CLAQS: Compact Learnable All-Quantum Token Mixer with Shared-ansatz for Text Classification

Junhao Chen[1,*], Yifan Zhou[1,*], Hanqi Jiang[1], Yi Pan[1],
Yiwei Li[1], Huaqin Zhao[1], Wei Zhang[2], Yingfeng Wang[3,†], Tianming Liu [1,†]
[1]University of Georgia, [2]Augusta University, [3]University of Tennessee at Chattanooga

Abstract

Quantum compute is scaling fast, from cloud QPUs to high throughput GPU simulators, making it timely to prototype quantum NLP beyond toy tasks. However, devices remain qubit limited and depth limited, training can be unstable, and classical attention is compute and memory heavy. This motivates compact, phase aware quantum token mixers that stabilize amplitudes and scale to long sequences. We present CLAQS, a compact, fully quantum token mixer for text classification that jointly learns complex-valued mixing and nonlinear transformations within a unified quantum circuit. To enable stable end-to-end optimization, we apply ℓ^1 normalization to regulate amplitude scaling and introduce a two-stage parameterized quantum architecture that decouples shared token embeddings from a window-level quantum feed-forward module. Operating under a sliding-window regime with document-level aggregation, CLAQS requires only eight data qubits and shallow circuits, yet achieves 91.64% accuracy on SST-2 and 87.08% on IMDB, outperforming both classical Transformer baselines and strong hybrid quantum–classical counterparts.

1 Introduction

Transformer-based language models have revolutionized natural language processing (NLP) by introducing the self-attention mechanism, which enables flexible token interactions and drives state-of-the-art performance across a wide range of tasks (Vaswani et al., 2017). However, attention-based architectures come at the cost of massive parameter counts and quadratic computational complexity, motivating the search for more efficient sequence mixers. Recent studies demonstrate that high-quality representations can arise from alternative mixing operations, such as Fourier or convolutional transforms (Lee-Thorp et al., 2022), suggesting that attention may not be the only viable paradigm for effective token integration.

Meanwhile, advances in quantum computing have opened new directions for sequence modeling. Quantum systems can represent information in exponentially large Hilbert spaces via superposition and entanglement, offering a unique mechanism for encoding complex linguistic dependencies beyond classical vector spaces. Yet, Noisy Intermediate-Scale Quantum (NISQ) hardware restricts circuit depth and qubit count, making fully quantum language models infeasible at scale. This has spurred the development of hybrid quantum–classical approaches, such as Adaptive Quantum–Classical Fusion (AQCF) (Pan et al., 2025), which dynamically alternate between quantum and classical processing. However, these methods often rely on static quantum components and do not fully exploit the potential of learnable quantum representations.

In this paper, we propose **CLAQS**, a compact, end-to-end quantum–classical model designed for text classification under realistic NISQ constraints. Our key innovations are threefold:

- **Learnable quantum token mixing:** CLAQS parameterizes both the quantum mixing weights and nonlinear transformations as complex-valued parameters, enabling data-driven discovery of optimal token interactions instead of relying on analytically fixed coefficients.
- **Stabilized training via normalization:** An ℓ_1 normalization constraint on mixing amplitudes ensures stable end-to-end optimization and mitigates coefficient explosion or vanishing, which were major limitations in earlier quantum designs.
- **Dual-stage quantum circuit architecture:** A shared token-level embedding ansatz and a window-level quantum feed-forward block jointly decouple representation construction

[*]Equal contribution [†]Corresponding author

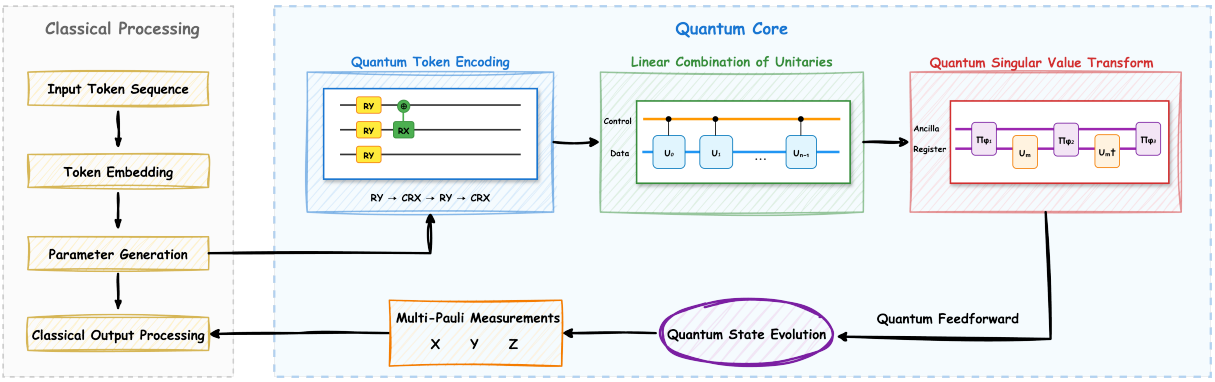

Figure 1: **Overall architecture of the CLAQS model.** The model integrates token-shared quantum embedding, LCU–QSVT mixing, and a quantum feed-forward block, mirroring the structure of a Transformer layer while operating on quantum representations.

and consumption—mirroring the structure of Transformer layers while maintaining qubit efficiency.

With only eight data qubits and shallow circuits, CLAQS achieves **91.64%** accuracy on SST-2 and **87.08%** on IMDB, outperforming both classical Transformers and strong hybrid baselines. To our knowledge, this is the first demonstration that a compact, learnable quantum circuit can act as an effective and stable sequence mixer for NLP, bridging the gap between classical attention and quantum superposition under tight resource budgets.

2 Related Work

Classical attention and alternative token mixers. The Transformer popularized self-attention and became the dominant sequence mixer in NLP (Vaswani et al., 2017). To reduce attention's quadratic cost, FNet mixes tokens with fixed Fourier transforms while remaining competitive on language tasks (Lee-Thorp et al., 2022). Beyond Fourier, recent *subquadratic* mixers—long-convolution operators and selective state-space models—have closed much of the quality gap to attention while improving scaling; representative examples include Hyena (implicit long convolutions with data-gated kernels) and Mamba (selective state spaces with input-dependent dynamics) (Poli et al., 2023; Gu and Dao, 2023). These efficiency trends motivate re-examining token mixing on alternative computational substrates, including quantum circuits.

Non-attention quantum text classifiers and encoders. Parallel to attention-based routes, a separate line studies end-to-end text classification without Transformer-style attention. Quantum-like semantic models have been proposed for sentiment and topic classification (Gruzdeva et al., 2025), while quantum n-gram language models target tweet categorization (Payares et al., 2023). On the representation side, task-oriented quantum text encoding, using amplitude-encoded features with QSVM back-ends, offers alternative front-ends for small-scale classification (Alexander and Widdows, 2022). These efforts contextualize our focus on a fully learnable LCU–QSVT token mixer that can be trained end-to-end and potentially plugged into larger classical pipelines.

Fully quantum NLP and quantum modules in classical models. Compositional QNLP (DisCoCirc) provides a circuit-level account of meaning (Coecke, 2021), with hardware demonstrations such as grammar-aware sentence classification and executions of compositional models on NISQ devices (Meichanetzidis et al., 2023; Lorenz et al., 2023). Closer to Transformers, quantum modules have been *inserted* into attention: QKSAN computes kernelized similarities on quantum circuits within self-attention (Zhao et al., 2024), and hybrid/quantized attention has been explored for molecular sequences (Smaldone et al., 2025). At the system level, dynamic hybrid co-design allocates quantum resources by input complexity (Pan et al., 2025), while Yang et al. (2022) attach quantum temporal convolutions to BERT—illustrating a pragmatic near-term path where small quantum sub-networks complement classical models.

Quantum-native attention mechanisms. Beyond replacing classical attention, several works implement token mixing directly in Hilbert space.

QSANN introduces Gaussian-projected quantum self-attention for text classification (Li et al., 2024). Quantum Vision Transformers define fully quantum attention layers (including compound-matrix attention) and argue asymptotic advantages in runtime and parameter count (Cherrat et al., 2024). QMSAN computes query–key similarities between mixed quantum states via SWAP-test-style overlaps (Chen et al., 2025). In this vein, *Quixer* prepares an LCU superposition over token states and applies a QSVT polynomial as a learned nonlinearity—effectively realizing attention in a quantum register (Khatri et al., 2024). Our **CLAQS** follows this quantum-native mixer direction for supervised classification and contributes a compact, fully learnable circuit: (i) both LCU mixing weights and the QSVT polynomial are trained end-to-end, (ii) LCU coefficients are ℓ_1-normalized to stabilize training, and (iii) a shared token-embedding ansatz is decoupled from a window-level quantum feed-forward module, yielding strong accuracy under an 8-data-qubit budget. [1]

3 Method

3.1 Problem Setup and Design Goals

We study supervised text classification with input token sequence $x = (w_0, \ldots, w_{n-1})$ and label $y \in \{1, \ldots, C\}$. Our goal is to replace dot-product attention with a *compact quantum mixer* that is trainable end-to-end, uses only q=8 data qubits, and preserves the principled *LCU* $\rightarrow$ *QSVT* $\rightarrow$ feed-forward $\rightarrow$ measurement pipeline introduced in Quixer while adapting it to classification tasks. Our design goals are (G1) compactness under strict qubit/depth budgets, (G2) phase-aware token interactions with explicit nonlinearity, (G3) stable optimization with regulated amplitudes, and (G4) scalability to long documents via sliding-window aggregation.

3.2 Model Overview

While maintaining the above pipeline, CLAQS reparameterizes all components to be trainable and differentiable, yielding a fully end-to-end learnable quantum circuit. The model proceeds in five stages. First, it maps each token embedding into a unitary circuit through a shared token-level ansatz. Second, it performs a complex-weighted unitary superposition with learnable parameters, normalized to ensure stability. Third, a trainable polynomial transformation acts as a quantum nonlinearity. Fourth, an independent feed-forward PQC refines the aggregated state at the window level. Finally, multi-axis measurements produce classical features for downstream classification through a lightweight MLP head.

[1] In classical simulation we evaluate $P_c(M)$ by repeated applications of M rather than materializing full block encodings, following Quixer's classical-simulation strategy and our implementation notes.

3.3 Unitary Token Embedding

Let $\mathbf{e}_w \in \mathbb{R}^{d_e}$ denote a classical token embedding. We map it to angles $\boldsymbol{\theta}_w = W_E\, \mathbf{e}_w$ (with Xavier initialization for W_E), which parameterize a q-qubit *ansatz-14* PQC $U(\boldsymbol{\theta}_w)$. As illustrated in Figure 2, a single *ansatz-14* layer alternates RY and controlled-RX gates and is repeated ℓ times; with q qubits this yields $4\ell q$ trainable angles while keeping depth modest. The circuit template is *shared* across tokens, where each token has its own $\boldsymbol{\theta}_w$ but the same gate pattern, improving sample efficiency and aligning with our unitary token-embedding design.

3.4 LCU Mixer with Learnable Complex Coefficients and L1 Normalization

For the window $\{U_j\}_{j=0}^{n-1}$ of token unitaries, we construct a linear combination of unitaries

$$M(\boldsymbol{b}) = \sum_{j=0}^{n-1} b_j\, U_j, \qquad \boldsymbol{b} \in \mathbb{C}^n, \tag{1}$$

where $\boldsymbol{b}$ is trainable and complex, representing the quantum magnitude and phase. To stabilize training and respect block-encoding assumptions, we apply L^1 normalization per forward pass:

$$\tilde{b}_j = \frac{b_j}{\sum_{k=0}^{n-1} |b_k|} \quad \Rightarrow \quad \sum_{j=0}^{n-1} |\tilde{b}_j| = 1. \tag{2}$$

On hardware, $M(\boldsymbol{b})$ is realized using a selection unitary U_{SEL} and a preparation unitary U_{PREP} with post-selection on the control register; in classical simulation we compute (1) directly, following Quixer's implementation strategy.

3.5 QSVT Nonlinearity with Learnable Polynomial

We apply Quantum Singular Value Transformation (QSVT) to the block-encoded $M(\tilde{\boldsymbol{b}})$, effecting a polynomial transform

$$P_{\boldsymbol{c}}(M) = \sum_{k=0}^{d} c_k\, M^k, \tag{3}$$

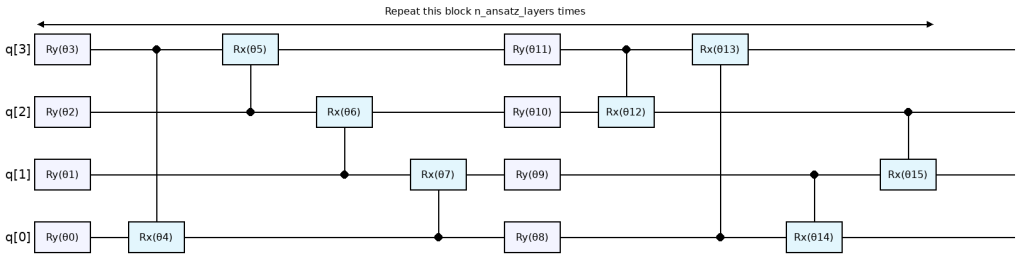

Figure 2: The detailed one layer of the shared *ansatz-14* for quantum token encoding (repeated ℓ times).

with learnable degree d and coefficients $\boldsymbol{c}$. Operationally, QSVT is implemented via a sequence of controlled phase rotations interleaved with U_M and $U_M^\dagger$; in classical simulation we evaluate the polynomial by repeated applications of M. Making $\boldsymbol{c}$ learnable allows the model to tailor its nonlinearity to data, rather than fixing $P_{\boldsymbol{c}}$ *a priori*. This follows the Quixer design where a QSVT polynomial acts as a nonlinearity on the LCU-prepared mix (Khatri et al., 2024).

3.6 Quantum Feed-Forward and Readout

After LCU and QSVT we apply a second, independent PQC $U_{\mathrm{FF}}(\boldsymbol{\phi})$ on the data register:

$$|\psi\rangle \;=\; U_{\mathrm{FF}}\, P_{\boldsymbol{c}}\big(M(\tilde{\boldsymbol{b}})\big)\; |0^q\rangle\,. \tag{4}$$

Unlike the embedding ansatz (shared per token), U_{FF} operates *once per window* on the mixed state, mirroring the role of position-wise feed-forward in Transformers but executed in quantum state space. Decoupling representation construction through the embedding ansatz from representation consumption through the quantum feed-forward increases modeling capacity without increasing token-wise parameters. This decoupling naturally leads to a clear transition from quantum feature transformation to classical readout and classification.

For output, we perform XYZ multi-axis readout. For each qubit, we measure the expectation values w.r.t. X, Y, and Z, yielding a feature vector $\mathbf{o} \in \mathbb{R}^{3q}$. A two-layer MLP with dropout maps $\mathbf{o}$ to logits $\mathbf{z} \in \mathbb{R}^C$. This readout-and-head design finally adapts the head to classification.

3.7 Objective, Regularization, and Optimization

We minimize the cross-entropy $\mathrm{CE}(z, y)$ over logits z and labels y, augmented with stability regularizers. Our objective is:

$$\mathcal{L} = \mathrm{CE}(z, y) + \mathrm{PSR} + \mathrm{L1C}. \tag{5}$$

The post-selection regularizer PSR steers the average success probability toward a target $\tau \in (0, 1)$:

$$\mathrm{PSR} = \lambda_{\mathrm{ps}}\Big(\overline{\big\|P_{\mathbf{c}}(M)|0^q\rangle\big\|^2} - \tau\Big)^2. \tag{6}$$

Here, $\overline{\|P_{\mathbf{c}}(M)|0^q\rangle\|^2}$ is the mean state norm (a proxy for post-selection success), M is a parameterized quantum circuit acting on the q-qubit zero state, and $P_{\mathbf{c}}$ is the projector onto the measurement outcome $\mathbf{c}$. The overline denotes an average (e.g., over minibatch examples and/or circuit shots), and the regularizer PSR nudges this mean success probability toward a target $\tau \in (0, 1)$.

The ℓ_1 constraint L1C softly encourages a unit-sum mixing vector $\tilde{\mathbf{b}}$:

$$\mathrm{L1C} = \lambda_{\ell_1}\big(\|\tilde{\mathbf{b}}\|_1 - 1\big)^2. \tag{7}$$

In practice we enforce ℓ_1 normalization in the forward pass (thus $\lambda_{\ell_1} = 0$) and use the mean state norm primarily for logging or light regularization. Unless otherwise specified, we train with AdamW and a cosine-annealed learning rate, following recent small-model quantum–classical NLP evaluation protocols.

3.8 Resource Model and Complexity

With window length n, QuixerClassifier uses q=8 data qubits, $\lceil \log_2 n \rceil$ control qubits, and a constant number of ancillae for QSVT and parity combination; total qubits scale as $\mathcal{O}(q + \log n)$. Let the embedding PQC have depth ℓ and the QSVT polynomial have degree d. Using standard controlled-gate constructions (or the ancilla-optimized variant), the gate complexity scales as

$$\mathcal{O}\big(d\, n\, q\, \ell\big), \tag{8}$$

matching Quixer's analysis and justifying our shallow-circuit, small-q design.

Compared with adaptive hybrids like AQCF that dynamically route between classical and quantum paths, our design fixes a compact quantum mixer and optimizes it end-to-end under the same small-model budget used for fair comparison on SST-2 and IMDB. The learnable LCU-QSVT mixer serves as an attention-like token mixer whose complex coefficients and polynomial parameters are directly tuned from data, yielding strong classification performance with only 8 qubits.

4 Experiments and Results

4.1 Tasks and datasets

We evaluate CLAQS on two standard sentiment classification benchmarks: SST-2, short sequences from the Stanford Sentiment Treebank (Socher et al., 2013) and IMDB, long reviews from Maas et al. (2011). We use the official train/validation/test splits and report Accuracy, Precision, Recall, and macro F1 on the test sets. These benchmarks are widely used to assess token-mixing mechanisms and attention alternatives in NLP. They further allow us to probe the sequence-length axis most relevant to token mixing.

RQ1. Model Effectiveness

Does the proposed **CLAQS** architecture outperform classical and hybrid baselines in text classification, given an equal or smaller parameter and qubit budget?

RQ2. Component Contribution

Which components of the **CLAQS** mixer (LCU, QSVT, and quantum feed-forward) most contribute to the observed performance gains, and how do aggregation or regularization schemes influence these results?

4.2 Implementation Details

We implement all quantum components in **TorchQuantum**, explicitly injecting rotation angles and realizing **LCU** and **QSVT** as dedicated differentiable layers. For classical simulation, we follow **Quixer**'s efficient strategy: repeatedly evaluating M and applying $P_{\boldsymbol{c}}$ through compositional updates rather than materializing full block-encoding circuits. Parameter initialization employs **Xavier** for W_E, small uniform noise for PQC angles, and near-isotropic initialization for $\boldsymbol{b}$ and $\boldsymbol{c}$, followed by end-to-end optimization. During the forward pass, $\boldsymbol{b}$ is batch-broadcast and $\mathbf{L}^1$**-normalized** to prevent coefficient explosion or collapse, ensuring stable success probabilities.

All models share the same tokenization and official train/validation/test splits as their respective datasets. The **CLAQS** mixer (LCU + QSVT) is optimized jointly with the two-stage PQC and the lightweight MLP head. We adopt the small-model evaluation protocol commonly used in recent quantum–classical NLP research, and conduct all experiments on an NVIDIA RTX A5000 (24 GB) GPU.

4.3 Main results

Table 1 summarizes the results on SST-2 and IMDB. On **SST-2**, **CLAQS** reaches **91.64%** accuracy (F1 = 0.915), outperforming both the *Classical Transformer* (79.36%) and the strongest hybrid baseline *AQCF* (81.88%) by **+12.28** and **+9.76** points, respectively. On **IMDB**, CLAQS achieves **87.08%** accuracy, surpassing *AQCF* (86.30%, +0.78), *Quantum Transformer* (82.30%, +4.78), and the *Classical Transformer* (82.18%, +4.90). Fully quantum methods underperform on both datasets, highlighting the need for principled quantum–classical integration. These gains are particularly notable given our strict **8-qubit** budget and shallow circuits, aligning with the resource-conscious design advocated in prior work (Khatri et al., 2024; Pan et al., 2025).

4.4 Ablation studies

Ablation 1: Aggregation mechanism We compare the CLAQS mixer (LCU+QSVT+U_{FF}) against two purely classical window aggregators that keep all other hyperparameters fixed: *mean_logits* and *attention pooling*. On **SST-2**, CLAQS improves over *mean_logits* by **+12.74** points (91.64 vs. 78.90) and over *attention pooling* by **+11.82** points (79.82). On **IMDB**, CLAQS also outperforms both baselines by **+2.69** (vs. mean_logits = 84.39) and **+3.45** (vs. attention = 83.63) points. This indicates that the *quantum* mixing induced by LCU (complex, normalized coefficients) plus QSVT (learnable polynomial nonlinearity) provides consistent benefits over classical pooling, with larger margins on the shorter sequence regime (SST-2).

Ablation 2: Measurement mask and quantum-side regularization We investigate two factors: (i) a measurement *mask* on the XYZ readout, and

Method	Qubits	SST-2 Dataset				IMDB Dataset			
		Acc. (%)	**Precision**	**Recall**	**F1-Score**	**Acc.** (%)	**Precision**	**Recall**	**F1-Score**
Quantum Methods									
DisCoCirC	20	53.44	0.536	0.536	0.534	50.00	0.750	0.500	0.333
VQC Classifier	20	52.29	0.262	0.500	0.343	54.36	0.544	0.544	0.543
Classical Baseline									
Classical Transformer	–	79.36	0.800	0.797	0.793	82.18	0.822	0.822	0.822
Hybrid Methods									
Hybrid Quantum CNN	20	72.02	0.729	0.715	0.714	77.91	0.781	0.779	0.779
Quantum Transformer	20	79.59	0.796	0.797	0.796	82.30	0.829	0.823	0.822
AQCF	20	81.88	0.819	0.818	0.818	86.30	0.863	0.863	0.863
Proposed Method									
CLAQS (Ours)	8	**91.64**	**0.915**	**0.915**	**0.915**	**87.08**	**0.871**	**0.871**	**0.871**

Table 1: Performance comparison on SST-2 and IMDB sentiment analysis benchmarks. "–" indicates that the reported qubit count is unspecified.

Model / Variant	SST-2				IMDB			
	Acc (%)	Prec.	Rec.	F1	Acc (%)	Prec.	Rec.	F1
Full CLAQS (Ours)	**91.64**	0.915	0.915	0.915	**87.08**	0.871	0.871	0.871
Ablation Study 1: Aggregation Mechanism								
mean_logits	78.90	0.790	0.789	0.789	84.39	0.844	0.844	0.844
attention pooling	79.82	0.798	0.798	0.798	83.63	0.836	0.836	0.836
Ablation Study 2: Mask and Quantum Regularization								
mean_logits	80.39	0.804	0.804	0.804	85.33	0.853	0.853	0.853
+mask +LCU/QSVT regs	70.24	0.702	0.702	0.702	85.83	0.858	0.858	0.858

Table 2: **Ablation Results.** Top rows report full CLAQS performance. Middle block (*Ablation Study 1*) compares aggregation mechanisms (mean_logits vs. attention) under identical hyperparameters. Bottom block (*Ablation Study 2*) examines effects of measurement masking and quantum-side regularization (LCU L^1, QSVT smoothness L^2 diff, overall QSVT L^2 penalties).

(ii) quantum-side *regularization* (LCU L^1, QSVT smoothness via L^2 on coefficient differences, and an overall L^2 penalty on the polynomial). On **SST-2**, the masked+regularized variant drops to 70.24% (vs. 91.64%), suggesting that strong constraints can *underfit* small datasets where flexible token interactions are beneficial. On **IMDB**, the same variant slightly exceeds the classical *mean_logits* baseline (85.83 vs. 85.33) but remains below full CLAQS (87.08). Overall, aggressive smoothing and masking can curb expressivity on short inputs, while mild smoothing may be tolerable or modestly helpful on longer reviews; the full CLAQS remains the most effective across both settings.

4.5 Parameter Budget

Table 3 compares trainable attention parameters excluding embeddings and qubit usage across models. A single 256-d, 8-head Transformer attention block contains 263,168 parameters. By contrast, our CLAQS attention uses only 454 parameters on IMDB and 326 on SST-2, which is, roughly $580\times$ and $807\times$ fewer than the Transformer block, respectively. Compared with AQCF (**8,268** parameters), CLAQS is about $18\times$ (IMDB) to $25\times$ (SST-2) smaller.

Transformer attention parameters grow quadratically with width and linearly with the number of layers. In contrast, CLAQS is independent of d_{model} and the number of heads; its attention budget scales linearly with the window length n and the polynomial degree d, plus a small constant from the quantum feed-forward:

$$P_{\text{attn}}^{\text{CLAQS}} = O(n) + O(d) + O(1).$$

Empirically, halving the window from n=256 to n=128 reduces the parameter count by exactly 128 ($454 \rightarrow 326$), matching the linear dependence on n.

Method	Configuration	Attention Params	Qubits
Classic Transformer	d_{model}=256, heads=8	263,168	–
AQCF (default)	embed_dim=128, L=2	8,268	20
Ours (IMDB)	window=256, degree=5	**454**	**8**
Ours (SST-2)	window=128, degree=5	**326**	**8**

Table 3: **Comparison of trainable attention parameters and qubit usage.** For **Ours**, the parameter count excludes token embedding parameters and includes only LCU coefficients ($2n$ for window length n), QSVT polynomial coefficients (d+1=6 for degree=5), and quantum feed-forward parameters. For **AQCF**, 8,268 already aggregates all trainable attention parameters under the stated defaults, where L is the number of transformer layers.

4.6 Discussion

Where do the gains come from? The main improvements arise from treating the attention-like mixer as a learnable quantum operator: (i) complex LCU coefficients (with L^1 normalization) provide phase-sensitive weighted mixing, and (ii) QSVT supplies a learnable polynomial nonlinearity that composes token unitaries into higher-order interactions, before a *window-level* quantum feed-forward refines the state. This design follows the block-encoding and QSVT principles of Quixer but exposes all mixer parameters to gradient-based learning and adapts the head to classification.

Compared to AQCF, which adaptively routes between classical and quantum paths, CLAQS fixes a compact quantum mixer and focuses the learning capacity on its complex coefficients and polynomial, yielding stronger results under the same small-model budget.

Resource considerations. All results are obtained with q=8 data qubits; the control register scales with $\lceil \log_2 n \rceil$, and a small number of ancillae support QSVT and parity combination. This adheres to the $\mathcal{O}(q+\log n)$ qubit count and $\mathcal{O}(d\,n\,q\,\ell)$ gate complexity characterized for LCU+QSVT mixers. We emphasize that our ablations keep the qubit budget fixed to isolate the contribution of the mixer itself.

With a strict 8-qubit budget and shallow depth, CLAQS delivers sizable gains over classical pooling and strong hybrid baselines on SST-2, and consistent improvements on IMDB. These results support a practical view of quantum components in NLP: compact qubit-based mixers can act as effective, learnable alternatives to classical attention within modern pipelines. Moreover, a limited number of qubits can effectively represent the input information, as a single quantum state inherently encodes multiple interdependent variables. In essence, a small set of qubits can span an informational space (Yosida, 2012; Royden and Fitzpatrick, 1988) equivalent to that of classical representations within the same dimensional framework, highlighting the expressive efficiency of quantum encoding.

5 Conclusion

We introduced CLAQS, a Compact Learnable All-Quantum Token Mixer with Shared-ansatz, demonstrating that a compact quantum module can effectively serve as an attention-like token mixer for text classification. Architecturally, CLAQS exposes both the LCU mixing weights and the QSVT polynomial as learnable complex parameters. It separates representation construction from consumption through a two-stage PQC design with a shared ansatz-14 embedding stage followed by a window-level quantum feed-forward stage. Finally, it applies L^1 normalization on the LCU coefficients to stabilize training and prevent parameter divergence. With only q=8 data qubits and shallow circuits, CLAQS attains **91.64%** accuracy on SST-2 and **87.08%** on IMDB, outperforming a classical Transformer and strong hybrid baselines (including AQCF), while remaining within realistic NISQ-style resource budgets.

Our analyses support three takeaways. First, making the LCU and QSVT learnable is key since phase-aware LCU mixing combined with a polynomial nonlinearity consistently outperforms classical window aggregation such as mean or attention pooling. Second, the separation between the token-shared embedding ansatz and the window-level quantum feed-forward improves adaptation to the task without increasing token-wise parameters. Third, the gains persist under a strict 8-qubit budget, indicating that small quantum mixers can be practical components in modern NLP pipelines.

Future directions. In future work, we will validate CLAQS on quantum hardware with shot-based readout under realistic noise. We will also broaden the empirical scope beyond sentiment analysis to natural language inference, topic classification, and token-level tasks, and explore scaling via multi-head and multi-layer variants alongside parameter-efficient training. In parallel, we will examine constrained and Chebyshev-based QSVT polynomials to understand their effects on optimization stability and calibration. Finally, we aim to develop a theoretical account of post-selection success and gradient scaling in compact QSVT–LCU mixers. More broadly, we view CLAQS as a step toward quantum–classical co-design, where compact qubit mixers can serve as learnable alternatives to attention under tight resource constraints.

References

Aaranya Alexander and Dominic Widdows. 2022. Quantum text encoding for classification tasks. In *2022 IEEE/ACM 7th Symposium on Edge Computing (SEC)*, pages 355–361. IEEE.

Fu Chen, Qinglin Zhao, Li Feng, Chuangtao Chen, Yangbin Lin, and Jianhong Lin. 2025. Quantum mixed-state self-attention network. *Neural Networks*, 185:107123.

El Amine Cherrat, Iordanis Kerenidis, Natansh Mathur, Jonas Landman, Martin Strahm, and Yun Yvonna Li. 2024. Quantum vision transformers. *Quantum*, 8(arXiv: 2209.08167):1265.

Bob Coecke. 2021. The mathematics of text structure. In *Joachim Lambek: The Interplay of Mathematics, Logic, and Linguistics*, pages 181–217. Springer.

Anastasia S Gruzdeva, Rodion N Iurev, Igor A Bessmertny, Andrei Y Khrennikov, and Alexander P Alodjants. 2025. A quantum-like approach to semantic text classification. *Entropy*, 27(7):767.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Nikhil Khatri, Gabriel Matos, Luuk Coopmans, and Stephen Clark. 2024. Quixer: A quantum transformer model. *arXiv preprint arXiv:2406.04305*.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2022. FNet: Mixing tokens with Fourier transforms. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4296–4313, Seattle, United States. Association for Computational Linguistics.

Guangxi Li, Xuanqiang Zhao, and Xin Wang. 2024. Quantum self-attention neural networks for text classification. *Science China Information Sciences*, 67(4):142501.

Robin Lorenz, Anna Pearson, Konstantinos Meichanetzidis, Dimitri Kartsaklis, and Bob Coecke. 2023. Qnlp in practice: Running compositional models of meaning on a quantum computer. *Journal of Artificial Intelligence Research*, 76:1305–1342.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Konstantinos Meichanetzidis, Alexis Toumi, Giovanni de Felice, and Bob Coecke. 2023. Grammar-aware sentence classification on quantum computers. *Quantum Machine Intelligence*, 5(1):10.

Yi Pan, Hanqi Jiang, Junhao Chen, Yiwei Li, Huaqin Zhao, Lin Zhao, Yohannes Abate, Yingfeng Wang, and Tianming Liu. 2025. Bridging classical and quantum computing for next-generation language models. *arXiv preprint arXiv:2508.07026*.

Esteban Payares, Edwin Puertas, and Juan C Martinez-Santos. 2023. Quantum n-gram language models for tweet classification. In *2023 IEEE 5th International Conference on Cognitive Machine Intelligence (CogMI)*, pages 69–74. IEEE.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. 2023. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, pages 28043–28078. PMLR.

Halsey Lawrence Royden and Patrick Fitzpatrick. 1988. *Real analysis*, volume 32. Macmillan New York.

Anthony M Smaldone, Yu Shee, Gregory W Kyro, Marwa H Farag, Zohim Chandani, Elica Kyoseva, and Victor S Batista. 2025. A hybrid transformer architecture with a quantized self-attention mechanism applied to molecular generation. *Journal of Chemical Theory and Computation*, 21(10):5143–5154.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Chao-Han Huck Yang, Jun Qi, Samuel Yen-Chi Chen, Yu Tsao, and Pin-Yu Chen. 2022. When bert meets quantum temporal convolution learning for text classification in heterogeneous computing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8602–8606. IEEE.

Kôsaku Yosida. 2012. *Functional analysis*, volume 123. Springer Science & Business Media.

Ren-Xin Zhao, Jinjing Shi, and Xuelong Li. 2024. Qksan: A quantum kernel self-attention network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.