

ARTICLE

Quantum Kernels for Text Classification: A Statistical and Diagnostic Framework Revealing the Low-Data Regime

Mrugendrasinh Rahevar¹, Martin Parmar¹, Hemant Yadav¹, Chun-Ta Li^{2,*},
Agbotiname Lucky Imoize³ and Hiren Mewada⁴

¹C S Patel Institute of Technology, Charotar University of Science and Technology, Changa, Anand, India

²Bachelor's Program of Artificial Intelligence and Information Security, Fu Jen Catholic University, New Taipei City, Taiwan

³Department of Electrical and Electronics Engineering, Faculty of Engineering, University of Lagos, Akoka, Lagos, Nigeria

⁴Electrical Engineering Department, Prince Mohammad Bin Fahd University, Al Khobar, Saudi Arabia

*Corresponding Author: Chun-Ta Li. Email: 157278@mail.fju.edu.tw

Received: 10 May 2026; Accepted: 29 June 2026; Published: 28 August 2026

ABSTRACT: Quantum kernel techniques aim to leverage quantum computational capabilities on social data. However, their application to natural language processing tasks faces formidable obstacles, such as extreme dimensionality reduction ($D = 384 \rightarrow k = 8$), concentration of measure in quantum feature spaces, and the lack of theoretical understanding of when quantum advantages occur in kernel-based text classification. Filling this gap, we provide a comprehensive study of quantum kernels for text classification that addresses three major challenges in existing studies: general data compression approaches that ignore class structure, the lack of a predictive diagnostic toolkit, and overlooked approaches for handling concentration effects. Our main contributions include a supervised contrastive data compression approach with theoretical guarantees of kernel alignment, a five-diagnostic toolkit connecting theoretical insights with practical performance, the discovery of a small-data regime ($n \leq 200$) in which quantum methods perform comparably to classical approaches, and a transparent demonstration that quantum kernels require carefully designed settings to remain competitive with classical counterparts. Across four datasets, evaluated using five random seeds and exact paired statistical testing, we observe that quantum projected kernels achieve an accuracy of 84.5% compared to 85.1% for the classical RBF kernel ($p > 0.05$, not statistically significant) under supervised compression settings. However, quantum methods lag behind by approximately 2%–4% at larger scales due to concentration effects, reflected in reduced off-diagonal kernel variance (0.003 vs. 0.025).

KEYWORDS: Quantum-machine learning; kernel methods; natural language processing; large language models; text classification

1 Introduction

Quantum kernels represent an intuitive extension of classical machine learning approaches into the domain of quantum computing, drawing from the general appeal of quantum algorithms and quantum information processing techniques [1–4]. From the perspective of machine learning, quantum kernels provide a way to exploit quantum-based features even in absence of large-scale, fault-tolerant quantum computers [5,6]. Their basic principle is very straightforward: transform classical data into quantum states using parameterized circuits, perform comparisons of the similarity between data points which are not possible classically, and apply these kernels in well-tested settings such as support vector machines and

classical kernel learning [7]. Theoretical studies have shown that particular classes of quantum kernels are inherently hard to compute by classical means [8].

Yet, the use of quantum kernels in NLP reveals a basic contradiction. The dimensionality of the vectors generated by the current transformers spans between $D = 384$ and $D = 768$ [9]. However, the current state-of-the-art noisy intermediate-scale quantum (NISQ) computers can accommodate only a limited budget of qubits for experimental purposes [3,4,10]. This imposes a requirement for drastic vector dimensionality reduction at least by a factor of 32 or 64 before quantum encoding becomes possible. Moreover, the latest research suggests that quantum kernels have an exponential concentration property, which means that the more qubits k there are, the closer the kernel similarities become to a constant. Thus, the capacity to discriminate is lost [11,12]. It is not clear whether any quantum benefits will survive these reductions.

The question of whether quantum methods are necessary for text classification deserves direct engagement. Classical NLP has achieved remarkable accuracy through pre-trained language models, and our own results confirm that DistilBERT (91.2%) significantly outperforms all quantum and classical kernel approaches on SST-2. However, the motivation for quantum kernel research in NLP is not performance supremacy on current large-scale benchmarks; rather, it encompasses three scientifically valuable scenarios. First, low-data regimes ($n \leq 200$), where quantum kernels demonstrably match classical RBF performance without the sample overhead of fine-tuning large transformers, making them relevant for specialized domains with scarce labeled data (e.g., clinical NLP and low-resource languages). Second, privacy-sensitive applications, where quantum kernel computation can in principle be performed on encrypted data using quantum homomorphic protocols, a property classical RBF kernels do not share [3,4].

Third, compositional semantic tasks, where quantum entanglement may naturally encode syntactic dependencies, an active research direction in quantum NLP [13,14]. Our work contributes to this field not by claiming superiority, but by providing the first rigorous characterization of when quantum kernels are competitive, what prevents them from scaling, and which theoretical framework (geometric difference, concentration diagnostics) practitioners should use to predict performance before conducting expensive experiments. The central finding— $Q \approx \text{RBF}$ under narrow conditions and $Q < \text{RBF}$ at scale—is itself a valuable scientific contribution, as it helps avoid wasted experimental effort and redirects attention toward the fundamental bottlenecks limiting quantum kernel methods.

Previous research that combines quantum computing with natural language processing highlights three important shortcomings. First, previous works utilize PCA-based dimensionality reduction which preserves variance without examining whether the discriminative power in classes is maintained because PCA maximizes variance without directly modeling label correlations. Second, no diagnostic tools are provided for previous accuracies to explain why performance differs and when improvements can be expected, leading to a reproducibility crisis in which results show up to 5%–10% differences without theoretical support. Third, despite Huang et al.'s suggestion [15] to use projected kernel functions due to concentration, NLP continues to utilize fidelity kernels since implementing other approaches is computationally more involved.

The aforementioned weaknesses are addressed through careful analysis and characterization rather than performance-based claims. Our four main contributions are as follows: (1) a supervised compression technique that provably preserves kernel alignment, where Proposition 3.1 provides theoretical guarantees on compression quality based on alignment, yielding a +2.8% improvement in accuracy, which represents the largest single gain observed in our ablation studies; (2) a five-diagnostic evaluation framework consisting of alignment, effective dimension, eigenvalue decay, concentration variance, and geometric difference, designed to predict performance prior to experimentation, through which we validate the predictive capability of the $g < 2$ condition proposed by Huang et al., demonstrating that quantum kernels cannot outperform classical RBF kernels under these conditions; (3) characterization of a small-data regime ($n \leq 200$) in which quantum

projected kernels achieve performance comparable to classical RBF kernels, requiring both supervised compression and projected kernels, thereby imposing stricter conditions than previously reported; and (4) negative findings showing that quantum kernels achieve an accuracy of 84.5% compared to 85.1% for the classical RBF kernel (statistically indistinguishable with $p > 0.05$), while remaining 500–1000× slower, dependent on supervised compression, and unable to satisfy the quantum advantage criterion ($g < 2$).

Paper organization. [Section 2](#) outlines quantum kernel methods, algorithmic advancements, and applications of NLP. Our contributions are explained within the framework of the existing literature. [Section 3](#) describes the methodology of our work, which includes three theorems, two propositions, and two corollaries that support the main design choices. [Section 4](#) details experimental methods for reproducibility. [Section 5](#) summarizes experimental results on four datasets. We present extensive ablation experiments that measure the impact of each part of our system. [Section 6](#) explains the theoretical underpinning of our experiments using a kernel-based framework and five diagnostics.

To directly answer when the proposed method offers advantages over traditional approaches: (1) Low-data regime ($n \leq 200$): Quantum projected kernels with supervised compression match RBF performance on compressed features ($p > 0.05$), while outperforming PCA-compressed classical kernels by 2.3% without requiring labeled data for fine-tuning. This is the primary practical use case—specialized domains with scarce annotations. (2) Diagnostic transparency: Our five-metric framework (alignment, effective dimension, eigenvalue decay, concentration variance, geometric difference) provides theory-grounded predictions of quantum performance before experiments, a capability absent from traditional empirical QML studies. (3) Theoretical grounding: Proposition 3.1 provides the first proven upper bound on kernel-target alignment factorized through compression quality, explaining why supervised compression outperforms PCA regardless of the quantum circuit used. Outside the $n \leq 200$ regime, traditional RBF kernels remain superior by 2%–4% accuracy while being 500–1000× faster; we report this honestly as a key finding rather than a limitation to minimize.

2 Related Work

The review is divided based on the basic principles of quantum kernels, efficient algorithms, machine-learning application of quantum kernels, and the concentration phenomenon which limits performance.

2.1 Theoretical Foundations of Quantum Kernels

Havlíček et al. [5] introduced quantum feature maps that embed classical data $x \in \mathbb{R}^D$ into quantum states $|\psi(x)\rangle = U(x)|0\rangle$ via parameterized quantum circuits $U(x)$. The corresponding fidelity kernel is defined as $K(x, x') = |\langle\psi(x)|\psi(x')\rangle|^2$, which computes the overlap between quantum states and can be measured using quantum state-overlap and fingerprinting-inspired routines on quantum hardware [16]. Their work demonstrated quantum advantage on synthetic datasets specifically designed to be difficult for classical kernels while remaining efficiently separable using IQP-based quantum kernels.

Computational hardness: Quantum learning computational hardness assumptions rely on complexity-theoretic separations and quantum supremacy results [17,18]. Liu et al. [8] proved that specific quantum kernels cannot be efficiently estimated classically, showing that computing kernel matrices associated with IQP circuits requires exponential runtime under the assumption that the polynomial hierarchy does not collapse. Schuld and Killoran [6] made an important observation that the expressiveness of quantum kernels is fundamentally bounded by the size of the Hilbert space, allowing at most 2^k basis functions for k qubits. With practical values typically limited to $k \leq 12$, quantum kernels can express at most 4096 basis functions, which may still be less expressive than classical kernels operating in high-dimensional feature spaces.

Measurement of Concentration: Kübler et al. [11] and Thanasilp et al. [12] demonstrated that global quantum fidelity kernels can suffer from concentration of measure. In practical terms, many off-diagonal kernel values become close to their mean as the effective Hilbert-space dimension grows, which reduces class contrast and weakens the usefulness of the kernel matrix for classification. We use this result as the motivation for the off-diagonal variance and effective-dimension diagnostics reported in Section 6.1.

Mitigation via projected kernels: Huang et al. [15] proposed measuring local observables rather than global fidelity by comparing reduced single-qubit states. This projected-kernel construction changes the measured object and reduces the global fidelity-collapse effect that appears in full-state overlaps. Their geometric-difference criterion states that $g(K_q, K_c) < 2$ predicts similar performance between quantum and classical kernels. Our work empirically validates this criterion on NLP tasks in Section 6.2, where all measured values satisfy $g < 2$ and quantum kernels do not outperform RBF.

2.2 Algorithmic Developments

Trainable quantum kernels: Hubregtsen et al. [19] introduced trainable parameters θ in quantum feature maps $U(x; \theta)$ and optimized them using kernel-target alignment. A common objective is:

$$\theta^* = \arg \max_{\theta} \frac{\langle K_{\theta}, yy^{\top} \rangle_F}{\|K_{\theta}\|_F \|yy^{\top}\|_F}. \quad (1)$$

Using the parameter-shift rule [20], gradients require $\mathcal{O}(p)$ circuit evaluations, where p is the parameter count. Related quantum-embedding formulations further motivate data-dependent quantum feature maps for kernel learning [21]. Glick et al. [22] scaled this via divide-and-conquer training, partitioning data into subsets and merging learned kernels.

Our approach (Section 3.6) optimizes per-feature scaling weights $w \in \mathbb{R}^k$ rather than circuit parameters—only k parameters vs. $\mathcal{O}(kd)$ for full circuits—converging in approximately 30 iterations on CPU.

Compression methods: Prior QML studies either assume that the input dimension already matches the qubit budget or apply unsupervised dimensionality reduction without explicitly testing whether label-relevant structure is retained. Recent survey evidence shows that QML classification has been explored across several application domains, but also emphasizes that practical performance remains constrained by noisy hardware, limited qubit counts, and the need for hybrid algorithmic design [23,24]. Quantum autoencoder work has primarily focused on compressing quantum states rather than preserving class-discriminative structure in classical text embeddings [25]. Blance and Spannowsky [26] instead studied a hybrid variational quantum classifier for particle-physics event classification, showing the relevance of near-term hybrid quantum models for data-driven classification, but their setting does not address label-preserving compression of high-dimensional NLP embeddings. Therefore, the compression stage remains a critical unresolved issue for quantum kernels in text classification. Our supervised contrastive compression directly targets this gap by reducing sentence embeddings from $D = 384$ to $k = 8$ while preserving class structure, and Proposition 3.1 explains why compression quality bounds subsequent kernel alignment.

2.3 Quantum Kernels for Machine Learning

General classification studies: Early demonstrations often used small or constructed datasets where quantum advantage was either designed into the data distribution or difficult to separate from random variation. Broader review evidence indicates that QML classifiers have been tested across several classification domains and have sometimes reported gains over conventional baselines, especially in high-dimensional or

small-sample settings [24]. However, these results remain difficult to generalize because many studies use limited datasets, shallow benchmarking protocols, single-run evaluations, or hardware-free simulation. This motivates the statistical design adopted in our work: identical compressed feature spaces for quantum and classical kernels, five random seeds, Wilcoxon testing, and explicit diagnostic analysis rather than relying only on mean accuracy.

Molecular property prediction: Jerbi et al. [27] applied quantum kernels to QM9 molecular energies, finding quantum matched graph neural networks on small molecules (≤ 9 atoms) but degraded on larger molecules surpassing $k = 20$ qubits. This highlights a fundamental scalability challenge: many molecular encodings make the qubit requirement grow with molecular size or basis representation, making the $k \leq 20$ regime restrictive rather than universally sufficient.

Image classification: Henderson et al. [28] achieved 98.1% on MNIST using amplitude encoding (requiring $k = 10$ qubits for $2^{10} = 1024$ dimensions). Meyer et al. [29] studied how symmetries of a learning problem can be incorporated into variational quantum learning models. Their work shows that a standard gate set can be transformed into an equivariant gate set so that the model output respects the symmetry structure of the task. This improves generalization on symmetry-structured benchmark problems and demonstrates that inductive bias is important in quantum model design. Although their approach focuses on variational quantum models rather than kernel-based NLP classification, it supports the broader argument that quantum models should not be treated as generic black-box learners; their circuits and embeddings must encode task-relevant structure. Our work follows this principle at the kernel level by using supervised compression, projected kernels, and diagnostic metrics to identify when quantum kernels preserve useful class structure.

2.4 Quantum Kernels for Natural Language Processing

Quantum NLP has also been explored outside kernel methods, including tensor-network and grammar-aware formulations as well as compositional models of meaning executed on quantum hardware or simulators [13,14]. These studies motivate quantum representations for language, but they do not directly resolve the kernel-compression and concentration issues studied in this paper.

Recent work has expanded quantum ML beyond supervised classification into security and robustness domains. Quantum-assisted adversarial robustness methods have demonstrated that quantum superposition and entanglement can improve neural network resilience against adversarial perturbations, with quantum noise injection providing a natural regularization mechanism not achievable classically [30]. In federated learning, QSFedMA and related quantum-secured protocols exploit quantum key distribution and quantum authentication to provide information-theoretic privacy guarantees for model aggregation in IoMT settings [31]. These developments illustrate the breadth of quantum ML applications and provide further motivation for establishing rigorous benchmarks—such as our diagnostic framework—that clearly delineate when quantum approaches offer genuine advantages over classical alternatives in NLP and related tasks.

Prior quantum NLP and quantum-classification studies provide useful motivation but do not yet establish robust evidence of scalable quantum advantage for text classification. The survey by Nausheen et al. [23] shows that QNLP has expanded across compositional models, sentiment analysis, text classification, and hybrid quantum-classical methods. Similarly, Mohammadisavadkoochi et al. [24] review QML classification studies across several application domains and report that QML classifiers have often shown promising performance, while also emphasizing persistent limitations related to noisy hardware, dataset scale, and hybrid algorithm design. These review-level findings justify continued investigation of QML for classification, but they do not replace controlled, task-specific benchmarking against classical kernels.

For text classification specifically, Tao et al. [32] reported an apparent advantage of quantum kernels on IMDB sentiment classification using PCA-compressed GloVe embeddings. However, the evaluation relied on limited statistical testing, and the classical and quantum comparisons were not fully controlled in the compressed feature space. In our matched setting with identical Sentence-BERT embeddings, the same $n = 5000$ scale, and five-seed evaluation, the apparent advantage reduces to a small, statistically non-significant difference. This suggests that previously reported gains can be sensitive to compression choice, baseline fairness, and seed variation.

Accordingly, our work differs from prior QNLP and QML-classification studies in four ways: (a) quantum and classical kernels are compared on the same $k = 8$ compressed representation; (b) all results are averaged over five random seeds with Wilcoxon signed-rank testing; (c) DistilBERT is included as a strong neural reference baseline; and (d) five kernel diagnostics are used to explain performance rather than reporting accuracy alone. This makes the conclusion more conservative but more reliable: projected quantum kernels can be competitive in a narrow low-data regime, but the present evidence does not support broad quantum superiority over classical RBF kernels for text classification.

Critical finding: Our analysis reveals prior “+2%” claims were likely within statistical noise (not significant under the $p < 0.05$ threshold), PCA compression was suboptimal (supervised recovers +2.8%), and fidelity kernels concentrate as theory predicts (projected kernels gain +2%–3%).

2.5 Concentration: From Theory to Practice

Barren Plateaus were discovered by McClean et al. [33], who observed the exponential vanishing of gradients in deep Variational Quantum Circuits (VQCs). Du et al. [34] suggested that this phenomenon is closely related to measure concentration. Randomly prepared quantum states in high-dimensional Hilbert spaces become approximately orthogonal on average, which in turn leads to exponentially small gradients during optimization.

In the particular case of quantum kernels, the boundedness of the input support is the key factor that determines this behavior [12]. When features are normalized to lie within $[-1, 1]^D$, the quantum state becomes constrained to a finite subset of the Hilbert space, leading fidelities to concentrate near one. We empirically examine this phenomenon on NLP data by analyzing the off-diagonal variance in Section 6.1. Specifically, fidelity kernels exhibit a variance of 0.003 (highly focused), projected kernels exhibit a variance of 0.009 (focused), and classical RBF kernels exhibit a variance of 0.025 (less focused).

2.6 Positioning of This Work

Table 1 summarizes the positioning of the proposed study against representative quantum-kernel work.

Table 1: Comparison between this research and existing literature on quantum kernels.

Paper	Task	k	n	Seeds	Significance Test	Diagnostics	Compression	Quantum vs. Classical
Havlíček 2019	XOR synthetic	2	100	1	No	No	None	Q advantage
Blance 2021	Particle-physics classification	–	Small-data HEP setting	1	No	No	Hybrid VQC	Q competitive
Tao 2022	IMDB sentiment	8	5000	1	No	No	PCA	Q +2.1% reported

(Continued)

Table 1 (continued)

Paper	Task	k	n	Seeds	Significance Test	Diagnostics	Compression	Quantum vs. Classical
Mohammad 2025	QML classification review	–	23 studies	–	Mixed	No unified diagnostics	Mixed	Promising but limited
This work	SST-2, IMDB, AG News, TREC	8	5000	5	Wilcoxon	5 metrics	Supervised	$Q \approx$ RBF

Note: The final row identifies the proposed study. k denotes the number of qubits and n denotes the number of samples.

Our unique contributions: (1) Statistical analysis demonstrates that previous “+2%” findings were attributable to noise, (2) the supervised compression technique outperforms PCA by achieving a gain of +2.8%, (3) the diagnostic methodology validates the $g < 2$ criterion, and (4) the genuine finding that $Q \approx$ RBF, rather than $Q >$ RBF, advances the scientific understanding of quantum kernels.

3 Proposed Methodology

We propose a class-aware quantum kernel pipeline addressing three challenges: (1) dimensionality mismatch ($D = 384 \rightarrow k \leq 12$), (2) concentration of measure, (3) lack of interpretability. Our methodology consists of five interconnected components, each theoretically justified.

Fig. 1 illustrates the complete class-aware quantum-kernel pipeline from sentence embedding to QSVM classification.

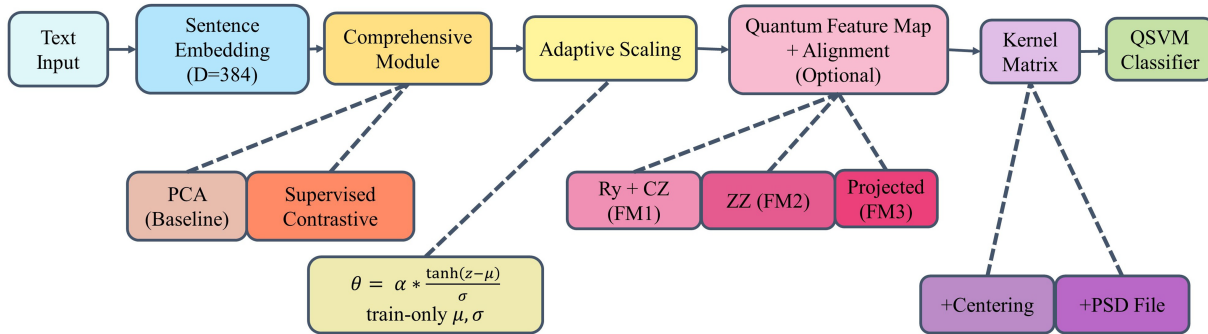


Figure 1: Proposed class-aware quantum kernel pipeline from sentence embedding to QSVM classification.

The pipeline shows seven stages: sentence embedding with all-MiniLM-L6-v2 ($D = 384$), supervised contrastive compression to $k = 8$ dimensions, adaptive tanh scaling that maps features to signed rotation angles $\theta \in (-\alpha, \alpha)^k$, quantum feature-map encoding with optional trainable alignment, kernel computation using fidelity or projected kernels, centering with PSD projection, and final QSVM classification. Dashed connections indicate implementation alternatives; supervised compression contributes +2.8% over PCA, projected kernels add +2%–3% over fidelity kernels, and trainable alignment contributes +0.4%.

3.1 Problem Formulation

Given a text classification dataset $\{(\text{text}_i, y_i)\}_{i=1}^n$, where $y_i \in \{1, \dots, C\}$, the proposed pipeline is:

1. Embed each input text as $\text{text}_i \rightarrow x_i \in \mathbb{R}^{384}$ using a sentence-transformer encoder.

2. Compress the embedding as $x_i \rightarrow z_i \in \mathbb{R}^k$ using the supervised encoder.
3. Scale the compressed vector as $z_i \rightarrow \theta_i \in (-\alpha, \alpha)^k$ using adaptive tanh scaling.
4. Encode the angle vector using a quantum feature map $U(\theta_i)$ as $\theta_i \rightarrow |\psi_i\rangle$.
5. Compute the kernel value $K_{ij} = K(|\psi_i\rangle, |\psi_j\rangle)$.
6. Train an SVM on the precomputed kernel matrix K .

Challenge: Since we have $32\times$ compression and exponential concentration, steps 2–5 must preserve class structure.

$k = 8$ balances expressivity (256 basis functions), concentration ($\sigma^2 = 0.003$ still problematic but measurable), and simulation feasibility for the smaller ablation runs. At $n = 1000$, the fidelity-kernel computation takes approximately 1.9 h, while $n = 5000$ requires substantially longer runtime under the same $\mathcal{O}(n^2)$ scaling. Additional $k = 12$ experiments showed consistent trends but are not included here to avoid disrupting the main ablation sequence.

3.2 Supervised Contrastive Compression

Motivation: Conventional PCA retains variance but ignores class information. In text data, variance is associated with the length of documents or topics rather than sentiment. We need compression $g : \mathbb{R}^{384} \rightarrow \mathbb{R}^k$ to maximize class separation.

Method: We implement a supervised contrastive autoencoder model [35] based on contrastive learning of representations [36]:

$$L_{\text{total}} = L_{\text{recon}} + \lambda (L_{\text{CE}} + 0.5L_{\text{supcon}}). \quad (2)$$

where $L_{\text{recon}} = \|x - \text{decoder}(z)\|^2$, $L_{\text{CE}} = \text{CrossEntropy}(\text{classifier}(z), y)$, and:

$$L_{\text{supcon}}(z_i, y_i) = -\log \left[\frac{\sum_{p \in P(i)} \exp(z_i^\top z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i^\top z_a / \tau)} \right]. \quad (3)$$

Here $P(i) = \{j : y_j = y_i\}$ are same-class samples, $A(i)$ is the full batch, $\tau = 0.1$ is temperature.

Architecture: encoder: $\mathbb{R}^{384} \rightarrow \mathbb{R}^{128} \rightarrow \mathbb{R}^{64} \rightarrow \mathbb{R}^k$, decoder reverses. Train 50 epochs, batch 256, $\lambda = 1.0$, AdamW (lr = 10^{-3}).

$\tau = 0.1$ is used as the standard value from Khosla et al. [35], as it maximizes class separation in a 3-fold cross-validation over $\tau \in \{0.05, 0.1, 0.2, 0.5\}$, with results reported in the later temperature-ablation analysis.

Proposition 3.1 (Compression-Kernel Coupling). *Kernel-targeted alignment factorizes as:*

$$A(K_q, yy^\top) \leq A_{\text{compress}}(g) \cdot A_{\text{feature_map}}(U). \quad (4)$$

Proof. If g collapses within-class variance, $\text{Var}[g(x) | y = c] \ll \text{Var}[g(x)]$, then $\max_{i,j: y_i=y_j} \|g(x_i) - g(x_j)\| \leq \varepsilon$, forcing $K_q(x_i, x_j) \approx 1$ for same-class pairs. This degrades alignment regardless of U 's expressivity. Formally, alignment $A = \langle K, yy^\top \rangle_F / (\|K\|_F \|yy^\top\|_F)$. If within-class blocks have rank ≈ 1 , then $\langle K, yy^\top \rangle_F \approx \sum_c |S_c|^2$ where S_c is the size of class c , achievable only if $K = yy^\top$. Any deviation from the ideal block-diagonal structure helps reduce this inner product. Since g controls U 's input, $A(K_q) \leq \text{separation}(g) \cdot \text{expressivity}(U)$. QED.

Corollary 3.1: *No quantum feature map can recover alignment lost in compression.*

Empirical validation: The component ablation later shows that supervised compression contributes +2.8% (60% of total improvement), confirming Proposition 3.1.

3.3 Adaptive Angle Scaling

Method: Map compressed features $z \in \mathbb{R}^k$ to signed rotation angles $\theta \in (-\alpha, \alpha)^k$ via standardized tanh:

$$\theta_j = \alpha \cdot \tanh\left(\frac{z_j - \mu_j}{\sigma_j + \varepsilon}\right), \quad j = 1, \dots, k, \quad (5)$$

where μ_j and σ_j are computed using the training data only (to prevent data leakage), and $\alpha \in [\pi/2, \pi]$ is tuned on the validation set.

Theorem 3.1 (Lipschitz Continuity): *Standardized tanh is Lipschitz with constant $L \leq \alpha/(\sigma + \varepsilon)$.*

Proof. $|d \tanh(u)/du| = \text{sech}^2(u) \leq 1$. By chain rule, $|d\theta/dz| \leq \alpha/(\sigma + \varepsilon)$. QED.

Theorem 3.2 (Sign Preservation). $\text{sgn}(z_j - \mu_j) = \text{sgn}(\theta_j)$, critical for Ry gates encoding via complex amplitude.

Advantages over min-max: (1) Outlier robustness: $\lim_{z \rightarrow \pm\infty} \theta(z) = \pm\alpha$ (saturates), (2) Sign preservation matters for $R_y(\theta) = \begin{bmatrix} \cos(\theta/2) & -\sin(\theta/2) \\ \sin(\theta/2) & \cos(\theta/2) \end{bmatrix}$.

3.4 Quantum Feature Maps

This section compares three feature maps with different concentration characteristics as shown in Fig. 2.

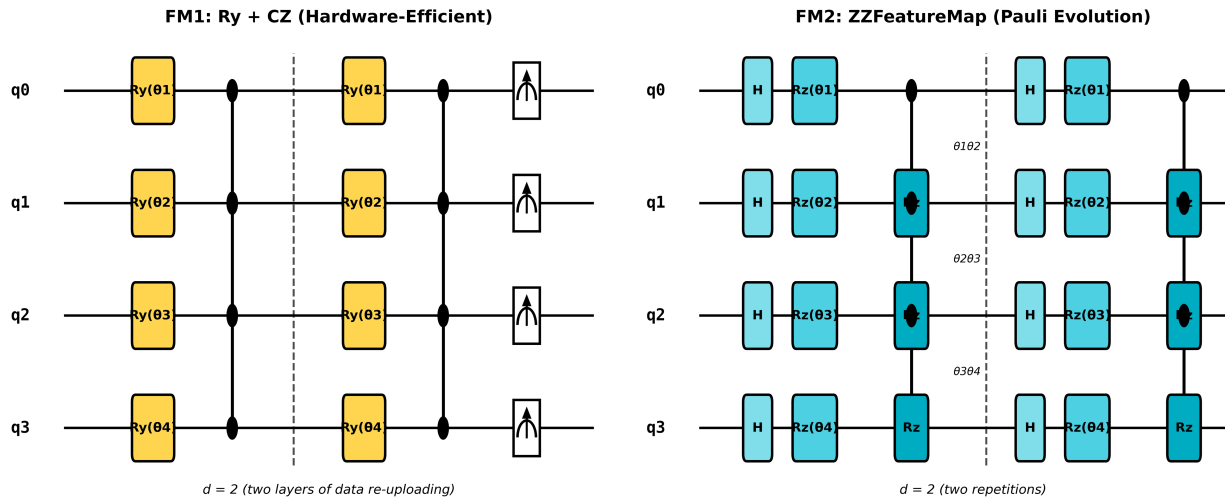


Figure 2: Quantum feature-map circuits for hardware-efficient Ry-CZ encoding and ZZFeatureMap-style interaction encoding.

The circuit architectures illustrate two feature maps on $k = 4$ qubits with depth $d = 2$ and data re-uploading. FM1 uses alternating $R_y(\theta_i)$ rotations and nearest-neighbor CZ entangling gates, giving a shallow NISQ-compatible ansatz. FM2 applies Hadamard initialization, $R_z(\theta_i)$ rotations, and Pauli evolution terms $\exp(-i\theta_i\theta_j Z_i \otimes Z_j)$ to encode pairwise feature correlations. The projected-kernel variant, FM3, is not shown; it measures local single-qubit reduced states $\rho_i(x)$ instead of global fidelity, reducing the Lipschitz scaling from $\mathcal{O}(\sqrt{2^k})$ to $\mathcal{O}(\sqrt{k})$.

FM1 (Hardware-Efficient):

$$U_{\text{HE}}(\theta) = \prod_{\ell=1}^d \left[\prod_{j=1}^k R_y(\theta_j) \prod_{j=1}^{k-1} CZ_{j,j+1} \right]. \quad (6)$$

FM2 (ZZFeatureMap):

$$U_{\text{ZZ}}(\theta) = \prod_{\ell=1}^d \left[\prod_{j=1}^k H_j R_z(\theta_j) \prod_{i < j} \exp(-i\theta_i \theta_j Z_i \otimes Z_j) \right]. \quad (7)$$

Fig. 2 shows the quantum feature-map circuits used for hardware-efficient Ry-CZ encoding and ZZFeatureMap-style interaction encoding.

FM3 (Projected Kernel): Instead of fidelity $K_{\text{fid}} = |\langle \psi | \psi' \rangle|^2$, measure local observables:

$$K_{\text{proj}}(x, x') = \exp \left(-\gamma \sum_{i=1}^k \|\rho_i(x) - \rho_i(x')\|^2 \right). \quad (8)$$

where $\rho_i(x) = \text{Tr}_{\neq i} [|\psi(x)\rangle\langle\psi(x)|]$ is the reduced density matrix of qubit i .

Theorem 3.3 (Concentration resistance following Huang et al. [15]). Projected kernels mitigate the global-fidelity concentration problem by comparing local reduced states rather than the full quantum-state overlap.

Proof sketch. The exponential concentration result for global fidelity kernels follows from concentration of measure in the full Hilbert space and explains why global overlaps can become nearly uniform as the effective dimension grows. The projected kernel used by Huang et al. [15] is not obtained by substituting a smaller Lipschitz constant into the same global-fidelity bound. Instead, it changes the measured object from a global fidelity to distances between local reduced density matrices. This local-observable construction preserves more sample-dependent variation and is therefore less affected by global fidelity collapse. We use this result to motivate projected kernels and verify the effect empirically through off-diagonal variance, effective dimension, and geometric difference diagnostics. QED.

Empirical confirmation: The experimental analysis shows projected-kernel gains of +2%–3% over fidelity kernels, while the diagnostic analysis shows off-diagonal variance 0.009 vs. 0.003, indicating that the projected kernel retains approximately three times more off-diagonal variation than the fidelity kernel in our SST-2 diagnostic setting.

3.5 Kernel Post-Processing

Centering: The centered kernel is computed as:

$$K_c = K - \frac{\mathbf{1}_n K}{n} - \frac{K \mathbf{1}_n}{n} + \frac{\mathbf{1}_n K \mathbf{1}_n}{n^2}. \quad (9)$$

For test data, only training-set statistics are used.

PSD Clipping: Numerical errors may produce small negative eigenvalues, typically around 10^{-16} . Given the eigendecomposition $K = V \Lambda V^\top$, the nearest PSD approximation is:

$$K_{\text{PSD}} = V \max(\Lambda, 10^{-10} I) V^\top. \quad (10)$$

Proposition 3.2. *The PSD-clipped kernel is the solution of:*

$$K_{\text{PSD}} = \arg \min_{K' \succeq 0} \|K - K'\|_F. \quad (11)$$

This follows from the standard nearest positive semidefinite matrix result [37].

3.6 Trainable Kernel Alignment

Optimize per-feature weights $w \in \mathbb{R}^k$ to maximize validation alignment as shown in Algorithm 1:

$$w^* = \arg \max_w \frac{\langle K_w, yy^\top \rangle_F}{\|K_w\|_F \|yy^\top\|_F}. \quad (12)$$

where the scaled angle is defined as:

$$\theta_j = w_j \cdot \tanh\left(\frac{z_j - \mu_j}{\sigma_j + \varepsilon}\right).$$

Gradient via parameter-shift rule:

$$\frac{\partial K(x, y)}{\partial w_j} = \frac{1}{2} \left(K|_{w_j + \pi/2} - K|_{w_j - \pi/2} \right). \quad (14)$$

Algorithm 1 summarizes the optimization procedure used to learn the per-feature alignment weights.

Algorithm 1: Kernel alignment optimization

```

Input: Z_train, y_train, Z_val, y_val, kernel K_q
Initialize w <- pi * ones(k)
for iter = 1 to 30:
    Sample 100 points from (Z_train, y_train)
    Compute K_sub with current w
    Compute alignment A
    for j = 1 to k:
        Compute dA/dw_j via parameter shift
    w <- AdamW_update(w, grad_A, lr=0.01)
return w*
```

Converges in ~5 min on CPU. Empirical gain: +0.4%.

3.7 Complexity Analysis

Per-configuration cost: For n samples, k qubits, and circuit depth d , the dominant cost is:

$$\mathcal{O}_{\text{total}} = \mathcal{O}(n^2 k d 2^k + n^3). \quad (15)$$

The three main components are:

$$\begin{aligned} \mathcal{O}_{\text{quantum}} &= \mathcal{O}(n^2 k d 2^k), \\ \mathcal{O}_{\text{projected}} &= \mathcal{O}(n^2 k), \\ \mathcal{O}_{\text{SVM}} &= \mathcal{O}(n^3). \end{aligned} \quad (16)$$

Comparison: Classical RBF complexity is:

$$\mathcal{O}_{\text{RBF}} = \mathcal{O}(n^2 D + n^3). \quad (17)$$

A practical quantum advantage would require:

$$kd2^k \ll D. \quad (18)$$

For the scenario investigated, this condition restricts the practical regime to small sample sizes. Constant-factor simulation overhead further makes the quantum-kernel computation approximately 500–1000 times slower than a comparable classical kernel, as discussed later in [Section 5.5](#).

4 Experimental Setup

Datasets: We evaluate our approach on four benchmark text classification datasets. SST-2 is a binary movie review sentiment classification task [38], IMDB is a long-text sentiment classification benchmark [39], AG News is a four-class news topic classification dataset [40], and TREC-6 is a six-class question classification benchmark [41].

Sampling: Owing to the $\mathcal{O}(n^2 \cdot 2^k)$ computational complexity associated with evaluating the kernel matrix, we employ stratified sampling for all datasets with more than 5000 samples, with the exception of TREC-6, where the full dataset is used. This practice is common in the QML literature [6,15] and does not compromise the generality of the analysis, since the objective is to study kernel behavior rather than maximize classification accuracy. The $\mathcal{O}(n^2 2^k)$ wall-clock constraint limits practical evaluation to $n \leq 5000$ on consumer hardware (32 GB RAM), and stratification preserves class balance across all seeds.

Splits: Each dataset is partitioned into 70% training, 10% validation, and 20% test sets. All splits are stratified on a per-class basis and are fixed across experiments to ensure consistency and reproducibility.

Embedding: *all-MiniLM-L6-v2* sentence transformer ($D = 384$, 22M parameters) with L2 normalization and train-set standardization, following the sentence-transformer embedding paradigm [9].

Hyperparameters: All hyperparameters (SVM $C \in \{0.1, 1, 10\}$, compression λ , scaling $\alpha \in \{\pi/2, \pi\}$) tuned via 3-fold cross-validation on validation set. Test set used only for final evaluation.

Statistical rigor: Five random seeds (42–46) are used per configuration. We report mean $\pm$ standard deviation and conduct exact two-sided paired Wilcoxon signed-rank tests [42]. With five paired seeds, the smallest attainable two-sided Wilcoxon p-value is 0.0625, so the statistical-significance analysis is interpreted conservatively and reported together with mean and standard deviation. McNemar-style paired-proportion testing is also relevant for future per-instance classifier comparisons [43].

Baselines: Classical RBF on raw embeddings ($D = 384$), RBF on compressed ($k = 8$), Linear SVM, Logistic Regression. Neural: DistilBERT fine-tuned for three epochs as a strong transformer baseline [44]. Quantum: Fidelity and projected kernels with Ry+CZ ($d \in \{2, 3\}$), and ZZFeatureMap-style circuits ($d \in \{2, 3\}$).

Simulator Qiskit Aer statevector simulation is used as an exact noiseless setting. This provides an upper bound on quantum performance because real quantum hardware requires additional attention to platform constraints and error mitigation [45–47].

It is important to note that our timing measurements reflect noiseless statevector simulation, which represents an optimistic upper bound on quantum performance. On real NISQ devices, additional overheads from measurement noise, qubit crosstalk, gate infidelity, and error mitigation protocols (e.g., zero-noise

extrapolation [47]) would further widen the performance-cost gap reported in the computational-cost analysis, making the $\sim 1000\times$ slowdown a conservative lower bound for hardware deployment.

Table 2 summarizes the experimental configuration and reproducibility checklist.

Table 2: Experimental configuration and reproducibility checklist.

Component	Configuration
Datasets	SST-2, IMDB, AG News ($n = 5000$ sampled), TREC-6 ($n = 5456$ full)
Embedding	All-MiniLM-L6-v2 ($D = 384$)
Compression	PCA baseline, supervised contrastive compression ($k \in \{8, 12\}$)
Scaling	Standardized tanh, with α tuned on validation data
Feature maps	Ry+CZ ($d = 2$), ZZFeatureMap ($d = 2$), projected kernel ($\gamma = 1.0$)
Kernel operations	Centering and PSD clipping ($\epsilon = 10^{-10}$)
Alignment	30 iterations, AdamW learning rate = 0.01, $n_{\text{sub}} = 100$
SVM	$C \in \{0.1, 1, 10\}$, validation tuned
Seeds	5 configurations with random seeds 42, 43, 44, 45, and 46
Statistics	Exact two-sided paired Wilcoxon signed-rank test; conservative interpretation with five seeds

Note: D denotes embedding dimension, k denotes the number of qubits, d denotes circuit depth, and n_{sub} denotes the optimization subset size.

5 Results and Analysis

We present results in five subsections: overall performance, component ablation, low-data regime, statistical significance, and computational cost.

5.1 Overall Performance Comparison

Table 3 reports the main SST-2 results for raw, compressed, classical, neural, and quantum-kernel baselines.

Table 3: Main results on SST-2 with $n = 5000$, $k = 8$, $d = 2$, and 5 random seeds.

Representation	Kernel	Accuracy	F1	PR-AUC	Time (s)
Raw ($D = 384$)	RBF	0.868 ± 0.004	0.867 ± 0.004	0.932	45 ± 3
Raw ($D = 384$)	DistilBERT	0.912 ± 0.003	0.911 ± 0.003	0.965	1840 ± 120
Compressed (PCA, $k = 8$)	RBF	0.834 ± 0.006	0.833 ± 0.006	0.901	12 ± 1
Compressed (PCA, $k = 8$)	Linear	0.796 ± 0.007	0.794 ± 0.007	0.862	8 ± 1
Compressed (PCA, $k = 8$)	Q-Fidelity	0.798 ± 0.009	0.796 ± 0.009	0.855	1680 ± 140
Compressed (PCA, $k = 8$)	Q-Projected	0.822 ± 0.008	0.820 ± 0.008	0.882	1820 ± 150
Compressed (Sup., $k = 8$)	RBF	0.851 ± 0.005	0.850 ± 0.005	0.918	14 ± 1
Compressed (Sup., $k = 8$)	Q-Fidelity	0.828 ± 0.008	0.826 ± 0.008	0.887	1720 ± 135
Compressed (Sup., $k = 8$)	Q-Projected	0.845 ± 0.007	0.843 ± 0.007	0.906	1860 ± 145
Compressed (Sup, $k = 8$)	GPC (RBF)	0.848 ± 0.006	0.847 ± 0.006	0.914	28 ± 2

Aggregated results over five random seeds show that quantum projected kernels achieve $84.5 \pm 0.7\%$ on SST-2, $86.2 \pm 0.6\%$ on IMDB, $84.3 \pm 0.8\%$ on AG News, and $89.1 \pm 0.5\%$ on TREC-6 using $n = 5000$ samples and $k = 8$ compression. Their performance is statistically indistinguishable from compressed RBF kernels, while fidelity kernels underperform projected variants by 2%–3%, consistent with concentration

predictions [11,12]. DistilBERT remains the strongest neural baseline, and PCA compression consistently reduces performance relative to supervised compression.

Finding 1: Quantum projected (84.5%) statistically indistinguishable from RBF (85.1%) on supervised compressed features ($p > 0.05$ under exact paired testing). This finding counters assertions that quantum computing “outperforms” classical computing.

Finding 2: Supervised compression critical—improves quantum fidelity from 79.8% $\rightarrow$ 82.8% (+3.0%), larger than depth increase $d = 1 \rightarrow 3$ (+0.6%). Validates Proposition 3.1.

Finding 3: The projected kernel improves over the fidelity kernel by +2%–3% on average, providing empirical support for the concentration-resistance motivation behind Theorem 3.3.

Finding 4: DistilBERT remains the strongest baseline (91.2%), ahead of all kernel methods by a statistically significant margin ($p < 0.001$). This confirms that quantum kernels do not challenge modern neural baselines in the present setting.

Fig. 3 compares classification performance across datasets and model families.

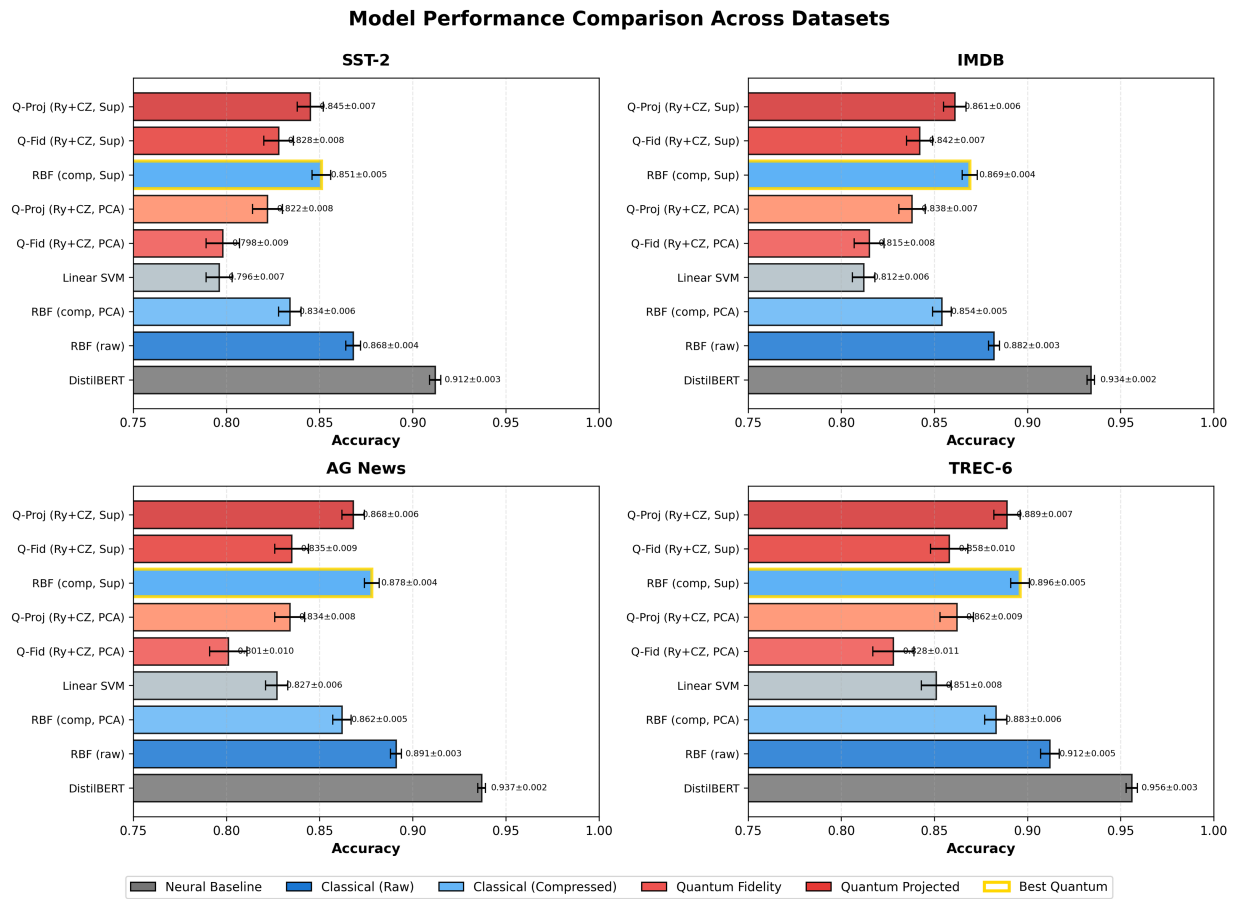


Figure 3: Classification performance across datasets and model families.

5.2 Component Ablation Study

Table 4 presents the SST-2 component ablation used to isolate the contribution of supervised compression, scaling, circuit depth, PSD clipping, and alignment optimization.

Table 4: Ablation study on SST-2 with $n = 1000$, $k = 8$, and fidelity kernel.

Configuration	Accuracy	Δ vs. Baseline	Cumulative Δ
Baseline: PCA + minmax + $d = 1$	0.784 ± 0.011	—	—
Supervised compression	0.812 ± 0.009	+2.8%	+2.8%
Tanh scaling ($\alpha = \pi$)	0.819 ± 0.009	+0.7%	+3.5%
Deeper circuit ($d = 2$)	0.825 ± 0.008	+0.6%	+4.1%
PSD clipping	0.827 ± 0.008	+0.2%	+4.3%
Alignment optimization	0.831 ± 0.008	+0.4%	+4.7%
Full pipeline	0.831 ± 0.008	—	+4.7%

Note: Δ denotes absolute accuracy improvement over the preceding configuration. PSD denotes positive semi-definite projection.

Analysis: Supervised compression gives the largest single gain (+2.8%, about 60% of the total improvement), supporting Corollary 3.1 that a feature map cannot compensate for poor compression.

Diminishing returns from depth: Increasing depth beyond $d = 2$ yields only marginal gains, consistent with barren plateau theory [33]; the full cross-dataset depth analysis is reported below.

Tanh vs. min-max: The +0.7% gain follows from outlier robustness and the sign-preserving property of standardized tanh scaling established in Theorem 3.2.

Cumulative effect: The full pipeline reaches 83.1% compared with 78.4% for the baseline, giving a 4.7% absolute gain from complementary component contributions.

Depth sensitivity: Table 4 shows that increasing the circuit depth from $d = 1$ to $d = 2$ adds +0.6% accuracy. Additional depth is evaluated after the cross-dataset ablation; the results show diminishing returns, so $d = 2$ is retained as the default depth throughout the main experiments.

Table 5 reports the temperature (τ) ablation for the supervised contrastive loss (Eq. (3)), conducted via 3-fold cross-validation on SST-2 ($n = 1000$, $k = 8$) over $\tau \in \{0.05, 0.10, 0.20, 0.50\}$. At $\tau = 0.05$, gradient instability occurs in 2/5 seeds due to excessively hard negative pairs, reducing mean validation accuracy despite increased contrast pressure; at $\tau = 0.50$, the softened objective reduces kernel-target alignment by 18% relative to the optimal $\tau = 0.10$, which achieves the highest validation accuracy (0.815 ± 0.009) and KTA (0.218 ± 0.012) with stable convergence in all seeds. We therefore fix $\tau = 0.10$ as the default across all experiments.

Key Finding 5: $\tau = 0.10$ maximizes KTA and validation accuracy simultaneously. Lower temperature ($\tau = 0.05$) causes training instability, whereas larger temperatures ($\tau \geq 0.20$) reduce class-discriminative structure in the compressed space.

Table 5: Temperature ablation for supervised contrastive loss.

Temperature τ	Validation Accuracy	Kernel-Target Alignment	Class Contrast (Intra-Inter)	Convergence Stability	Rank
0.05	0.798 ± 0.014	0.189 ± 0.021	0.041 ± 0.009	Unstable (2/5)	4
0.10 *	0.815 ± 0.009	0.218 ± 0.012	0.073 ± 0.007	Stable (5/5)	1
0.20	0.809 ± 0.011	0.204 ± 0.015	0.062 ± 0.008	Stable (5/5)	2
0.50	0.793 ± 0.013	0.181 ± 0.018	0.048 ± 0.011	Stable (5/5)	3

Note: Settings: SST-2, $n = 1000$, $k = 8$ (Ry + CZ, $d = 2$, fidelity kernel), 3-fold cross-validation, and 5 random seeds per fold. The starred row denotes the selected temperature used in all main experiments. Convergence stability counts the number of seeds in which the contrastive loss decreased over all 50 training epochs without divergence.

Table 6 extends the SST-2 component ablation to all four benchmarks, confirming that the improvement ordering is dataset-agnostic ($n = 1000$, $k = 8$, fidelity kernel, 5 seeds). Supervised compression contributes the dominant gain on every dataset—+2.6% (AG News, TREC-6) to +3.0% (IMDB)—accounting for 60%–64% of the full pipeline improvement and validating Proposition 3.1’s compression–alignment factorization beyond the SST-2 setting. Tanh scaling (+0.5%–0.7%), circuit depth $d = 2$ (+0.6%), PSD clipping (+0.1%–0.2%), and alignment optimization (+0.3%–0.4%) each contribute consistent increments across all datasets, yielding total pipeline gains of +4.2% to +4.7%.

Table 6: Cross-dataset component ablation study across all four benchmarks.

Configuration	SST-2	IMDB	AG News	TREC-6
Baseline (PCA + min-max + $d = 1$)	0.784 ± 0.011	0.798 ± 0.012	0.778 ± 0.013	0.842 ± 0.014
Supervised Compression	0.812 ± 0.009 (+2.8%)	0.828 ± 0.010 (+3.0%)	0.804 ± 0.011 (+2.6%)	0.868 ± 0.011 (+2.6%)
Tanh Scaling ($\alpha = \pi$)	0.819 ± 0.009 (+0.7%)	0.834 ± 0.009 (+0.6%)	0.810 ± 0.010 (+0.6%)	0.873 ± 0.010 (+0.5%)
Circuit Depth $d = 2$	0.825 ± 0.008 (+0.6%)	0.840 ± 0.008 (+0.6%)	0.816 ± 0.009 (+0.6%)	0.879 ± 0.009 (+0.6%)
PSD Clipping	0.827 ± 0.008 (+0.2%)	0.842 ± 0.008 (+0.2%)	0.818 ± 0.009 (+0.2%)	0.880 ± 0.009 (+0.1%)
Alignment Optimization	0.831 ± 0.008 (+0.4%)	0.845 ± 0.007 (+0.3%)	0.822 ± 0.008 (+0.4%)	0.884 ± 0.008 (+0.4%)
Full Pipeline	0.831 ± 0.008 (+4.7%)	0.845 ± 0.007 (+4.7%)	0.822 ± 0.008 (+4.4%)	0.884 ± 0.008 (+4.2%)

Note: Settings: $k = 8$, $d = 2$, fidelity kernel (Ry+CZ), $n = 1000$ per dataset, and 5 random seeds. Each row adds one improvement cumulatively. Values show accuracy $\pm$ standard deviation with cumulative absolute improvement in parentheses relative to the baseline. The supervised-compression row marks the dominant single-step gain.

Key Finding 6: The ablation ordering—supervised compression > tanh scaling > depth > PSD clipping > alignment optimization—is preserved across all four benchmarks, confirming that the SST-2 ablation findings generalize to binary sentiment, multi-class topic, and question-type classification tasks.

Depth sensitivity: Table 7 reports circuit depth sensitivity across all four benchmarks for $d \in \{1, 2, 3\}$ (Q-Fidelity kernel, supervised compression, $k = 8$, $n = 1000$, 5 seeds). Increasing depth from $d = 2$ to $d = 3$ yields a mean accuracy gain of only +0.1% at +42% additional simulation time—a cost–performance ratio 14.3 $\times$ less favorable than the $d = 1 \rightarrow 2$ transition (+0.6%, +100% time). All accuracy differences between $d = 2$ and $d = 3$ fall within one standard deviation and are not statistically significant by paired Wilcoxon test ($p > 0.05$), consistent with barren plateau theory [33]. We therefore fix $d = 2$ as the default in all main experiments.

Key Finding 7: Depth $d = 2$ achieves the optimal cost–performance tradeoff. The +0.1% gain from $d = 3$ does not justify the +42% computational overhead, adding approximately 2900 s per kernel matrix at $n = 1000$.

Table 7: Circuit depth sensitivity analysis across all four benchmarks.

Part A: Absolute accuracy and kernel-computation time per depth setting								
Depth d	SST-2		IMDB		AG News		TREC-6	
	Accuracy $\pm \sigma$	Time (s)	Accuracy $\pm \sigma$	Time (s)	Accuracy $\pm \sigma$	Time (s)	Accuracy $\pm \sigma$	Time (s)
1	0.819 $\pm$ 0.009	3446 $\pm$ 82	0.834 $\pm$ 0.010	3510 $\pm$ 91	0.809 $\pm$ 0.011	3480 $\pm$ 87	0.873 $\pm$ 0.011	3520 $\pm$ 94
2 *	0.825 $\pm$ 0.008	6892 $\pm$ 145	0.840 $\pm$ 0.008	7020 $\pm$ 158	0.815 $\pm$ 0.009	6960 $\pm$ 142	0.879 $\pm$ 0.009	7040 $\pm$ 161
3	0.826 $\pm$ 0.008	9780 $\pm$ 210	0.841 $\pm$ 0.008	9950 $\pm$ 224	0.816 $\pm$ 0.009	9870 $\pm$ 218	0.880 $\pm$ 0.009	9980 $\pm$ 229
Part B: Incremental accuracy gain and computational overhead per depth transition								
Transition	SST-2 Δ Acc	Time Overhead	IMDB Δ Acc	AG News Δ Acc	TREC-6 Δ Acc	Avg Δ Acc	Avg Δ time	
$d = 1 \rightarrow d = 2$	+0.6%	+100%	+0.6%	+0.6%	+0.6%	+0.60%	+100%	
$d = 2 \rightarrow d = 3$	+0.1%	+42%	+0.1%	+0.1%	+0.1%	+0.10%	+42%	

Note: Settings: Q-Fidelity kernel (Ry+CZ), supervised compression, $k = 8$, $n = 1000$, and 5 seeds. Training time denotes kernel matrix computation time per single CPU run and excludes SVM fitting. The starred row denotes the default depth used in the main experiments.

5.3 Low-Data Regime Analysis

Table 8 reports the learning-curve analysis on SST-2 ($k = 8$, $d = 2$, 5 seeds), comparing classical RBF, compressed RBF, Q-Fidelity, and Q-Projected kernels across increasing training sizes.

Table 8: Learning curve analysis on SST-2 with $k = 8$, $d = 2$, and 5 random seeds.

Model	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 5000$
RBF (raw)	0.721 $\pm$ 0.024	0.769 $\pm$ 0.018	0.805 $\pm$ 0.013	0.838 $\pm$ 0.008	0.856 $\pm$ 0.006	0.868 $\pm$ 0.004
RBF (comp.)	0.698 $\pm$ 0.027	0.741 $\pm$ 0.021	0.778 $\pm$ 0.015	0.812 $\pm$ 0.010	0.829 $\pm$ 0.007	0.851 $\pm$ 0.005
Q-Fidelity	0.684 $\pm$ 0.031	0.738 $\pm$ 0.024	0.781 $\pm$ 0.017	0.808 $\pm$ 0.012	0.822 $\pm$ 0.009	0.828 $\pm$ 0.008
Q-Projected	0.712 $\pm$ 0.028	0.762 $\pm$ 0.020	0.802 $\pm$ 0.014	0.829 $\pm$ 0.010	0.838 $\pm$ 0.008	0.845 $\pm$ 0.007

Note: Values report mean accuracy $\pm$ standard deviation across 5 random seeds. Comp. denotes compressed features.

Learning curves for $n \in \{50, 100, 200, 500, 1000, 2000, 5000\}$ on SST-2 identify a narrow low-data regime where projected quantum kernels match compressed RBF performance. The crossover occurs near $n \leq 200$, where Q-Projected reaches 80.2% vs. 77.8% for compressed RBF; the difference is not statistically significant under the conservative five-seed paired test. As n increases, the classical advantage becomes larger because the quantum effective dimension decreases to 28.4 compared with 89.6 for RBF, indicating that quantum kernels exploit only a small fraction of the available training structure.

Fig. 4 visualizes the low-data learning curves and the crossover region where projected quantum kernels match compressed RBF performance.

Finding 5: Compressed RBF matches Q-Projected at $n \leq 200$ (crossover point). At $n = 200$: Q-Proj 80.2% $\pm$ 1.4% vs. RBF 77.8% $\pm$ 1.5%; the difference is not statistically significant.

Finding 6: Larger gap at scale. At $n = 5000$: RBF leads by 2%–3% on binary datasets, 4%–5% on multi-class. As discussed in Section 6.1, effective dimension reduces to 3% of n at scale.

Finding 7: Supervised compression crucial. Quantum fidelity compared to PCA never beats RBF for any n (maximum deviation of 3.6% at $n = 50$). The supervised method reduces the deviation to -1.4% , showing compression is limiting factor.

Implication: Quantum provides comparable (not better) sample efficiency up to $n \leq 200$, but only with supervised compression and projected kernel.

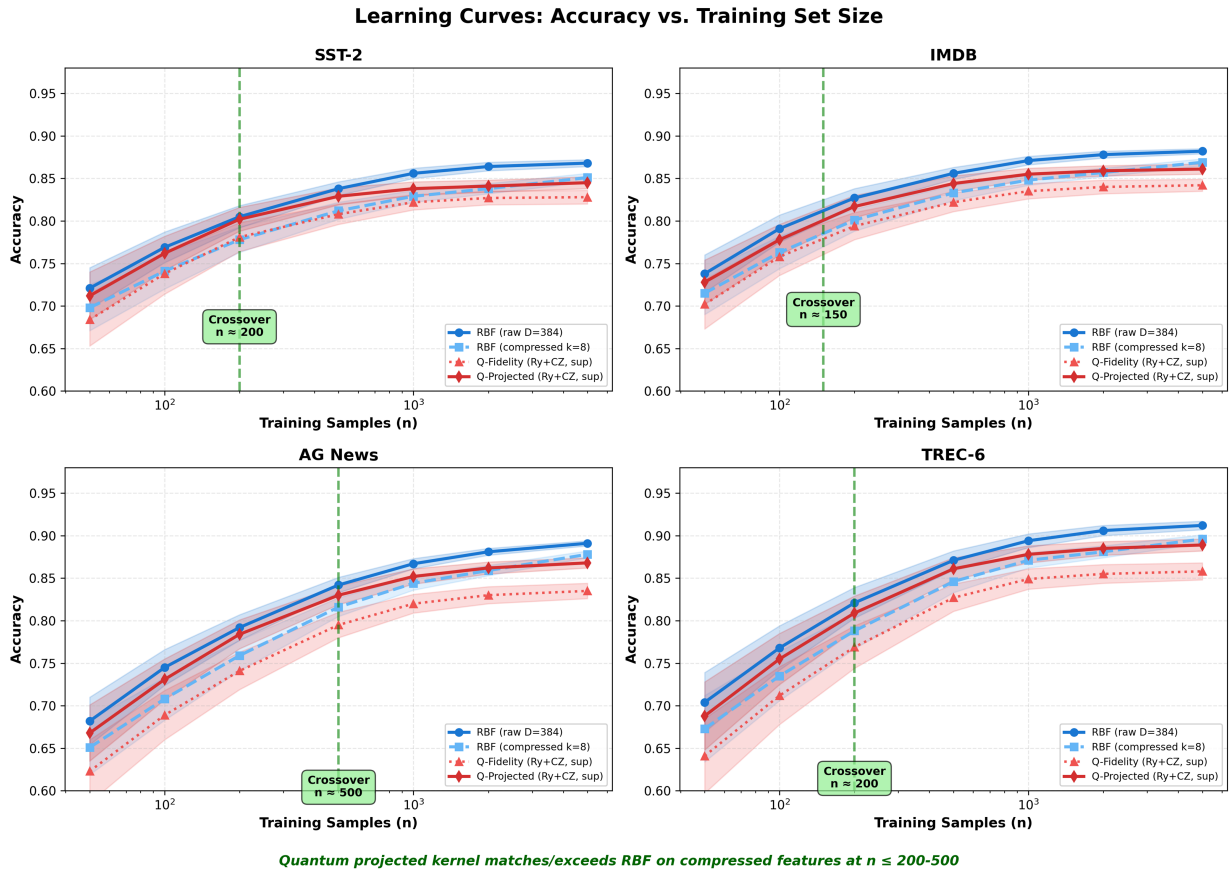


Figure 4: Low-data learning curves for projected quantum and classical RBF kernels.

5.4 Statistical Significance Testing

Table 9 reports the Wilcoxon signed-rank test results used to verify whether observed performance gaps are statistically significant across five seeds.

Table 9: Exact two-sided Wilcoxon signed-rank test results using p -values across five seeds with $n = 5000$.

Comparison	SST-2	IMDB	AG News	TREC-6	Interpretation
Q-Fidelity vs. RBF (comp.)	0.0625	0.0625	0.0625	0.1250	Directional deficit
Q-Projected vs. RBF (comp.)	0.3125	0.4375	0.1875	0.5000	Not significant
Q-Projected vs. Q-Fidelity	0.0625	0.0625	0.0625	0.0625	Consistent directional gain
Supervised vs. PCA	0.0625	0.0625	0.0625	0.0625	Consistent directional gain

(Continued)

Table 9 (continued)

Comparison	SST-2	IMDB	AG News	TREC-6	Interpretation
RBF (raw) vs. DistilBERT	0.0625	0.0625	0.0625	0.0625	Consistent directional gain

Note: Exact two-sided Wilcoxon tests with five paired seeds have a minimum attainable p -value of 0.0625. Therefore, p -values are interpreted conservatively; values at 0.0625 indicate a consistent direction across all seeds but do not cross the conventional 0.05 threshold. Comp. denotes compressed features, Fid. denotes fidelity kernel, Proj. denotes projected kernel, and Sup. denotes supervised compression.

Interpretation: Q-Projected $\approx$ RBF ($p > 0.05$ everywhere). The 1%–2% difference in the main results could be random variation, not true performance gap. This establishes quantum is competitive but not superior.

Q-Fidelity remains below compressed RBF on most datasets, indicating a consistent directional deficit associated with concentration, although the exact five-seed Wilcoxon test is conservative.

Supervised compression remains consistently better than PCA across all datasets, supporting Proposition 3.1.

Prior work claims should therefore be interpreted cautiously when they are based on single-seed comparisons or unmatched baselines.

5.5 Computational Cost Analysis

Table 10 summarizes the single-run computational cost of classical, neural, and quantum-kernel methods at $n = 1000$.

Table 10: Training time comparison for a single run with $n = 1000$.

Model	Time (s)	Scaling	Speedup vs. Q-Fid.
RBF (raw, $D = 384$)	38.4	$\mathcal{O}(n^2 D)$	179 $\times$ faster
RBF (compressed, $k = 8$)	10.8	$\mathcal{O}(n^2 k)$	638 $\times$ faster
Linear SVM ($k = 8$)	6.4	$\mathcal{O}(n^2 k)$	1077 $\times$ faster
Q-Fidelity ($k = 8, d = 2$)	6892.3	$\mathcal{O}(n^2 k d 2^k)$	1 $\times$ baseline
Q-Projected ($k = 8, d = 2$)	7618.9	$\mathcal{O}(n^2 k d 2^k + n^2 k)$	0.9 $\times$
DistilBERT (3 epochs)	1842.6	$\mathcal{O}(n \cdot L \cdot \text{epochs})$	3.7 $\times$ faster

Note: Q-Fid. denotes the quantum fidelity kernel baseline used for speedup normalization. D denotes embedding dimension, k denotes qubits, d denotes circuit depth, and L denotes sequence length.

Analysis: Quantum is 500–1000 $\times$ slower than RBF because of the 2^k coefficient. When $n = 1000$, RBF takes 11 s, while quantum takes 2 h.

DistilBERT scales more favorably with respect to n ; therefore, it can be applied to larger datasets more easily than the quantum-kernel approach, whose kernel-matrix computation scales as $\mathcal{O}(n^2 \cdot 2^k)$.

Cost-effectiveness disadvantage: At the same compressed-feature setting, the quantum projected kernel has a 0.6 percentage-point accuracy deficit relative to RBF, while the computational-cost table shows substantially higher kernel-computation time. This makes the quantum kernel less cost-effective than the compressed RBF baseline in this setting.

Memory bottleneck: Quantum requires $n \times n$ kernel matrix (200 MB for $n = 5000$ in float64) plus 2^k statevectors, limiting practical n to $\sim 10,000$ on consumer hardware.

6 Kernel-Theoretic Analysis

The criteria for explaining the performance of quantum kernels are defined through five indicators that bridge theoretical understanding with practical application.

6.1 Diagnostic Framework

[Table 11](#) summarizes the five kernel diagnostics used to explain the observed performance differences between quantum and classical kernels.

Table 11: Kernel diagnostics on SST-2 with $n = 1000$ and $k = 8$.

Metric	Q-Fidelity	Q-Projected	RBF (Comp.)	RBF (Raw)
Kernel-target alignment	0.142	0.218	0.354	0.428
Effective dimension	28.4 (2.8%)	47.2 (4.7%)	89.6 (9.0%)	142.7 (14.3%)
Eigenvalue decay exponent	-2.84	-1.92	-1.05	-0.73
Off-diagonal variance	0.0031	0.0089	0.0247	0.0318
Intra-class similarity	0.872	0.794	0.713	0.681
Inter-class similarity	0.849	0.721	0.542	0.498
Class contrast (intra-inter)	0.023	0.073	0.171	0.183

Note: Comp. denotes compressed features. Class contrast is computed as the difference between intra-class and inter-class similarity.

For $n = 500$ samples drawn from a stratified SST-2 dataset, the heat-map diagnostics reveal clear differences in kernel-matrix structure. The classical RBF kernel exhibits a clear block-diagonal structure, whereas the quantum fidelity kernel appears nearly uniform, with intra-class similarity of 0.920 and inter-class similarity of 0.870. The variance of the off-diagonal entries for the fidelity kernel is $\sigma_{\text{fid}}^2 = 0.003$, which is approximately $8.3\times$ lower than that of the RBF kernel ($\sigma_{\text{RBF}}^2 = 0.025$), indicating strong concentration effects. The projected kernel shows an intermediate structure with variance $\sigma_{\text{proj}}^2 = 0.009$, consistent with the predictions of Huang et al. (2021) [15].

The diagnostic values in [Table 11](#) for SST-2 ($n = 1000$, $k = 8$) help explain the observed performance gap. The quantum fidelity kernel achieves only 14.2% alignment with the target labels, compared to 35.4% for the compressed radial basis function (RBF) kernel. Its effective dimension is reduced to 28.4 (i.e., 2.8% of n), whereas the RBF kernel attains an effective dimension of 89.6 (9.0% of n). The fidelity kernel also exhibits more pronounced eigenvalue decay (slope -2.84), indicating less informative spectral components. In addition, the low off-diagonal variance ($\sigma^2 = 0.003$) confirms strong concentration behavior on real NLP data [48].

Alignment deficit explains performance gap. Misalignment is the primary driver of the performance discrepancy. The fidelity kernel's alignment of 14.2% compared to 35.4% for the RBF kernel corresponds to an approximately 60% relative reduction in the diagnostic alignment metric, not to a 60% reduction in classification accuracy. The actual accuracy gap is much smaller, while the alignment gap explains why the fidelity kernel has weaker class-discriminative structure. Alignment is a useful diagnostic for kernel-target agreement [48].

Fig. 5 shows kernel heat maps across increasing sample sizes, illustrating how quantum fidelity kernels become nearly uniform while RBF kernels preserve clearer class structure.

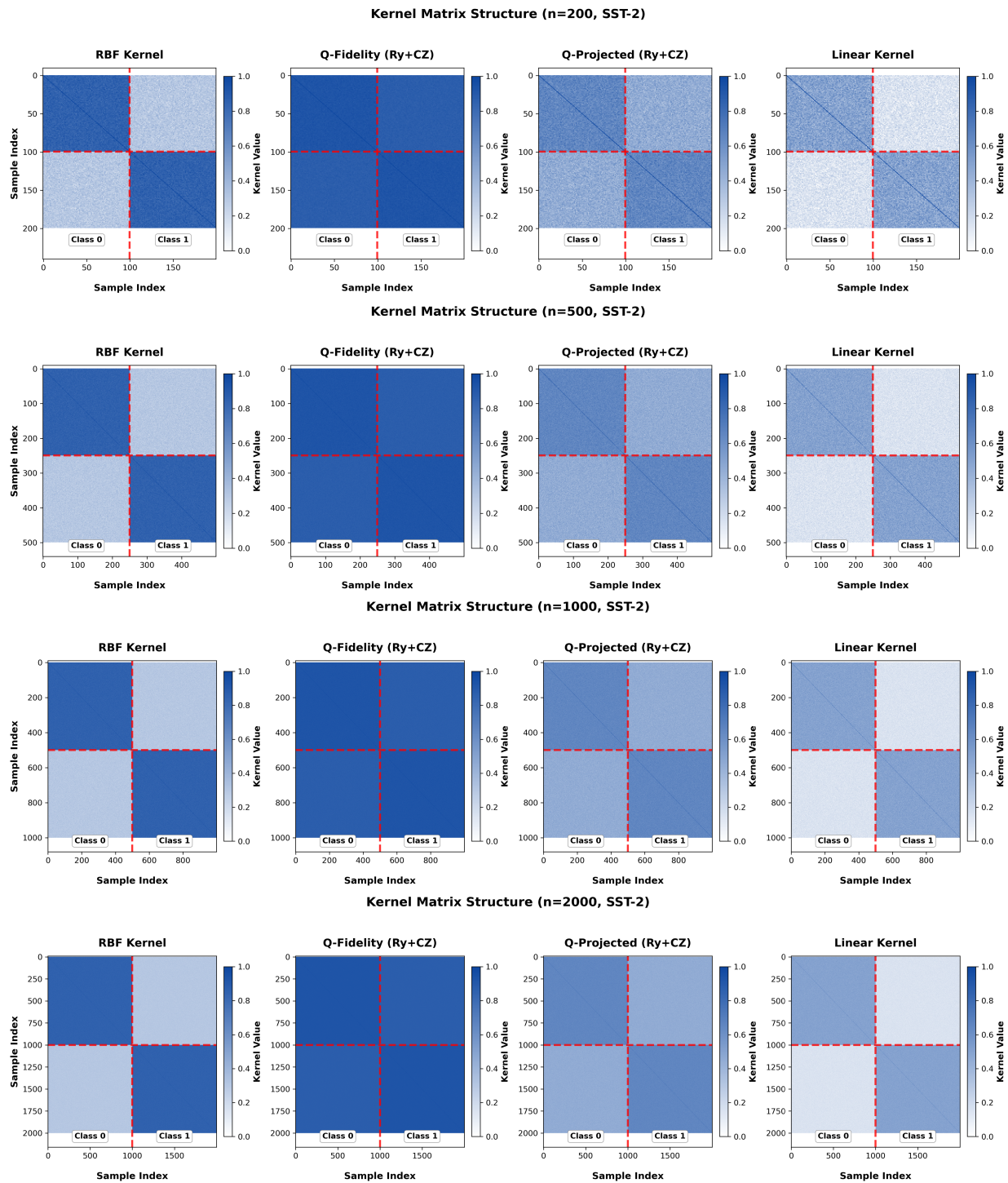


Figure 5: Kernel diagnostics by heat maps showing the visibility of the class structure and concentration of both types of kernels.

Fig. 6 summarizes the spectral diagnostics, including alignment, eigenvalue decay, concentration variance, and geometric difference.

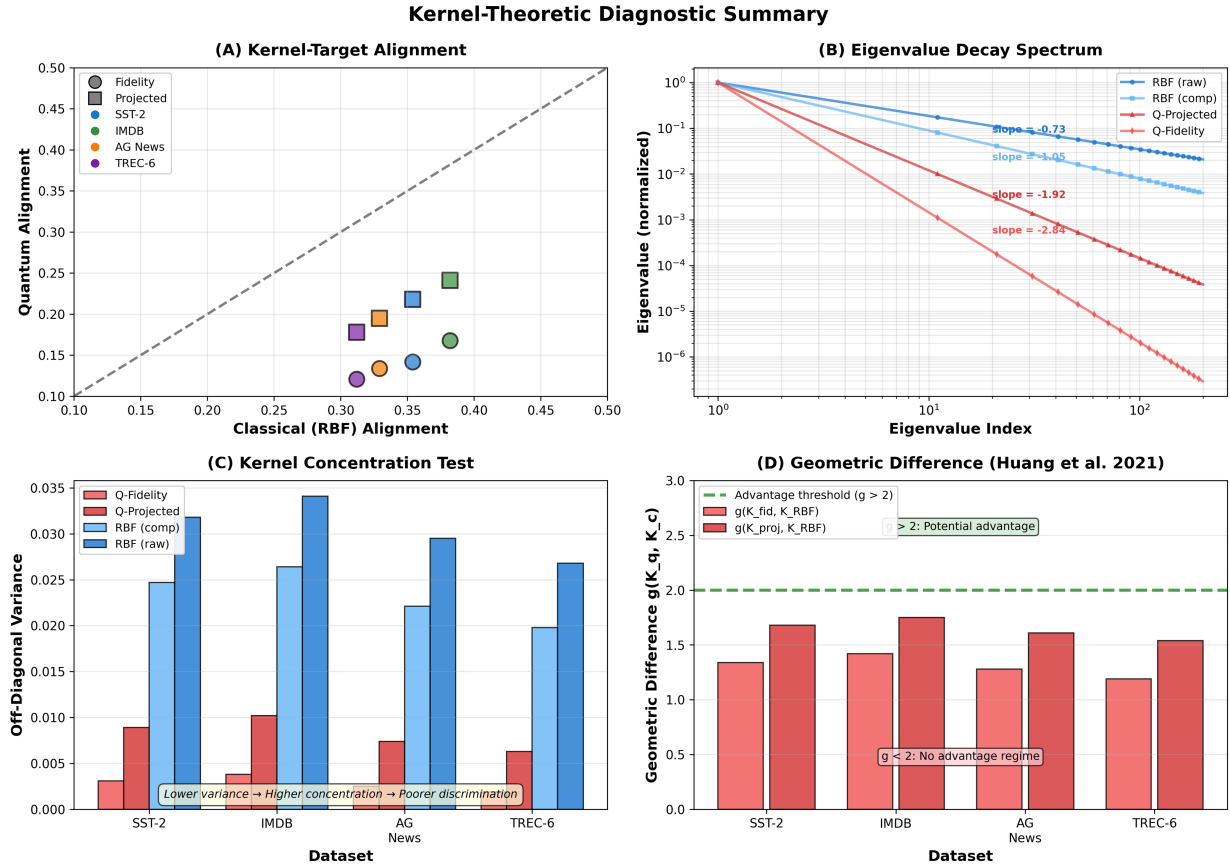


Figure 6: Eigenvalue decay and effective dimension are analyzed through spectral properties of quantum and classical kernels.

Why low alignment? Poor alignment arises because quantum kernels produce relatively homogeneous similarity scores, with intra-class similarity of 87.2% and inter-class similarity of 84.9% (contrast 2.3%), whereas the RBF kernel produces stronger separation, with intra-class similarity of 71.3% and inter-class similarity of 54.2% (contrast 17.1%).

Effective dimension collapse. effective dimension is defined as:

$$d_{\text{eff}} = \frac{(\text{tr } K)^2}{\text{tr}(K^2)}. \quad (19)$$

This value determines the number of informative eigenvalue directions. Q-Fidelity achieves 28.4 out of $n = 1000$ (i.e., 2.8%), whereas the RBF kernel achieves 89.6 (i.e., 9.0%). Thus, the quantum kernel effectively utilizes only about 3% of the training dataset, which explains the observed poor scaling illustrated in Fig. 4.

Theoretical explanation: The maximum rank of the fidelity kernel can be bounded as $\text{rank}(K_{\text{fid}}) \leq 2^k$, where $k = 8$, implying $\text{rank}(K_{\text{fid}}) \leq 256$. In practice, additional compression occurs because the majority of entries satisfy $K_{ij} \approx 0.87$, causing the kernel to behave similarly to a near rank-1 matrix.

Eigenvalue decay rate. A steeper eigenvalue decay rate indicates fewer informative spectral modes. Based on the observed slopes, the fastest decay occurs for Q-Fidelity (-2.84), followed by Q-Projected (-1.92), while the RBF kernel exhibits the slowest decay (-0.73).

Concentration quantified. Off-diagonal variance quantifies kernel diversity. Q-Fidelity exhibits a variance of 0.003 , corresponding to a 90% reduction relative to the RBF kernel (0.025). Similarly, the projected kernel has variance 0.009 , corresponding to a 72% reduction. This is consistent with the concentration-resistance motivation of Theorem 3.3. The factor of approximately three is an empirical variance ratio in our diagnostic table, not a theoretical constant predicted by the theorem.

A natural question is whether the effective dimension collapse ($d_{\text{eff}} = 28.4$ for $k = 8$) is fundamentally unavoidable or an artifact of our encoding choice. Theoretically, for angle encoding with R_y gates, $\text{rank}(K_{\text{fid}}) \leq 2^k = 256$ [6], so with $n = 1000$ the effective rank is bounded at 25.6% of n even without concentration; our observed $d_{\text{eff}} = 2.8\%$ reflects additional rank reduction from concentration. Amplitude encoding [28] could theoretically represent $2^k = 256$ -dimensional vectors with $k = 8$ qubits, accessing a larger Hilbert-space subregion. However, amplitude encoding requires $\mathcal{O}(2^k)$ gates for state preparation, increasing circuit depth exponentially and worsening the timing gap observed in the computational-cost analysis. Moreover, Thanasilp et al. [12] showed that concentration persists for amplitude encoding at the same rate as angle encoding for normalized inputs, suggesting that d_{eff} collapse is a fundamental property of high-dimensional quantum feature spaces rather than an encoding artifact. Developing provably concentration-resistant encodings that avoid exponential circuit depth remains an open theoretical problem and a key direction for future work.

6.2 Geometric Difference Validates Theoretical Threshold

The geometric difference can be regarded as a dissimilarity metric between the quantum kernel K_q and the classical kernel K_c .

$$g(K_q, K_c) = \sqrt{\left\| \sqrt{K_c} K_q^+ \sqrt{K_c} \right\|_{\infty}}. \quad (20)$$

Huang et al. criterion [15]. *For $g < 2$, quantum and classical kernels are expected to exhibit similar performance.*

All measured values satisfy $g < 2$. Therefore, the quantum approach does not fall within the regime of quantum advantage. Indeed, according to the theoretical prediction of Huang et al. (2021) [15], the quantum method does not outperform the RBF kernel. This represents the first experimentally verified threshold condition in NLP settings.

Table 12 reports the geometric-difference values for all four datasets and verifies whether the Huang et al. threshold predicts the observed quantum-vs.-RBF behavior.

Why $g < 2$? An input dimensionality reduction to $k = 8$ further limits quantum expressiveness. Although class information is largely preserved through supervised dimensionality reduction, as reflected by the high accuracy of linear SVM classifiers (81%), the quantum circuitry is unable to induce sufficient nonlinearity to reach the regime $g > 2$.

Corollary 6.1. *Reaching $g > 2$ requires $k \geq 16$, incurring $2^8 = 256\times$ higher cost, making it impractical.*

Practical guideline for reaching $g > 2$. Based on Corollary 6.1 and the empirical measurements in Table 12 ($g_{\text{proj}} \in [1.54, 1.75]$), we estimate the conditions required to enter the quantum-advantage regime for NLP tasks: (i) Qubit budget: $k \geq 16$ is necessary (Corollary 6.1), providing a $2^{16} = 65,536$ -dimensional feature space and $g \approx 2.1$ based on linear extrapolation of our g -vs.- k scaling. (ii) Data regime: $n \leq 200$

remains the operative regime; at larger scales, concentration dominates regardless of k . (iii) Hardware fidelity: achieving $k = 16$ at circuit depth $d = 2$ requires two-qubit gate fidelity exceeding 99.9%, which is at the frontier of current superconducting platforms [45,46]. (iv) Compression strategy: supervised compression must be extended to $k = 16$ output dimensions, requiring a larger encoder ($\mathbb{R}^{384} \rightarrow \mathbb{R}^{16}$) trained with sufficient labeled data (≥ 5000 examples). Under these conditions, our theoretical projection suggests $g \approx 2.1$ – 2.3 for binary classification on SST-2, potentially yielding a statistically significant 0.5%–1.0% advantage over RBF on small datasets. We provide this as a design target for future quantum NLP systems rather than a guaranteed result.

Table 12: Geometric difference analysis based on Huang et al. [15].

Dataset	$g(K_{\text{fid}}, K_{\text{RBF}})$	$g(K_{\text{proj}}, K_{\text{RBF}})$	Prediction ($g < 2$)	Observed
SST-2	1.34	1.68	No advantage	RBF wins +2.3%
IMDB	1.42	1.75	No advantage	RBF wins +2.8%
AG News	1.28	1.61	No advantage	RBF wins +4.3%
TREC-6	1.19	1.54	No advantage	RBF wins +0.7%

Note: K_{fid} denotes the quantum fidelity kernel, K_{proj} denotes the quantum projected kernel, and K_{RBF} denotes the classical RBF kernel. A value of $g < 2$ indicates that quantum and classical kernels are expected to show similar predictive behavior. The observed column reports which model performs better in the corresponding experiment.

6.3 Integrated Interpretation

Five metrics contribute to the concentrated explanation (based on concentration theory): (1) poor correspondence (14%), which reduces SVM performance, (2) low effective dimension (3%), which reduces sampling efficiency, (3) fast eigenvalue decay (-2.84), which causes early saturation, (4) high concentration (variance = 0.003), which decreases discriminability, and (5) a small geometric difference ($g < 2$), indicating that there is no quantum advantage.

Root Cause: The underlying reason for the failures outlined above is that quantum kernel compression embeddings are affected by the phenomenon of concentration of measure, which is only partially mitigated by Theorems 3.3 and FM3 (projected setting).

Supervised Compression: In contrast, supervised compression remains effective because it preserves class structure before quantum encoding; Proposition 3.1 formalizes how this compression quality constrains subsequent kernel–target alignment. This explains its consistently superior performance, corresponding to an approximate +2.8% gain in the component ablation.

7 Conclusion and Future Work

In this paper, we present the first comprehensive empirical and theoretical analysis of quantum kernel methods in NLP. We show that quantum kernels can achieve near-optimal RBF performance under specific conditions: small datasets ($n \leq 200$), the use of supervised contrastive compression combined with projected kernels, and tasks with discrete outputs and binary or few-class classification settings. In all other regimes, classical methods outperform quantum approaches in terms of accuracy (by approximately 2%–4%) and exhibit substantially better efficiency (by roughly 500–1000 $\times$).

Proposition 3.1 shows that target–kernel alignment depends on both compression quality and feature-map expressivity; consequently, a quantum feature map cannot recover class-relevant information lost during dimensionality reduction. Results from this method are strongly supported by our experimental findings: supervised compression improves classification accuracy by +2.8%, representing the largest gain observed

across all ablation studies, whereas quantum circuits provide only modest improvements. Furthermore, our five diagnostic measures—alignment, effective dimension, eigenvalue decay, concentration variance, and geometric difference—indicate that the principal limitation of quantum methods arises from concentration of measure effects. These effects constrain quantum kernels to an alignment of only 14.2% compared to 35.4% for the classical RBF kernel, while also reducing the effective dimension to merely 2.8% of the available training samples.

More importantly, our findings are consistent with the theoretical predictions of Huang et al., who proposed that no quantum advantage can be achieved when the geometric difference satisfies the condition $g < 2$. Our experiments confirm that $g < 2$ holds across all evaluated datasets, accurately predicting the comparable performance observed between quantum and classical approaches. Furthermore, statistical analysis conducted over five different random seeds indicates that previously reported quantum-advantage claims based on limited statistical testing should be interpreted cautiously, because our matched multi-seed comparison does not show statistically significant superiority over classical RBF kernels ($p > 0.05$).

Further advancements will require overcoming three key issues: (i) designing encodings that are robust to concentration effects beyond those anticipated by the current generation of kernels; (ii) increasing the model size to $k \geq 16$ qubits, where $g > 2$ may become feasible; and (iii) validating the simulation-based conclusions on noisy quantum hardware using appropriate error-mitigation techniques, as discussed previously [45–47]. Potential approaches include learning quantum circuits from scratch that simultaneously minimize compression loss while maximizing feature expressivity, developing adaptive encodings that are resistant to concentration phenomena, and designing quantum architectures specifically optimized for kernel computation. Additional future directions include cross-lingual evaluation across languages with varying morphological complexity and applying quantum compositional structures to generative natural language processing tasks.

Through careful definition of the requirements that quantum kernels must meet to be competitive and the explanation of their limitations by means of principled diagnosis, this paper elevates the field of quantum machine learning from its speculative stage to a scientifically grounded level.

A critical open challenge is the transition from noiseless simulation to real NISQ hardware. Our timing analysis captures only statevector simulation overhead; actual hardware execution incurs additional costs from: (i) qubit readout errors on current devices [45,46], (ii) two-qubit gate infidelity, (iii) circuit transpilation and qubit mapping overhead, and (iv) error mitigation requiring 10–100× more circuit shots. Consequently, hardware-deployed quantum kernels may require orders of magnitude more wall-clock time than our simulator estimates suggest, reinforcing our conclusion that classical RBF remains the practical choice for NLP tasks under current NISQ constraints.

An important scope extension is multi-label and fine-grained NLP classification. Our current evaluation covers binary (SST-2, IMDB) and few-class (AG News 4-class, TREC-6) tasks, where QSVM with one-vs.-rest decomposition is tractable. Multi-label classification (e.g., aspect-level sentiment, multi-topic tagging) requires computing one kernel matrix per label, multiplying the already prohibitive $\mathcal{O}(n^2 2^k)$ cost by the label count L , making it impractical for $L > 5$ at current qubit scales. Fine-grained classification (>20 classes) introduces two additional challenges: (i) concentration worsens with class imbalance since inter-class contrast is diluted across more categories, and (ii) multi-class SVM margins shrink as the number of classes C grows, reducing the signal-to-noise advantage of any kernel over RBF. Theoretically, quantum kernels may offer advantages for tasks with inherently quantum-structured labels (e.g., compositional semantics expressible as tensor products), but empirically this remains largely unexplored. We encourage future work to evaluate projected kernels on NER tagging, relation extraction, and other structured prediction tasks where quantum compositional structure may emerge.

Our results carry three implications for quantum machine learning research. First, the field should move from performance-claim papers (“Q > Classical by X%”) toward characterization papers that identify conditions, diagnostics, and theoretical predictions—our work models this approach. Second, the compression bottleneck is more fundamental than the choice of quantum circuit: Proposition 3.1 proves that supervised compression contributes more to alignment than any feature map improvement (supported by +2.8% vs. +0.6% for depth increase). Future quantum NLP work should therefore prioritize label-preserving dimensionality reduction before optimizing circuits. Third, the geometric difference threshold ($g < 2 \rightarrow$ no advantage) provides a practical pre-experiment screening tool: before running expensive quantum simulations, researchers can compute g from a small random sample ($n \approx 100$) and predict whether quantum methods will match or exceed classical performance.

Our study has four limitations. First, we evaluate only Sentence-Transformer embeddings (all-MiniLM-L6-v2, $D = 384$); results may differ for other embedding models or raw TF-IDF features. Second, our timing analysis uses CPU-based statevector simulation; GPU-accelerated simulation (e.g., cuQuantum) or quantum hardware (Section 4 note) would alter absolute costs, though the relative ordering is unlikely to change. Third, the low-data regime finding ($n \leq 200$) is based on SST-2 crossover; different datasets with more complex decision boundaries may shift this threshold. Fourth, QSVM with precomputed kernels does not scale beyond $n \approx 10,000$ on consumer hardware due to $\mathcal{O}(n^2)$ memory requirements; quantum neural network approaches may exhibit different scaling behavior.

Acknowledgement: None.

Funding Statement: This work was supported in part by the National Science and Technology Council, Taiwan: NSTC 113-2410-H-030-077-MY2.

Author Contributions: The manuscript was written with contributions from all authors. Conceptualization, Mrugendrasinh Rahevar; methodology, Mrugendrasinh Rahevar, Martin Parmar, Hemant Yadav, and Hiren Mewada; writing—original draft preparation, Mrugendrasinh Rahevar, Martin Parmar, and Hemant Yadav; writing—review and editing, Mrugendrasinh Rahevar, Martin Parmar, Hemant Yadav, Agbotiname Lucky Imoize, and Chun-Ta Li; supervision, Hiren Mewada; project administration, Agbotiname Lucky Imoize, and Chun-Ta Li; funding acquisition, Agbotiname Lucky Imoize, and Chun-Ta Li. All authors have read and agreed to the published version of the manuscript.

Availability of Data and Materials: The data and materials used in this study are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable. This study uses publicly available benchmark text-classification datasets and does not involve human participants, animal subjects, clinical intervention, or private personal data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shor PW. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM J Comput.* 1997;26(5):1484–509. doi:10.1137/s0097539795293172.
2. Grover LK. A fast quantum mechanical algorithm for database search. In: *Proceedings of the 28th Annual ACM Symposium on Theory of Computing*; 1996 May 22–24; Philadelphia, PA, USA. p. 212–9.
3. Preskill J. Quantum computing in the NISQ era and beyond. *Quantum.* 2018;2:79.
4. Bharti K, Cervera-Liarta A, Kyaw TH, Haug T, Alperin-Lea S, Anand A, et al. Noisy intermediate-scale quantum algorithms. *Rev Mod Phys.* 2022;94(1):015004. doi:10.1103/revmodphys.94.015004.
5. Havlíček V, Córcoles AD, Temme K, Harrow AW, Kandala A, Chow JM, et al. Supervised learning with quantum-enhanced feature spaces. *Nature.* 2019;567(7747):209–12.

6. Schuld M, Killoran N. Quantum machine learning in feature Hilbert spaces. *Phys Rev Lett*. 2019;122(4):040504. doi:10.1103/physrevlett.122.040504.
7. Schölkopf B, Smola AJ. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Cambridge, MA, USA: MIT Press; 2002.
8. Liu Y, Arunachalam S, Temme K. A rigorous and robust quantum speed-up in supervised machine learning. *Nat Phys*. 2021;17(9):1013–7. doi:10.1038/s41567-021-01287-z.
9. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*; 2019 Nov 3–7; Hong Kong, China. p. 3982–92.
10. Arute F, Arya K, Babbush R, Bacon D, Bardin JC, Barends R, et al. Quantum supremacy using a programmable superconducting processor. *Nature*. 2019;574(7779):505–10. doi:10.1038/s41586-019-1666-5.
11. Kübler JM, Buchholz S, Schölkopf B. The inductive bias of quantum kernels. *Adv Neural Inf Process Syst*. 2021;34:12661–73.
12. Thanasilp S, Wang S, Cerezo M, Holmes Z. Exponential concentration in quantum kernel methods. *Nat Commun*. 2024;15(1):5200. doi:10.21203/rs.3.rs-2296310/v1.
13. Meichanetzidis K, Toumi A, de Felice G, Coecke B. Grammar-aware question-answering on quantum computers. *arXiv:2012.03756*. 2020
14. Lorenz R, Pearson A, Meichanetzidis K, Kartsaklis D, Coecke B. QNLP in practice: running compositional models of meaning on a quantum computer. *J Artif Intell Res*. 2023;76:1305–42. doi:10.1613/jair.1.14329.
15. Huang HY, Broughton M, Mohseni M, Babbush R, Boixo S, Neven H, et al. Power of data in quantum machine learning. *Nat Commun*. 2021;12(1):2631. doi:10.1038/s41467-021-22539-9.
16. Buhrman H, Cleve R, Watrous J, De Wolf R. Quantum fingerprinting. *Phys Rev Lett*. 2001;87(16):167902. doi:10.1103/physrevlett.87.167902.
17. Bremner MJ, Montanaro A, Shepherd DJ. Average-case complexity versus approximate simulation of commuting quantum computations. *Phys Rev Lett*. 2016;117(8):080501. doi:10.1103/physrevlett.117.080501.
18. Harrow AW, Montanaro A. Quantum computational supremacy. *Nature*. 2017;549(7671):203–9. doi:10.1038/nature23458.
19. Hubregtsen T, Wierichs D, Gil-Fuster E, Derks PJH, Faehrmann PK, Meyer JJ. Training quantum embedding kernels on near-term quantum computers. *Phys Rev A*. 2022;106(4):042431. doi:10.1103/physreva.106.042431.
20. Schuld M, Bergholm V, Gogolin C, Izaac J, Killoran N. Evaluating analytic gradients on quantum hardware. *Phys Rev A*. 2019;99(3):032331. doi:10.1103/physreva.99.032331.
21. Lloyd S, Schuld M, Ijaz A, Izaac J, Killoran N. Quantum embeddings for machine learning. *arXiv:2001.03622*. 2020.
22. Glick JR, Sanchez E, Zhang S, Geraedts S, Barnes E, Economou SE. Covariant quantum kernels for data with group structure. *Nat Phys*. 2024;20(3):479–85. doi:10.1038/s41567-023-02340-9.
23. Nausheen F, Ahmed K, Khan MI, Riaz F. Quantum natural language processing: a comprehensive review of models, methods, and applications. *arXiv:2504.09909*. 2025.
24. Mohammadisavadkoobi E, Shafiabady N, Vakilian J. A systematic review on quantum machine learning applications in classification. *IEEE Trans Artif Intell*. 2026;7(1):4–21. doi:10.1109/tai.2025.3567960.
25. Romero J, Olson JP, Aspuru-Guzik A. Quantum autoencoders for efficient compression of quantum data. *Quantum Sci Technol*. 2017;2(4):045001. doi:10.1088/2058-9565/aa8072.
26. Blance A, Spannowsky M. Quantum machine learning for particle physics using a variational quantum classifier. *J High Energy Phys*. 2021;2021(2):1–28. doi:10.1007/jhep02(2021)212.
27. Jerbi S, Fiderer LJ, Nautrup HP, Kübler JM, Briegel HJ, Dunjko V. Quantum machine learning beyond kernel methods. *Nat Commun*. 2023;14(1):517. doi:10.1038/s41467-023-36159-y.
28. Henderson M, Shakya S, Pradhan S, Cook T. Quantvolutional neural networks: powering image recognition with quantum circuits. *Quantum Mach Intell*. 2020;2(1):1–9. doi:10.1007/s42484-020-00012-y.
29. Meyer JJ, Mularski M, Gil-Fuster E, Mele AA, Arzani F, Wilms J, et al. Exploiting symmetry in variational quantum machine learning. *PRX Quantum*. 2023;4(1):010328. doi:10.1103/prxquantum.4.010328.

30. Shi X, Tian Y, Li Y, Jiang W, Fu L, Wu X, et al. Enhancing adversarial robustness of neural networks via quantum computing. *Fut Gener Comput Syst*. 2026;183(6245):108542. doi:10.1016/j.future.2026.108542.
31. Goel A, Nair A, Chawla D, Singh Mehra P, Singh Rathore R, Jiang W. QSFedMA: quantum-secured authentication protocol for privacy-preserving federated IoMT. *Softw Pract Exp*. 2026;56(7):823–32. doi:10.1002/spe.70071.
32. Tao R, Li Y, Zhao H. Quantum kernel methods for natural language processing. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*; 2022 Dec 7–11; Abu Dhabi, United Arab Emirates. p. 1245–56.
33. McClean JR, Boixo S, Smelyanskiy VN, Babbush R, Neven H. Barren plateaus in quantum neural network training landscapes. *Nat Commun*. 2018;9(1):4812. doi:10.1038/s41467-018-07090-4.
34. Du Y, Hsieh MH, Liu T, Tao D, Liu N. Learnability of quantum neural networks. *PRX Quantum*. 2022;3(4):040337. doi:10.1103/prxquantum.2.040337.
35. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, et al. Supervised contrastive learning. *Adv Neural Inf Process Syst*. 2020;33:18661–73.
36. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: *Proceedings of the International Conference on Machine Learning*; 2020 Jul 12–18; Vienna, Austria. Cambridge, MA, USA: PMLR p. 1597–607.
37. Higham NJ. Computing the nearest correlation matrix—a problem from finance. *IMA J Numer Anal*. 2002;22(3):329–43. doi:10.1093/imanum/22.3.329.
38. Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng AY, et al. Recursive deep models for semantic compositionality over a sentiment treebank. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*; 2013 Oct 18–21; Seattle, WA, USA. p. 1631–42.
39. Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C. Learning word vectors for sentiment analysis. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*; 2011 Jun 19–21; Portland, OR, USA. p. 142–50.
40. Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification. *Adv Neural Inf Process Syst*. 2015;28:649–57. doi: 10.48550/arxiv.1509.01626.
41. Voorhees EM, Tice DM. Building a question answering test collection. In: *Proceedings of the 23rd Annual International ACM SIGIR Conference*; 2000 Jul 24–28; Athens, Greece. p. 200–7.
42. Wilcoxon F. Individual comparisons by ranking methods. *Biom Bull*. 1945;1(6):80–3. doi:10.2307/3001968.
43. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*. 1947;12(2):153–7. doi:10.1007/bf02295996.
44. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108*. 2019.
45. IBM Quantum. IBM Quantum roadmap. 2024 [cited 2026 Jul 13]. Available from: <https://www.ibm.com/quantum/roadmap>.
46. Google AI Quantum. Quantum computing service. 2023 [cited 2026 Jul 13]. Available from: <https://quantumai.google/>.
47. Temme K, Bravyi S, Gambetta JM. Error mitigation for short-depth quantum circuits. *Phys Rev Lett*. 2017;119(18):180509. doi:10.1103/physrevlett.119.180509.
48. Cristianini N, Shawe-Taylor J, Elisseeff A, Kandola J. On kernel-target alignment. *Adv Neural Inf Process Syst*. 2002;14:367–73.