When Quantum Entanglement Meets Transformer Attention: A Controlled Study of Quantum Attention Heads

Mohd Farhan Israk Soumik*, Hussein Zangoti†, Arun Babu Aryal*, Ahmed Imteaj‡, Abdur R. Shahid*
*School of Computing, Southern Illinois University Carbondale, IL, USA
†College of Engineering and Computer Science, Jazan University, Saudi Arabia
‡Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL, USA
*{mohdfarhanisrak.soumik, arunbabu.aryal}@siu.edu, †hmzangoti@jazanu.edu.sa, ‡aimteaj@fau.edu, *shahid@cs.siu.edu

***Abstract*—Quantum attention mechanisms are widely proposed as next-generation primitives for augmenting the capabilities of transformers but it comes at a cost. Existing research works and designs typically replace entire modules at once,conflating the geometric properties of quantum entanglement with structural shifts in model depth, parameter density, and architectural nonlinearity. This paper isolates query-key entanglement as a strict independent variable. By replacing only the scoring function of a single attention head with a parameterized quantum circuit, we keep the surrounding transformer framework entirely classical and identical across conditions. This experimental design is anchored by two formal structural propositions: the first establishes the algebraic boundaries where entangling layers remain invisible to a Z-basis readout, and the second proves that only cross-register gates can generate a non-separable query-key score. Evaluating five distinct topological conditions on TREC-6 and AG News across multiple structural controls, we find that no non-separable quantum condition yields a statistically significant performance gain over the classical baseline. However, the quantum scorer does induce a measurable increase in attention entropy. We demonstrate that this shift is a purely statistical byproduct of the nonlinear data encoding shared by all quantum variants, rather than the entangling topology that distinguishes them. Ultimately, these findings reveal that while query-key entanglement can be formally verified inside the circuit, it remains invisible to text classification, implying that nonlinearity fundamentally alters the attention landscape, but changes due to entanglement bears minimal significance.**

***Index Terms*—Quantum Machine Learning, Parameterized Quantum Circuits, Attention Mechanism, Entanglement, Text Classification**

I. Introduction

Quantum computing promises a model of computation that departs fundamentally from the classical paradigm. A register of n qubits spans a 2^n-dimensional Hilbert space, and through entanglement a quantum state can encode correlations among its subsystems that no classical product representation can express compactly [1]. This exponential state space, together with the non-classical correlations entanglement provides, has motivated a sustained effort to determine whether quantum systems can offer computational advantages for machine learning [2]. The most practical embodiment of this effort on near-term hardware is the parameterized quantum circuit, a sequence of tunable gates whose parameters are optimized by gradient descent in direct analogy to a neural network [3], [4]. Such circuits can be embedded as differentiable components inside otherwise classical models. This in turn raises a concrete and testable question regarding whether the distinct inductive bias of a quantum circuit, the very joint distributions over measurement outcomes that classical arithmetic cannot reproduce efficiently, translate into any measurable benefit on tasks that classical deep learning already handles well. Researchers have yet to come out with an obvious answer to this question due to the fact that training these circuits is itself delicate because of the arising of vanishing gradients with growing the number of qubits [5]. As a result, any benefit must be weighed against this cost. Such a question is,therefore,best approached empirically rather than in principle, by isolating a single, well-understood component where a quantum circuit could plausibly help and testing whether it in fact does

Natural language processing represents a compelling setting to explore this question, and not merely because it is difficult but the argument is more specific. Attention mechanisms now underlie most state-of-the-art NLP models [6], and at their core each attention head reduces to an inner product between a query and a key vector that determines how strongly one token attends to another. Analyses of trained Transformers have shown that individual heads specialize, with some capturing interpretable syntactic and positional relations while others appear redundant and can be pruned without harming performance [7], [8]. The scoring function of a single head is thus both a small, well-isolated component and one whose behaviour is known to vary, which makes it an ideal site to introduce and test an alternative primitive. Quantum measurement by characteristics, is itself natively an inner product computation [9]. This implies that a quantum circuit, acting on a query-key pair offers a principled generalization of the attention score, replacing the dot product with a richer function evaluated in an exponentially large Hilbert space. Apart from that, language also possesses compositional structure, with words combining into phrases and phrases into sentences through rules that mirror the tensor-product composition native to quantum mechanics, a correspondence developed formally in the DisCoCat framework [10]. These arguments, structural

alignment on one hand and the inner-product form of attention on the other, have made quantum models an attractive candidate for linguistic computation. Motivated by this alignment, a growing body of work has embedded quantum components into NLP architectures [11]. A recent survey of this literature observes that most QNLP models remain confined to small datasets and that only a handful of architectures have been explored in any depth, so the relative contribution of their individual quantum ingredients remains largely unexplored. Proposals span compositional models that compile grammatical structure directly into quantum circuits [10], quantum recurrent networks in which classical gate operations are replaced by parameterized circuits [12], and quantum attention mechanisms in which query, key, and value representations are derived from quantum measurements rather than classical projections [13]–[15]. Across this range of approaches, quantum components have been shown to train stably within NLP pipelines and to reach or approach the accuracy of their classical counterparts on standard benchmarks.

A limitation shared across these approaches is the absence of any experimental control over what makes them quantum, making it difficult to attribute any observed difference to the quantum part specifically. When an entire attention block or recurrent cell is replaced, the resulting model differs in depth, parameter count, and internal structure alongside its computational primitive, so observed gains or losses conflate many variables. An observed difference in performance is therefore uninterpretable which means it cannot be attributed to entanglement, to the changed architecture, or to any other single factor. Entanglement in particular has never been isolated as the variable under test. As a result, whether the correlations that quantum circuits can represent are actually useful for a given task, and whether a circuit that entangles query and key registers produces a meaningfully different attention score than one that does not, remains an open and largely unexamined question. Without controlling for this, it is impossible to say whether quantum NLP models succeed or fail because of their quantum character or in spite of it.

This paper addresses that gap directly by replacing only the dot-product scoring function of a single attention head with a parameterized quantum circuit, while every other component, including the remaining heads, both transformer blocks, and the feed-forward networks, stays classical and unchanged. Entanglement is then varied as an explicit control by changing only which two-qubit CZ gates are applied, making it the sole independent variable across all comparisons. The findings show that entangling and non-entangling circuits achieve statistically indistinguishable accuracy on text classification while differing consistently in attention entropy. Two structural results explain why: a commutation argument shows that a naively placed entangling layer is invisible to the readout, and a separability analysis proves that only cross-register gates yield a non-separable score. A placement control that moves the quantum head to a deeper layer, where it operates on contextualised representations rather than raw embeddings, confirms that the null persists regardless of input quality, framing the result as a property of text classification rather than an artifact of circuit position.

II. Related Work

A. Parameterized Quantum Circuits for Machine Learning

The use of parameterized quantum circuits as trainable machine learning components was established through several parallel developments. Biamonte et al. [2] surveyed the landscape of quantum machine learning and identified classification, generative modelling, and reinforcement learning as areas where quantum systems might offer structural advantages. Schuld and Killoran [4] formalised the connection between quantum circuits and kernel methods, showing that a PQC implicitly defines a feature map into an exponentially large Hilbert space where the kernel trick can be applied. Benedetti et al. [16] studied PQCs specifically as supervised learning models, establishing training protocols and gradient estimation strategies that made end-to-end optimisation practical on near-term hardware. The variational quantum algorithm framework [3] unifies these efforts under a common structure in which a classical optimiser iteratively adjusts circuit parameters to minimise a cost function evaluated via quantum measurement.

Two properties of PQCs have received particular attention. Sim et al. [17] introduced quantitative measures of *expressibility* and *entangling capability*, showing that circuit structure and qubit connectivity determine how much of the Hilbert space the circuit can access. Their analysis demonstrated that all-to-all two-qubit connectivity substantially outperforms linear connectivity, a finding relevant to any hybrid design that places a PQC inside a larger model. Separately, McClean et al. [5] showed that randomly initialised circuits with sufficient depth exhibit barren plateaus in which gradients vanish exponentially with qubit count, posing a fundamental trainability challenge. Subsequent work refined this picture by linking trainability directly to entanglement: Ortiz Marrero et al. [18] proved that circuits producing volume-law entanglement between subsystems concentrate toward maximally mixed states, inducing barren plateaus and suggesting that more entanglement is not always beneficial. Together, these results categorically define the design space within which quantum components must operate: the circuit must be expressive enough to represent useful functions while remaining shallow and modestly entangled enough to train.

A related body of work concerns how classical data enters the circuit. The choice of feature map, the circuit that encodes a classical vector into a quantum state, strongly determines what a quantum model can learn. Havlíček et al. [19] demonstrated supervised classification in quantum-enhanced feature spaces, encoding data through a fixed feature map and learning a decision boundary in the induced Hilbert space. Lloyd et al. [20] argued that the embedding itself should be trained to separate data classes, framing the encoding as a metric-learning problem. Schuld et al. [21] showed that the data-encoding strategy bounds the set of functions a variational model can represent, establishing that encoding is not a neutral

preprocessing step but a determinant of model expressivity. These findings are directly relevant to attention scoring, where the query and key vectors must be encoded into the circuit before any entangling operation can act on them.

B. Quantum Approaches to Natural Language Processing

Quantum models of language have been explored along two broad lines. The first is grounded in compositional semantics. Coecke et al. [22] proposed the DisCoCat framework, which represents word meanings as tensors and derives sentence meaning through compositional contraction rules that map directly onto the tensor-product structure of quantum mechanics. Lorenz et al. [10] implemented this framework on quantum hardware, compiling grammatical string diagrams into parameterized circuits and demonstrating small-scale sentence classification on real devices.

The second line embeds quantum components into established neural architectures. Chen et al. [12] replaced the internal gate operations of an LSTM with parameterized quantum circuits, introducing a quantum recurrent model that was trained stably on sequence tasks while reducing the overall parameter count. In the attention domain, Li et al. [13] proposed the Quantum Self-Attention Neural Network (QSANN), which derives query and key representations from quantum circuit measurements and computes Gaussian-projected self-attention coefficients. Zhao et al. [14] introduced an alternative quantum self-attention mechanism (QSAN) based on quantum logical similarity, replacing the classical inner-product similarity with logical gate operations to enable fully quantum processing of the attention computation. Beyond text, Cherrat et al. [23] extended the quantum transformer concept to vision, designing quantum attention layers for image classification and analysing their theoretical expressiveness.

Across these attention related body of works, reported accuracies are high but are obtained under heterogeneous conditions, for example, QSANN [13], QMSAN [24], QAKSAN [25] Quixer [26] and QSAN [14] each replace the full attention mechanism, use different base architectures and embedding strategies, and evaluate on different benchmarks. Direct numerical comparison with the present results is therefore not meaningful, since any accuracy difference would conflate the quantum primitive with every other architectural variable, the very confound this paper is designed to eliminate. The relevant comparison here is internal, across conditions that share every component except the entangling topology.

C. Attention Head Analysis

A distinct line of classical NLP research has characterised what individual attention heads learn and how much each contributes. Voita et al. [7] showed that most heads in a trained Transformer can be pruned with little accuracy loss, while a small minority capture interpretable syntactic and positional relations. Clark et al. [8] analysed BERT's attention patterns and found that specific heads specialise in dependency relations, coreference, and other linguistic functions. These findings establish that the scoring function of a single head is a small, well-isolated, and behaviourally-varying component suitable for quantum circuit application without disruptions. Our work targets exactly this site, adopting the PQC framework of variational quantum algorithms and addressing the methodological gap left by prior quantum NLP work by making entanglement the sole independent variable.

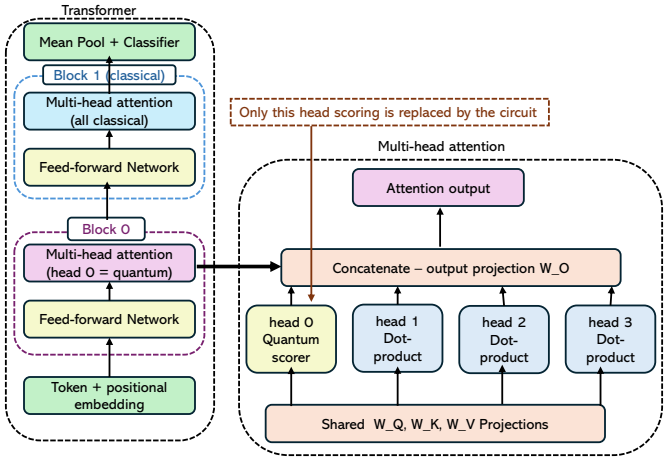


Fig. 1. The model is a standard two-block Transformer in which only the scoring function of head 0 in block 0 is replaced by a parameterized quantum circuit (Fig. 2). All other heads, blocks, and feed-forward networks remain classical, isolating entanglement as the sole variable under test.

III. Methodology

A. Classical Attention Baseline

In standard multi-head attention [6], a sequence $X \in \mathbb{R}^{L\times d}$ is projected into H heads of dimension $d_h = d/H$, and each head produces

$$\mathrm{Attn}(Q_m, K_m, V_m) = \mathrm{softmax}\left(\frac{Q_m K_m^\top}{\sqrt{d_h}}\right) V_m. \qquad (1)$$

The pairwise score $s_{ij} = q_i^\top k_j/\sqrt{d_h}$ is a bilinear function of q_i and k_j. The baseline model is a two-block Transformer (d=64, H=4, d_h=16, two-layer FFN) with learned token and positional embeddings, masked mean pooling, and a linear classifier, trained from scratch without pretrained embeddings. The architecture is deliberately small: rather than maximising accuracy, the aim is a setting in which a single head's scoring function constitutes a non-trivial fraction of the model's capacity, so that the effect of replacing it is not diluted beyond detection.

B. Quantum-Scored Attention Head

The scoring function of head 0 is replaced by a parameterized quantum circuit (PQC) in exactly one block at a time, either block 0 in the primary configuration or block 1 in the placement control, while the remaining three heads in that modified block, the other block in its entirety, the feed-forward networks, and all projection matrices are left classical and identical across all conditions, with a clear motivation to isolate the effect of quantum circuit, so that any observed difference is clearly attributable to the circuit parameters and the entanglement topology they operate under, not to any other architectural change.

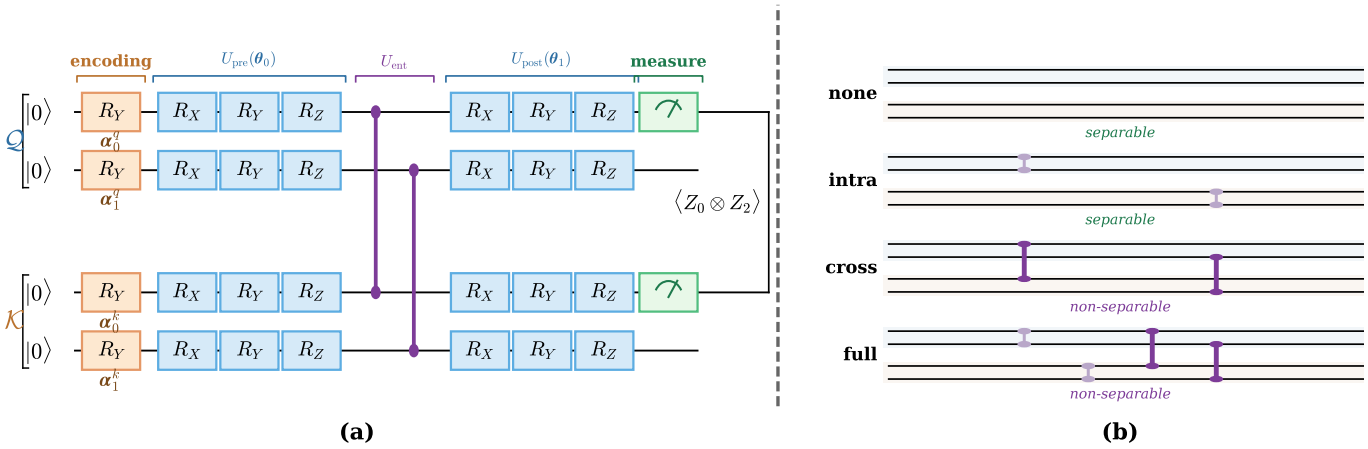


Fig. 2. (a) The quantum-scored attention head. The query and key vectors are encoded as R_Y rotations, transformed by parameterized layers U_{pre} and U_{post} with an intervening entangling layer U_{ent} of CZ gates, and read out as $\langle Z_0 \otimes Z_2 \rangle$. (b) The four entanglement modes, which differ only in the CZ gates applied within U_{ent}. Cross-register gates (dark) couple the query and key registers and make the score non-separable; intra-register gates (light) act within a single register and leave the score separable.

Encoding.: The d_h-dimensional query and key vectors are projected to $h = n/2$ angles each by learned linear maps, since quantum rotation gates accept scalar angles rather than high-dimensional vectors:

$$\boldsymbol{\alpha}_i^q = \pi \tanh(W_q q_i) \in \mathbb{R}^h, \quad \boldsymbol{\alpha}_j^k = \pi \tanh(W_k k_j) \in \mathbb{R}^h, \quad (2)$$

where tanh keeps every angle in $(-\pi, \pi)$. Query angles are loaded onto wires $\mathcal{Q} = \{0, \ldots, h-1\}$ and key angles onto $\mathcal{K} = \{h, \ldots, n-1\}$ via Y-rotation angle embedding:

$$|\phi_0\rangle = \bigotimes_{w=0}^{h-1} R_Y\left(\alpha_{i,w}^q\right) \bigotimes_{w=0}^{h-1} R_Y\left(\alpha_{j,w}^k\right) |0\rangle^{\otimes n}. \quad (3)$$

The state $|\phi_0\rangle$ is a product across $\mathcal{Q}|\mathcal{K}$; entanglement is introduced by the subsequent circuit.

Variational circuit.: A pre-entangler rotation layer $U_{\text{pre}}(\boldsymbol{\theta}_0)$, an entangling layer U_{ent} of CZ gates, and a post-entangler rotation layer $U_{\text{post}}(\boldsymbol{\theta}_1)$ are applied in sequence:

$$|\psi\rangle = U_{\text{post}}(\boldsymbol{\theta}_1)\, U_{\text{ent}}\, U_{\text{pre}}(\boldsymbol{\theta}_0)\, |\phi_0\rangle, \quad (4)$$

where each rotation layer applies $R_X R_Y R_Z$ per wire. The post-entangler layer is not optional: its necessity for making the entangler visible to the readout is established in Section III-C. The $6n$ rotation angles (24 for n=4), together with the projection matrices W_q, W_k and the temperature γ, are trained end-to-end with the rest of the model.

Score and attention.: The score is the expectation of a two-qubit Z observable, one qubit from each register:

$$s_{ij} = \gamma\, \langle \psi \mid Z_0 \otimes Z_h \mid \psi \rangle, \quad (5)$$

with learnable temperature γ. Because $s_{ij} \in (-\gamma, \gamma)$ is bounded, unlike the dot product which is unbounded, scores cannot spread as far apart, so the softmax peaks less sharply and produces a structurally more diffuse attention distribution. The head output is $o_i = \sum_j \text{softmax}_j(s_{ij})\, v_j$ with classical values v_j.

Whether the entangling layer in (4) actually affects the score (5) is not obvious—and, as shown next, requires care.

C. The CZ Visibility Problem

A CZ gate is diagonal in the computational basis, $\text{CZ} = \text{diag}(1, 1, 1, -1)$, as is the observable $Z_0 \otimes Z_h$. Diagonal matrices commute, so

$$[U_{\text{ent}},\ Z_0 \otimes Z_h] = 0 \quad (6)$$

for any product of CZ gates.

Proposition 1 (Entangler invisibility)**.** *If $U_{\text{post}} = \mathbb{I}$, the score* (5) *is independent of U_{ent} for all inputs and parameters.*

Proof. Let $|\phi\rangle = U_{\text{pre}}(\boldsymbol{\theta}_0)|\phi_0\rangle$. Without U_{post}, the score is $\langle\phi|\, U_{\text{ent}}^\dagger (Z_0 \otimes Z_h)\, U_{\text{ent}}\, |\phi\rangle$. By (6) this equals $\langle\phi|\, Z_0 \otimes Z_h\, |\phi\rangle$, independent of U_{ent}. □

In other words, CZ gates placed directly before a Z-basis readout have no effect on the output. The post-entangler layer U_{post} contains non-diagonal R_X, R_Y rotations that break this commutativity, making the entanglement detectable. This is verified numerically: over 300 random $(q, k, \boldsymbol{\theta})$ samples, removing U_{post} brings the mean absolute score difference between entangled and non-entangled circuits to exactly 0.000000; restoring it raises it to 0.399 (cross) and 0.297 (full).

D. Separability Analysis and Four-Mode Ablation

A naive comparison between a quantum model and the classical baseline conflates entanglement with all other differences between the scoring functions. To isolate the contribution of query–key entanglement specifically, a separability analysis is needed.

Definition 1 (Separable score)**.** *$s(q, k)$ is separable if $s(q, k) = f(q)\, g(k)$, or equivalently $\langle Z_0 \otimes Z_h \rangle = \langle Z_0 \rangle \langle Z_h \rangle$.*

TABLE I
ENTANGLEMENT MODES (n=4, h=2, $\mathcal{Q}$=$\{0,1\}$, $\mathcal{K}$=$\{2,3\}$). SEPARABILITY GAP IS $|\langle -|Z_0 Z_h|-\rangle \langle Z_0\rangle \langle Z_h\rangle|$, AVERAGED OVER 200 RANDOM SAMPLES.

Mode	CZ gates applied	Score type	Sep. gap
none	—	separable	0.000
intra	$(0,1),\ (2,3)$	separable	0.000
cross	$(0,2),\ (1,3)$	non-separable	0.261
full	$(0,1),(0,2),(1,3),(2,3)$	non-separable	0.109

Proposition 2 (Separability under intra-register entanglement)**.** *In modes* none *and* intra*, the score* (5) *is separable for all inputs and parameters. In modes* cross *and* full *it is not.*

Proof. In modes none and intra, every CZ gate acts within a single register, so the full unitary factors as $U = U_{\mathcal{Q}} \otimes U_{\mathcal{K}}$ and the output state is a product, $|\psi\rangle = |\psi_{\mathcal{Q}}\rangle \otimes |\psi_{\mathcal{K}}\rangle$. Since Z_0 acts only on $\mathcal{Q}$ and Z_h only on $\mathcal{K}$,

$$\langle Z_0 \otimes Z_h\rangle = \langle\psi_{\mathcal{Q}}|Z_0|\psi_{\mathcal{Q}}\rangle\langle\psi_{\mathcal{K}}|Z_h|\psi_{\mathcal{K}}\rangle = \langle Z_0\rangle \langle Z_h\rangle\,. \quad (7)$$

In modes cross and full, CZ gates couple $\mathcal{Q}$ and $\mathcal{K}$; the circuit no longer factors and (7) fails in general. □

Intra-register CZ gates entangle qubits within a register but leave the query–key interaction unchanged. Only cross-register gates create genuine query–key entanglement, so any accuracy difference between cross/full and none/intra is attributable specifically to that entanglement. The four modes and their numerically verified separability gaps are summarised in Table I.

IV. EXPERIMENTAL SETUP

A. Datasets

For this study, two text classification benchmarks are chosen that together span a range of input lengths and label structures while remaining small enough for quantum circuit simulation. TREC-6 [27] is a question-type classification dataset with relatively short text size (median 10 tokens) and six fine-grained classes(ABBR, ENTY, DESC, HUM, LOC and NUM) making it a natural fit for probing attention behaviour on minimal sequences where each token carries high informational weight. On the other hand, AG News [28] is a dataset focused on topic classification task with relatively longer passages (median 38 tokens) and four coarse labels(World, Sports, Business and Sci/Tech), testing whether any effect observed on short sequences persists as the attention window grows. Both datasets are solvable from scratch with lexical features alone, which is a deliberate requirement because pretrained embeddings are highly likely to dominate the representation and obscure any contribution of the scoring function. The full AG News training set contains 120,000 examples; a 5,000-example subset is used because quantum statevector simulation scales with both sequence length and batch size, making the full set computationally prohibitive under exact simulation. A fixed validation set of 500 examples is held out from each training split for model selection. Table II demonstrates the full statistics of this setup.

TABLE II
DATASET STATISTICS. VOCABULARY BUILT FROM THE TRAINING SPLIT WITH MINIMUM FREQUENCY 2. †SUBSET OF THE FULL TRAINING SET.

Property	TREC-6 [27]	AG News [28]
Task	Question type	News topic
Classes	6	4
Full / used train	5,452 / 4,952	120k / 5,000†
Validation / test	500 / 500	500 / 7,600
Vocabulary	3,234	12,218
Median / 95th len.	10 / 17	38 / 57
Max len. used	20	50

B. Training

All models are optimised with Adam [29] for 20 epochs using two parameter groups: classical weights at 2×10^{-3} and the 24 circuit parameters at 10^{-2}, the latter has been set higher to compensate for the characteristically smaller gradients of variational quantum circuits [5]. The batch size is kept as 64 while applying no learning-rate scheduler or weight decay, keeping the training protocol as simple as possible so that differences between conditions cannot be attributed to optimiser tuning. The checkpoint with the highest validation accuracy is selected and test accuracy is reported only at that checkpoint, ensuring the test set plays no role during training. Each condition is run over five random seeds and results are reported as mean ± standard deviation. Statistical significance is assessed with exact two-sided permutation tests over the pooled seeds of each pair of conditions ($\binom{10}{5}$=252 arrangements, p<0.05).

C. Experiments

In this experiment, all the concerned circuits are simulated using PennyLane's [30] `default.qubit` backend with backpropagation-based gradients. The five experiments that have been performed are described sequentially as follows.

Primary ablation. Experiment A compares all five conditions (classical, qnone, qintra, qcross, qfull) in the default four-head setting, providing the primary test of whether entanglement affects classification accuracy.

Dilution control. Because three classical heads share the workload in this configuration, any quantum effect could be masked by compensation from the majority. Experiment B therefore reduces the model to a single head (H=1), making the quantum circuit the sole attention mechanism.

Data-efficiency control. If entanglement still shows no effect, the explanation might be that sufficient training data allows the classical scorer to saturate. Experiment C varies training size through $N \in \{100, 250, 500, 1000, 2000\}$ to test whether a quantum advantage emerges in the low-data regime.

Encoding-bottleneck control. A further possibility is that projecting d_h=16 dimensions down to h=2 rotation angles discards too much information for entanglement to act on. Experiment D widens the circuit to $n \in \{2, 4, 8\}$ qubits, encoding 1, 2, or 4 angles per register. Because the statevector grows as 2^n, conservative settings are used (batch 16, max lengthof 14, 800 training examples, 3 seeds) to keep the entire simulation tractable.

Placement control. A final possibility is that block 0 operates on raw token embeddings, where entanglement between query and key may have little to exploit, whereas a deeper layer sees representations that are already contextualised. Experiment E, therefore, moves the quantum head from block 0 to block 1, to test whether entanglement becomes useful beyond raw embeddings.

V. Results

A. Classification Accuracy

Table III reports test accuracy for every entanglement condition across the three control experiments reported in the table and both datasets. The central finding is visible without statistical testing which shows all five conditions fall within a narrow band on every experiment, and no quantum condition consistently leads or lags the classical baseline.

In the primary ablation on TREC-6, qcross achieves the highest mean at 83.0% and classical sits at 82.4%, yet the permutation test yields p=0.44, well above the significance threshold. On AG News the pattern is even flatter, with classical, qnone, and qcross all tied at 77.1%. The remaining quantum conditions sit within a point of this value, and no permutation p falls below 0.69 for any entangling condition on this dataset.

Removing the three classical heads that could compensate for the quantum head does not change the picture. In the dilution control, which is focused on a single attention head, the results for every condition remain within roughly one point of classical approaches on both datasets, and the entangling conditions stay non-significant throughout the experiment ($p \geq 0.44$ on TREC-6, $p \geq 0.25$ on AG News). Moving the quantum head to block 1, where it operates on already-contextualised representations rather than raw embeddings, likewise produces no separation as it can be see that, the lowest p among entangling conditions across both datasets in this setting is 0.11.

Two cells in Table III do fall below 0.05 and require direct comment. On AG News, qintra reaches p=0.04 in the primary ablation and qnone reaches p=0.01 in the dilution control. Both are marked with † in the table. Notably, both belong to the separable group (green rows) as qintra applies CZ gates only within each register and produces no query-key entanglement by Proposition 2, while qnone contains no two-qubit gates at all. Moreover, the qintra result goes in the wrong direction (76.2% vs. a classical 77.1%), and neither value survives a Bonferroni correction for the four comparisons drawn in each experiment. The genuinely entangling conditions, qcross and qfull, remain tied with classical in every cell of the table. The two significant values thus localise the small detectable effects to the nonlinear encoding shared by all quantum variants, not to entanglement.

B. Attention Entropy and Scaling Controls

Although the accuracy remains unchanged, the quantum circuit still behaves differently from its classical counterpart. Figure 3(a) shows the mean attention entropy of the modified head on the test set. On both datasets, every quantum condition

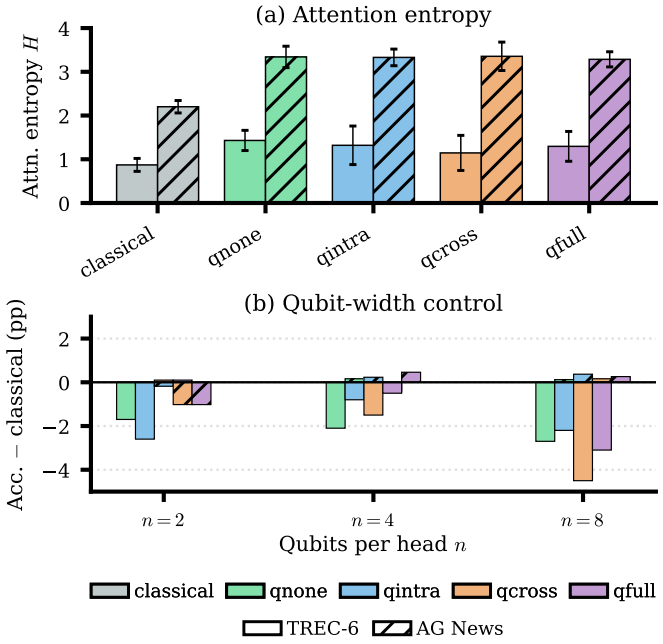


Fig. 3. (a) Mean attention entropy by condition and dataset (solid bars: TREC-6; hatched bars: AG News). (b) Qubit-width control, shown as each condition's deviation from its own dataset's classical baseline.

produces substantially higher entropy than classical: roughly 1.1 to 1.4 nats versus 0.87 on TREC-6, and 3.3 nats versus 2.2 on AG News. The quantum scorer thus spreads attention more broadly than the dot product, yet this broader distribution does not translate into an accuracy difference. The entropy elevation also does not track entanglement within the quantum group: qnone, which has no CZ gates, produces entropy comparable to or above the entangling conditions, confirming that the effect originates in the nonlinear encoding rather than the entangling layer. Figure 4 tests whether the null is an artifact of data abundance, plotting each dataset on its own axis since TREC-6 and AG News span different training-set sizes. As training size decreases from 2000 to 100 examples on TREC-6 and from 5000 down to 250 on AG News, all five conditions track one another closely, and at the smallest sizes the classical scorer is if anything marginally ahead (58.8% vs. 50.9% for qfull at N=100 on TREC-6). There is no low-data regime in which entanglement pays off. At the largest AG News size, N=5000, every condition's accuracy sits within half a point of its value in the primary ablation on the full training set, indicating the curve has converged to the same result obtained independently in Section V-A. Finally, Figure 3(b) varies the circuit width from two to eight qubits per head and plots each condition's deviation from its own classical baseline, so that TREC-6 and AG News, which differ substantially in absolute accuracy at this reduced training size, can be read on one axis. At every width, on both datasets, the two entangling conditions (qcross, qfull) remain interleaved with the two non-entangling conditions (qnone, qintra) rather than separating from them as none of the four deviations exceeds

TABLE III

TEST ACCURACY (%, MEAN ± STD OVER FIVE SEEDS) AND PERMUTATION P-VALUES AGAINST THE CLASSICAL BASELINE, ACROSS THE THREE CONTROL EXPERIMENTS REPORTED HERE AND BOTH DATASETS. ROWS ARE GROUPED BY THE SEPARABILITY STATUS ESTABLISHED IN TABLE I: QNONE AND QINTRA PRODUCE A SEPARABLE SCORE, QCROSS AND QFULL PRODUCE A NON-SEPARABLE ONE. NO NON-SEPARABLE CONDITION DIFFERS SIGNIFICANTLY FROM CLASSICAL IN ANY SETTING. THE TWO VALUES MARKED † FALL BELOW 0.05 BUT BOTH OCCUR IN THE *separable* GROUP, NEITHER SURVIVES A MULTIPLE-COMPARISON CORRECTION, AND THE NON-SEPARABLE CONDITIONS REMAIN TIED WITH CLASSICAL THROUGHOUT, THE OPPOSITE OF WHAT AN ENTANGLEMENT-DRIVEN EFFECT WOULD PREDICT.

	TREC-6						**AG News**					
	Primary		**Dilution**		**Placement**		**Primary**		**Dilution**		**Placement**	
Condition	4 heads, block 0		1 head, block 0		4 heads, block 1		4 heads, block 0		1 head, block 0		4 heads, block 1	
	Acc	*p*	*Acc*	*p*	*Acc*	*p*	*Acc*	*p*	*Acc*	*p*	*Acc*	*p*
classical *(baseline)*	82.4±1.1	—	81.9±2.0	—	82.5±1.3	—	77.1±0.3	—	77.8±0.1	—	76.5±0.7	—
qnone	81.8±0.7	0.35	82.5±0.9	0.57	81.8±1.4	0.39	77.1±0.8	0.94	**78.5±0.4**	0.01†	**77.3±0.3**	0.06
qintra	82.5±1.2	0.91	**83.2±1.4**	0.31	82.9±1.8	0.72	76.2±0.7	0.04†	78.3±0.9	0.32	76.2±0.8	0.56
qcross	**83.0±1.4**	0.44	82.3±1.1	0.68	**83.6±1.5**	0.29	77.1±0.7	0.98	78.2±0.9	0.25	76.6±0.7	0.87
qfull	81.6±1.1	0.35	82.8±1.2	0.44	82.9±1.3	0.71	76.6±1.3	0.69	77.7±0.8	0.97	77.2±0.5	0.11

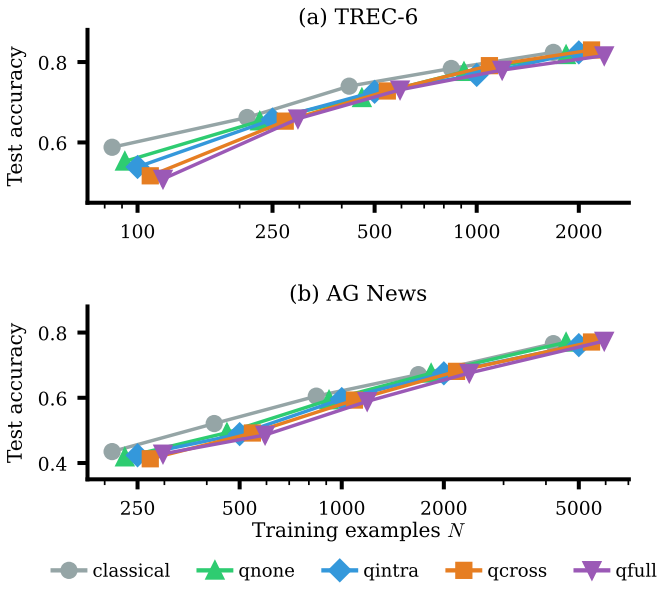


Fig. 4. Data efficiency on (a) TREC-6 and (b) AG News datasets. Markers offset horizontally for visibility; each panel uses its own y-axis range.

0.4 points on AG News at any width, and on TREC-6 the largest deviation in either direction is a 4.5-point *drop* for qcross at n=8, the opposite of an entanglement benefit. At two qubits each register holds a single qubit, so the intra-register gates are empty and qfull coincides with qcross, as the matching bars confirm on both datasets. Widening the encoding does not unlock an entanglement advantage on either dataset; if anything, the larger statevector makes optimisation harder without offering a compensating benefit.

VI. DISCUSSION AND LIMITATIONS

The experiments reveal a clean dissociation: a parameterised quantum circuit that verifiably entangles the query and key registers gives no accuracy advantage over a separable classical score, on two datasets, across five independent controls, yet it measurably alters the attention distribution through nonlinear encoding alone. Most quantum attention proposals conflate nonlinearity with non-separability across registers, but the five conditions in this study pull these apart. Proposition 1 keeps the entangling gates visible to the readout via a trainable post-entanglement layer, and Proposition 2 in turn establishes that qcross and qfull produce a genuinely non-separable score while qnone and qintra do not. Against this verified split, Table III shows no dependence on separability, while Figure 3(a) shows a large dependence on nonlinearity, with every quantum condition raising attention entropy by roughly 0.5 nat on TREC-6 and over a full nat on AG News. Accordingly, nonlinearity changes the attention distribution whereas changes due to entanglement remain insignificant. This separation, not the null alone, is the paper's central claim, and the two sub-threshold p-values reinforce it, as both land in separable conditions, exactly where a true null predicts and a real effect would not. More broadly, the result is best read as a statement about where entanglement is representationally necessary. Separable scoring functions evidently suffice for topic and question-type classification from short text, thereby leaving a non-separable interaction nothing to capture. This in turn supplies a mechanistic account of the pattern in prior work whereby quantum-circuit advantages concentrate on data with relational or quantum-native structure rather than on standard language representations. Several limitations, however, bound the scope of these findings. The circuits are small and classically simulated, the entangling layer occupies a single head, and both datasets classify English text by lexical content rather than by structural relations between tokens. Moreover, AG News is subsampled to 5,000 examples to keep statevector simulation tractable, and with five seeds per condition the permutation tests can detect differences of two to three percentage points but may miss smaller effects. The null is also specific to CZ-based entangling layers with R_Y angle encoding and Z-basis readout, so other gate sets, encoding strategies, or measurement schemes could yield a different

outcome. These choices are standard primitives, and Proposition 1 shows the post-entangler layer ensures the entangler remains visible under this readout. Constructing a task whose label provably requires a non-separable query-key interaction, and verifying that it survives the same nonlinearity-versus-entanglement decomposition, is therefore the logical next step.

VII. Conclusion and Future Work

This study demonstrates that replacing the dot-product scorer in a single Transformer attention head with a parameterized quantum circuit does not improve classification accuracy on either of the two text classification benchmarks. The quantum head does, however, change the resulting attention patterns through its nonlinear encoding, thus making this null accuracy result more informative than a simple failure to improve performance. The analysis separates the effects of nonlinearity from those of query-key entanglement using five entanglement conditions supported by two formal propositions. Across these conditions, entanglement produces minimum change in performance. In contrast, the nonlinear encoding leaves a detectable effect on the attention distribution. These findings suggest three directions for future work. First, a task requiring non-separable query-key interactions could test whether the quantum circuit benefits from entanglement when the task demands it. Second, applying quantum circuits to more attention heads and deeper Transformers could reveal whether entanglement effects emerge at greater quantum-to-classical ratios. Finally, running the same experiments on quantum hardware could show how noise and finite-shot sampling affect the balance between nonlinearity and entanglement.

VIII. Acknowledgment

This material is partly based upon work supported by the U.S. NSF Grant CRII-IIS-RI-2553868, NSF NAIRR support (NAIRR260238). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

[1] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell *et al.*, "Quantum supremacy using a programmable superconducting processor," *Nature*, vol. 574, no. 7779, pp. 505–510, 2019.

[2] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, "Quantum machine learning," *Nature*, vol. 549, no. 7671, pp. 195–202, 2017.

[3] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio *et al.*, "Variational quantum algorithms," *Nature Reviews Physics*, vol. 3, no. 9, pp. 625–644, 2021.

[4] M. Schuld and N. Killoran, "Quantum machine learning in feature hilbert spaces," *Physical review letters*, vol. 122, no. 4, p. 040504, 2019.

[5] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, "Barren plateaus in quantum neural network training landscapes," *Nature communications*, vol. 9, no. 1, p. 4812, 2018.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 5797–5808.

[8] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does bert look at? an analysis of bert's attention," in *Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, 2019, pp. 276–286.

[9] A. M. Smaldone, Y. Shee, G. W. Kyro, M. H. Farag, Z. Chandani, E. Kyoseva, and V. S. Batista, "A hybrid transformer architecture with a quantized self-attention mechanism applied to molecular generation," *Journal of chemical theory and computation*, vol. 21, no. 10, pp. 5143–5154, 2025.

[10] R. Lorenz, A. Pearson, K. Meichanetzidis, D. Kartsaklis, and B. Coecke, "Qnlp in practice: Running compositional models of meaning on a quantum computer," *Journal of Artificial Intelligence Research*, vol. 76, pp. 1305–1342, 2023.

[11] F. Nausheen, K. Ahmed, M. I. Khan, and F. Riaz, "Quantum natural language processing: A comprehensive review of models, methods, and applications," *arXiv preprint arXiv:2504.09909*, 2025.

[12] S. Y.-C. Chen, S. Yoo, and Y.-L. L. Fang, "Quantum long short-term memory," in *Icassp 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 8622–8626.

[13] G. Li, X. Zhao, and X. Wang, "Quantum self-attention neural networks for text classification," *Science China Information Sciences*, vol. 67, no. 4, p. 142501, 2024.

[14] J. Shi, R.-X. Zhao, W. Wang, S. Zhang, and X. Li, "Qsan: A near-term achievable quantum self-attention network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 8, pp. 13 995–14 008, 2025.

[15] C.-S. Chen and E.-J. Kuo, "Quantum adaptive self-attention for quantum transformer models," *arXiv preprint arXiv:2504.05336*, 2025.

[16] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, "Parameterized quantum circuits as machine learning models," *Quantum science and technology*, vol. 4, no. 4, p. 043001, 2019.

[17] S. Sim, P. D. Johnson, and A. Aspuru-Guzik, "Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms," *Advanced Quantum Technologies*, vol. 2, no. 12, p. 1900070, 2019.

[18] C. Ortiz Marrero, M. Kieferová, and N. Wiebe, "Entanglement-induced barren plateaus," *PRX quantum*, vol. 2, no. 4, p. 040316, 2021.

[19] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, "Supervised learning with quantum-enhanced feature spaces," *Nature*, vol. 567, no. 7747, pp. 209–212, 2019.

[20] S. Lloyd, M. Schuld, A. Ijaz, J. Izaac, and N. Killoran, "Quantum embeddings for machine learning," *arXiv preprint arXiv:2001.03622*, 2020.

[21] M. Schuld, R. Sweke, and J. J. Meyer, "Effect of data encoding on the expressive power of variational quantum-machine-learning models," *Physical Review A*, vol. 103, no. 3, p. 032430, 2021.

[22] B. Coecke, M. Sadrzadeh, and S. Clark, "Mathematical foundations for a compositional distributional model of meaning," *arXiv preprint arXiv:1003.4394*, 2010.

[23] E. A. Cherrat, I. Kerenidis, N. Mathur, J. Landman, M. Strahm, and Y. Y. Li, "Quantum vision transformers," *Quantum*, vol. 8, no. arXiv: 2209.08167, p. 1265, 2024.

[24] F. Chen, Q. Zhao, L. Feng, C. Chen, Y. Lin, and J. Lin, "Quantum mixed-state self-attention network," *Neural Networks*, vol. 185, p. 107123, 2025.

[25] R.-X. Zhao, J. Shi, and X. Li, "Qksan: A quantum kernel self-attention network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 184–10 195, 2024.

[26] N. Khatri, G. Matos, L. Coopmans, and S. Clark, "Quixer: A quantum transformer model," *arXiv preprint arXiv:2406.04305*, 2024.

[27] X. Li and D. Roth, "Learning question classifiers," in *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

[28] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, 2015.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, S. Ahmed, V. Ajith, M. S. Alam, G. Alonso-Linaje, B. AkashNarayanan, A. Asadi *et al.*, "Pennylane: Automatic differentiation of hybrid quantum-classical computations," *arXiv preprint arXiv:1811.04968*, 2018.