity and multi-class configuration. Although the destructive approach exhibits weaker training scores in certain cases (e.g., MILK-3), it achieves superior validation and testing performance, revealing stronger generalization capabilities that are an essential quality for real-world deployment on noisy hardware.

Dataset	KAGG						MILK					
	Training		Validation	Testing			Training		Validation	Testing		
	Time	F1-score			Accuracy	AUC	Time	F1-score			Accuracy	AUC
Binary Classification												
Destructive Swap	1.4 h/100	**99.72%**	**83.33**	80.33%	76.00%	**89.00%**	1.3 h/100	63.95%	66.67%	66.67%	59.38%	70.62%
Standard Swap	-/78	97.22%	64.86%	57.89%	68.00%	79.24%	15 h/100	79.32%	**95.83**%	93.07%	92.71%	94.75%
Hadamard Test	-/47	**99.72%**	**83.33%**	**81.33%**	**78.00%**	83.88%	21 h/100	91.07%	78.69%	84.21%	81.25%	86.89%
Quantum Kernel	-/76	97.22%	64.86%	57.89%	68.00%	79.24%	-/49	66.86%	62.22%	62.92%	65.62%	73.22%
Dot-product	**0.2 h/100**	**99.72%**	74.63%	72.99%	63%	66.52%	**0.2 h/100**	**93.98**%	95.65%	**96.77**%	**96.88**%	**100**%
Three-Class Classification												
Destructive Swap	-/67	99.52%	**73.33%**	79.17%	79.17%	91.29%	0.9 h/100	65.48%	33.33%	36.11%	36.11%	52.75%
Standard Swap	-/63	**100.00%**	68.33%	80.00%	80.00%	91.37%	11 h/100	65.08%	**66.67%**	52.78%	52.78%	80.06%
Hadamard Test	-/38	95.24%	50.00	69.17%	69.17%	91.18%	16 h/100	**72.62%**	**66.67%**	**66.67%**	**66.67%**	84.87%
Quantum Kernel	-/61	**100.00%**	71.67%	**80.83%**	**80.83%**	**92.47%**	12 h/100	68.65%	41.67%	59.72%	59.72%	**87.56%**
Dot-product	**0.2 h/ 100**	**100.00%**	71.67%	**80.83%**	**80.83%**	92.12%	0.16 h/100	65.08%	66.67%	52.78%	52.78%	79.63%
Four-Class Classification												
Destructive Swap	-/12	66.25%	46.25%	50.00%	50.00%	74.85%	-/91	53.43%	64.00%	58.00%	58.00%	84.35%
Standard Swap	-/6	57.50%	47.50%	53.75%	53.75%	73.89%	-/45	54.71%	41.00%	44.00%	44.00%	67.41%
Hadamard Test	-/4	45.18%	27.50%	41.25%	41.25%	75.01%	-/52	71.14%	**85.00%**	**76.00%**	**76.00%**	**95.55%**
Quantum Kernel	-/10	79.46%	50%	61.25%	61.25%	86.11%	-/72	**75.71%**	62.00%	62.00%	62.00%	83.15%
Dot-product	0.4 h /100	**97.50%**	**58.75%**	**65.00%**	**65.00%**	**86.17%**	0.45 h/100	52.26%	61.00%	63.00%	63.00%	84.88%

Table 4: Training Time and Performance Benchmarking of Quantum Pairwise Attention Computation Methods under Noisy Quantum Simulation for Binary, Three-Class, and Four-Class Classification Tasks, Including the Classical Scaled Dot-Product Baseline



(a) KAGG-2 (b) MILK-2

(c) KAGG-3 (d) MILK-3

(e) KAGG-4 (f) MILK-4

Fig. 8: Training accuracy convergence curves across six benchmark datasets under a noisy simulation environment, comparing different quantum attention computation circuits, with additional results highlighting the classical dot-product and destructive swap circuits under noiseless conditions.

Overall, the models demonstrate robust performance on KAGG but face greater challenges on MILK, where a quantum advantage becomes more apparent as the classical model's performance declines. Among quantum counterparts, the kernel and destructive swap methods stand out for their robustness and scalability, with the destructive swap circuit achieving competitive accuracy at minimal computational cost. These findings confirm it's suitability for noise-resilient, shallow-depth quantum attention, offering a promising path toward efficient and practical implementations on near-term quantum devices.

From Figure 8 in the KAGG side, for binary classification task we observe that all models converge after about 20 epochs with near-perfect accuracy (0.99). The Kernel and Standard Swap methods converge at a slightly lower level (0.95). In KAGG-3, convergence requires around 40 epochs, indicating increased dataset complexity. Here, the Hadamard method is the only one that plateaus early without reaching full convergence. In KAGG-4, most noisy variants stop before converging, which may impair their generalization on unseen data. Overall, convergence is clearly influenced by dataset complexity, and the noiseless experiments Classical and Destructive Swap (Noiseless) consistently achieve the highest near-perfect results.

In the MILK side, for binary classification models converge at different accuracy levels. Noiseless experiments lead at 0.9, followed by Hadamard, while both Standard Swap and Destructive Swap variants converge at the lowest level (0.6). In MILK-3, noiseless experiments are no longer superior; quantum Kernel and Hadamard test achieve the highest convergence (0.7). Interestingly, the Destructive Swap (noisy) outperforms its noiseless counterpart here, suggesting more then a degree of noise-resilience but even noise-enhanced performance. In MILK-4, noiseless experiments fall to the weakest level (0.5), while Hadamard and Kernel maintain 0.7, and Destructive Swap again benefits from noise, surpassing its noiseless version.

The results suggest that quantum advantage may be context-dependent: while classical methods and noise-free experiments dominate in ideal settings, certain quantum circuits can match or exceed them in more complex, noisy regimes—hinting at potential noise-robust quantum learning in suitable problem domains.

Model	**KAGG**			**MILK**		
	Training	**Testing**		**Training**	**Testing**	
	Accuracy		**F1-score**	**Accuracy**		**F1-score**
Binary Classification						
Lambeq	76.25%	/	/	95.83%	/	/
SQO-LSTM	99.71%	78.00%	**81.97%**	**100**	**100%**	**100%**
Our approach	**99.72**	**78.74%**	73.00%	90.00%	97.92%	97.96%
Three-Class Classification						
Lambeq	73.968%	/	/	71.88	/	/
SQO-LSTM	**99.29%**	**80.00%**	**80.00%**	**98.40%**	**85.71%**	**85.71%**
Our approach	87.62%	78.33%	78.33%	64.68%	51.39%	51.39%
Four-Class Classification						
Lambeq	70.57%	/	/	72.81%	/	/
SQO-LSTM	**98.75%**	61.25%	61.25%	**96.00%**	**80.50%**	**80.50%**
Our approach	93.39%	**75.71%**	**75.71%**	50.71%	48.50%	48.50%

Table 5: Best performance comparison of hybrid quantum-classical attention mechanisms in a noiseless simulation environment for binary, three-class, and four-class classification tasks.

6.6 Comparative Analysis

This section presents a comparative analysis of our proposed quantum Destructive attention Transformer against established methodologies in the literature that have been evaluated on the same datasets. The selected studies, detailed below, serve as benchmarks to contextualize its performance and innovation within the field of QNLP.

1. **Lambeq Comparative Study :** This foundational study introduces a suite of synthetic, grammatically-structured datasets for sentiment analysis (e.g., AD2–AD5) and Named Entity Recognition (e.g., KAGG, MILK), providing balanced, formal-text corpora in English and Spanish. It conducts a systematic evaluation using the DisCoPy and lambeq toolkits to construct and test various quantum circuit ansatzes (e.g., IQP, Sim14, Sim15, Strong Entangling) on these datasets. The key findings indicate that while quantum simulations can achieve binary classification accuracies exceeding 83%, outperforming certain classical methods, this comes at the cost of substantially longer training times. Furthermore, experiments on real quantum hardware reveal performance limitations due to current constraints such as qubit count and noise. These insights underscore that although QNLP is promising, further advancements in both quantum software and hardware are necessary for scaling to more complex tasks.
2. **SQO-LSTM:** [31] This work proposes an optimized Quantum Long Short-Term Memory (SQO-LSTM) model designed to reduce computational overhead by consolidating multiple Variational Quantum Circuits (VQCs) into a single shared circuit while preserving core LSTM gating functionality. This architecture demonstrates consistent and near-perfect performance across all datasets, offering competitive advantages over classical and QLSTM baselines using multiple VQCs under identical experimental setting. The performance gains are attributed to its parameter-sharing mechanism, which acts as an implicit regularizer by limiting redundant degrees of freedom. This constraint mitigates overfitting in low-data regimes and stabilizes training by ensuring a unified gradient flow, thereby maintaining expressive temporal dynamics with enhanced computational efficiency.

Table 5 presents a comparative benchmark of three hybrid quantum-classical models: Lambeq, SQO-LSTM, and Our approach evaluated in a noiseless simulation environment. The comparison is structured across two primary datasets (KAGG and MILK) and three classification tasks: binary, three-class, and four-class. For each model and task, performance is reported in terms of Training Accuracy and Testing Accuracy/F1-score. Notably, the Lambeq study only reports training accuracy.

The experimental results reveal several systematic trends. First, all models exhibit a performance decrease as the number of classes increases, with a more pronounced degradation observed on the MILK dataset compared to KAGG. This significant disparity is attributable to MILK's variable syntactic sentence structure, which presents a greater linguistic challenge than KAGG's fixed template. Second, the SQO-LSTM model consistently demonstrates the highest training performance, achieving near-perfect or perfect accuracy across tasks (e.g., 100% on MILK-2). Third, the performance of the remaining two models diverges based on the dataset:

1. On KAGG, our model presents competitive training accuracy relative to SQO-LSTM and achieves superior test performance as task complexity increases (e.g., 75.71% vs. 61.25% on KAGG-4), suggesting stronger generalization capabilities, whereas the Lambeq baseline shows the lowest training performance.
2. On MILK, the trend differs markedly: our model performs significantly lower in both training and testing compared to the two benchmarks, while Lambeq maintains moderate training accuracy, though the absence of its test metrics limits any conclusion regarding its generalization.

This comparative study reveals that SQO-LSTM excels at fitting training data, whereas our proposed model exhibits stronger generalization, particularly on structured English (KAGG) tasks, where it sustains robust performance even as complexity rises. The pronounced performance gap between KAGG and MILK datasets underscores the sensitivity of quantum models to linguistic and structural variations in the data. Overall, while certain quantum architectures achieve high accuracy, their efficacy is highly contingent on both task complexity and dataset characteristics.

7 Conclusion

This work introduced a hybrid quantum–classical Transformer architecture in which the Destructive Swap test serves as a quantum similarity primitive for pairwise attention computation. Through systematic benchmarking across multiple linguistic datasets for classification tasks under both noiseless and noisy quantum simulations, we demonstrated that Destructive approach achieves a favorable balance between computational efficiency, noise resilience, and architectural scalability, aligning with the practical constraints of NISQ-era hardware. Compared with Hadamard-based, standard swap, and kernel-based similarity circuits, the destructive variant consistently delivered competitive or superior performance while significantly reducing training time, establishing its practical advantage for resource-constrained quantum execution.

The empirical results further indicate that although classical attention mechanisms remain effective in ideal settings, their performance degrades with increasing task complexity. In contrast, quantum similarity circuits, particularly the destructive approach exhibit enhanced robustness and, in several cases, improved generalization. These findings provide concrete evidence of a quantum advantage, especially in complex or noise-affected regimes where classical baselines weaken.

Overall, the study validates the feasibility of quantum-enhanced attention for natural language processing and highlights the destructive swap test as an efficient, noise-aware solution for scalable hybrid quantum architectures.

Future research will focus on experimental validation on real quantum hardware, hardware-aware circuit compilation and depth reduction, and the extension of quantum attention mechanisms to generative modeling tasks.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural

information processing systems **30** (2017)

[2] He, W., Wu, Y., Li, X.: Attention mechanism for neural machine translation: A survey. In: 2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), vol. 5, pp. 1485–1489 (2021). IEEE

[3] Zhang, H., Zhao, Q.: A survey of quantum transformers: Technical approaches, challenges and outlooks. arXiv preprint arXiv:2504.03192 (2025)

[4] Guo, N., Yu, Z., Choi, M., Agrawal, A., Nakaji, K., Aspuru-Guzik, A., Rebentrost, P.: Quantum linear algebra is all you need for transformer architectures. arXiv preprint arXiv:2402.16714 (2024)

[5] Xue, C., Chen, Z.-Y., Zhuang, X.-N., Wang, Y.-J., Sun, T.-P., Wang, J.-C., Liu, H.-Y., Wu, Y.-C., Wang, Z.-L., Guo, G.-P.: End-to-end quantum vision transformer: Towards practical quantum speedup in large-scale models. arXiv preprint arXiv:2402.18940 (2024)

[6] Li, G., Zhao, X., Wang, X.: Quantum self-attention neural networks for text classification. Science China Information Sciences **67**(4), 142501 (2024)

[7] Gao, Y., Song, Z., Yang, X., Zhang, R.: Fast quantum algorithm for attention computation. arXiv preprint arXiv:2307.08045 (2023)

[8] Child, R., Gray, S., Radford, A., Sutskever, I.: Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509 (2019)

[9] Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 (2020)

[10] Zaheer, M., Guruganesh, G., Dubey, K.A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., *et al.*: Big bird: Transformers for longer sequences. Advances in neural information processing systems **33**, 17283–17297 (2020)

[11] Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al.: Rethinking attention with performers. arXiv preprint arXiv:2009.14794 (2020)

[12] Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)

[13] Cerezo, M., Arrasmith, A., Babbush, R., Benjamin, S.C., Endo, S., Fujii, K., McClean, J.R., Mitarai, K., Yuan, X., Cincio, L., *et al.*: Variational quantum algorithms. Nature Reviews Physics **3**(9), 625–644 (2021)

[14] Benedetti, M., Lloyd, E., Sack, S., Fiorentini, M.: Parameterized quantum circuits as machine learning models. Quantum science and technology **4**(4), 043001 (2019)

[15] Aharonov, D., Kitaev, A., Nisan, N.: Quantum circuits with mixed states. In: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, pp. 20–30 (1998)

[16] Buhrman, H., Cleve, R., Watrous, J., De Wolf, R.: Quantum fingerprinting. Physical review letters **87**(16), 167902 (2001)

[17] Schuld, M., Killoran, N.: Quantum machine learning in feature hilbert spaces. Physical review letters **122**(4), 040504 (2019)

[18] Liao, Y., Ferrie, C.: Gpt on a quantum computer. arXiv preprint arXiv:2403.09418 (2024)

[19] Khatri, N., Matos, G., Coopmans, L., Clark, S.: Quixer: A quantum transformer model. arXiv preprint arXiv:2406.04305 (2024)

[20] Chen, F., Zhao, Q., Feng, L., Chen, C., Lin, Y., Lin, J.: Quantum mixed-state self-attention network. Neural Networks **185**, 107123 (2025)

[21] Smaldone, A.M., Shee, Y., Kyro, G.W., Farag, M.H., Chandani, Z., Kyoseva, E., Batista, V.S.: A hybrid transformer architecture with a quantized self-attention mechanism applied to molecular generation. Journal of Chemical Theory and Computation **21**(10), 5143–5154 (2025)

[22] Zhao, R.-X., Shi, J., Li, X.: Qksan: A quantum kernel self-attention network. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)

[23] Kerenidis, I., Mathur, N., Landman, J., Strahm, M., Li, Y.Y., *et al.*: Quantum vision transformers. Quantum **8**, 1265 (2024)

[24] Evans, E.N., Cook, M., Bradshaw, Z.P., LaBorde, M.L.: Learning with sasquatch: a novel variational quantum transformer architecture with kernel-based self-attention. arXiv preprint arXiv:2403.14753 (2024)

[25] Zhang, H., Zhao, Q., Zhou, M., Feng, L.: Hqvit: Hybrid quantum vision transformer for image classification. arXiv preprint arXiv:2504.02730 (2025)

[26] Garcia-Escartin, J.C., Chamorro-Posada, P.: Swap test and hong-ou-mandel effect are equivalent. Physical Review A—Atomic, Molecular, and Optical Physics **87**(5), 052330 (2013)

[27] Das, S., Zhang, J., Martina, S., Suter, D., Caruso, F.: Quantum pattern recognition on real quantum processing units. Quantum Machine Intelligence **5**(1), 16 (2023)

[28] Peral-García, D., Cruz-Benito, J., García-Peñalvo, F.J.: Comparing natural language processing and quantum natural processing approaches in text classification tasks. Expert Systems with Applications, 124427 (2024)

[29] Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management **45**(4), 427–437 (2009)

[30] Nakas, C.T., Bantis, L.E., Gatsonis, C.A.: ROC Analysis for Classification and Prediction in Practice. Chapman and Hall/CRC, ??? (2023)

[31] Bouakba, Y., Belhadef, H., Saadi, A.: Sqo-lstm: a single quantum-output long short-term memory for classification tasks: Y. bouakba et al. The Journal of Supercomputing **81**(18), 1637 (2025)