Structured Supervision from Quantum Models: Distilling Robustness into Classical Networks

Syed Rameez Naqvi, Zhengming Ding, Fang Qi, Md Mostafizer Rahman and Lu Peng

Department of Computer Science, Tulane University
6823 Saint Charles Ave, New Orleans, LA 70118 USA
{snaqvi, zding1, fqi2, mrahman9, lpeng3}@tulane.edu

Abstract

Quantum machine learning offers a practical opportunity to leverage near-term quantum devices without requiring quantum inference at deployment. We propose a quantum-classical hybrid knowledge distillation framework in which a variational quantum circuit, equipped with a Quantum Fourier Transform-inspired positional encoding, acts as a soft-label teacher for a compact classical student. Rather than serving as a standalone classifier, the quantum model is used offline to generate structured probability distributions that encode global spectral information and uncertainty. These quantum-derived soft labels are distilled into a lightweight classical network using a hybrid loss that combines hard and soft supervision. Experiments on MNIST demonstrate that students trained with quantum soft labels exhibit consistent and statistically meaningful robustness improvements to Gaussian noise and in-plane rotations compared to classical distillation baselines, while maintaining strong clean accuracy and calibration. These results highlight a distinct role for NISQ-era quantum models as supervisory signal generators, enabling quantum-informed learning within fully classical deployment pipelines.

Introduction

Deep neural networks have achieved remarkable success across vision, speech, and language tasks, yet their growing size and complexity pose challenges for practical deployment. Beyond raw accuracy, real-world systems must balance efficiency, robustness to distribution shifts, and reliable uncertainty estimates (Paleyes, Urma, and Lawrence 2022; Han, Mao, and Dally 2015; Wang 2023). Knowledge distillation (KD) addresses this tension by training compact *student* models to match the predictive behavior of larger *teacher* networks, transferring not only accuracy but also aspects of the teacher's predictive structure through soft supervision (Hinton, Vinyals, and Dean 2015).

In parallel, quantum machine learning (QML) has emerged as a promising paradigm for constructing expressive models using near-term quantum devices. Prior work has shown that even shallow variational quantum circuits[1] (VQCs) can induce rich, non-classical feature representations and probability distributions (Havlíček et al. 2019; Schuld and Killoran 2019; Benedetti et al. 2019). However, most existing QML studies evaluate quantum models primarily as end-task classifiers and emphasize clean accuracy on small benchmarks. This framing underutilizes a key capability of quantum models: their ability to generate structured predictive distributions that encode uncertainty and global correlations, even when their standalone accuracy is limited.

In this work, we argue that near-term quantum models are better viewed not as deployable predictors, but as *supervisory signal generators*. We introduce a quantum–classical hybrid KD framework in which a VQC acts as a soft-label teacher for a compact classical student. Rather than performing quantum inference at deployment, the quantum model is used offline to produce expressive probability distributions that serve as training targets. These quantum-derived soft labels are then distilled into a classical network using a hybrid loss that combines hard labels with soft supervision, enabling quantum-informed learning while preserving fully classical inference.

Crucially, the goal of this work is not to establish quantum advantage or to rank quantum encodings by predictive performance. Instead, we study how the inductive biases induced by a specific quantum encoding are transferred to classical models through soft-label distillation. In this sense, our contribution is mechanism-driven rather than benchmark-driven: we treat the quantum model as a controllable source of supervisory geometry and analyze how this geometry shapes robustness in distilled classical students.

A central component of our framework is a Quantum Fourier Transform (QFT)-inspired positional encoding (PE) that injects global spectral structure into the quantum circuit's input representation (Coppersmith 2002). This encoding induces a distinctive geometry over the teacher's output distributions, shaping the uncertainty structure presented to the student during training. Although the resulting quantum teachers are weak classifiers in isolation, their soft-label geometry contains encoding-specific inductive biases that classical students can exploit. As a result, students distilled from quantum teachers achieve strong clean accuracy while exhibiting improved robustness to Gaussian noise and in-plane rotations compared to classical distillation baselines.

Our contributions are summarized as follows:

[1] A variational quantum circuit (VQC) is a parameterized quantum circuit whose rotation angles are optimized using classical gradient-based methods.

- We propose a practical quantum–classical KD framework that uses VQCs as soft-label teachers, enabling quantum-informed supervision without requiring quantum inference at deployment.
- We introduce QFT-PE − an encoding method for quantum teachers that induces global spectral structure in soft-label distributions, shaping the robustness and calibration behavior of distilled classical students.
- Through controlled experiments on MNIST, we demonstrate that quantum-distilled students outperform classical KD under Gaussian noise and rotational perturbations, despite the limited standalone accuracy of the quantum teachers.

Taken together, these results suggest a shift in how near-term quantum models should be evaluated: not solely as classifiers competing on accuracy, but as structured supervisors whose probabilistic outputs can instill robustness and reliability into lightweight classical networks.

Scope of empirical evaluation. We evaluate our framework on MNIST as a controlled testbed to isolate how structured quantum soft-label geometries influence robustness and calibration in distilled classical models. This choice reflects current NISQ constraints and aligns with prior work that uses small benchmarks to study representation geometry and inductive bias transfer. Our contribution is methodological rather than benchmark-driven, focusing on the role of quantum models as supervisory signal generators rather than as deployable predictors.

Our framework uses a small quantum circuit to generate structured uncertainty patterns that are then transferred to a classical neural network via KD. The quantum component does not perform final inference; rather, it shapes the geometry of supervisory signals during training. Our objective is not to outperform standalone neural networks in clean accuracy, but to examine how quantum-induced inductive biases influence robustness and calibration in classical students.

Background and Related Work

KD Concept. KD was originally introduced as a model compression technique, enabling compact student networks to approximate the predictive behavior of larger teachers through softened output distributions (Hinton, Vinyals, and Dean 2015). Intuitively, KD trains a smaller model to mimic the probability outputs of a larger model, allowing the student to inherit aspects of the teacher's predictive structure. By transferring probabilistic information rather than one-hot labels, KD exposes aspects of teacher's inductive bias and uncertainty structure, often improving generalization beyond what is achievable with hard supervision alone.

Subsequent work has extended KD beyond efficiency, showing that soft-label supervision can also improve calibration and robustness. Prior studies demonstrate that distilled students may inherit resilience to distribution shifts and adversarial perturbations from their teachers, and that temperature-scaled supervision can mitigate overconfidence in modern neural networks (Guo et al. 2017; Furlanello et al. 2018; Goldblum et al. 2020; Yuan et al. 2020). These findings suggest that KD is not merely a compression tool, but a mechanism for transferring predictive geometry and uncertainty structure from teacher to student. Our work builds on this perspective, focusing on how the structure of the teacher's output distribution shapes student robustness.

QML for Supervised Tasks. QML explores how quantum circuits can be used to process classical data and learn predictive mappings under near-term hardware constraints (Schuld and Petruccione 2018; Benedetti et al. 2019). A central line of work studies quantum feature maps and VQCs as expressive models that embed classical inputs into high-dimensional Hilbert spaces, enabling non-classical representations and kernel-based learning (Havlíček et al. 2019). While these models are often evaluated as end-task classifiers, their performance is typically limited by circuit depth, noise, and qubit count in the NISQ regime.

Recent perspectives emphasize hybrid quantum–classical pipelines as a practical route forward, combining quantum models with classical optimization and inference (Broughton et al. 2020; Jerbi et al. 2023). However, most prior work still treats quantum circuits as predictors rather than as sources of supervisory information. Only limited attention has been paid to how the probabilistic outputs of quantum models might be exploited as training signals for classical networks.

Positioning of This Work. Our approach bridges these two lines of research by treating a VQC as a *soft-label teacher* within a KD framework. Rather than deploying the quantum model at inference time, we use it offline to generate structured probability distributions that encode global correlations and uncertainty. These quantum-derived soft labels are then distilled into a compact classical student. In contrast to prior KD methods that rely on classical teachers, and prior QML work that evaluates quantum models primarily on accuracy, we study how the geometry of quantum soft labels influences robustness and calibration in classical students under realistic NISQ constraints.

Proposed Framework

Motivation and Framework Overview

Our goal is to combine the expressive representational geometry of quantum models with the efficiency and scalability of classical neural networks. Rather than deploying a quantum model at inference time, we use a VQC as an offline *soft-label teacher*. The quantum model produces informative probability distributions that capture richer class relationships and uncertainty structure than one-hot targets. These distributions are distilled into a compact classical *student* using a hybrid loss, enabling the student to inherit robustness and calibration properties from the quantum teacher while remaining fully classical at deployment.

Conceptually, the framework realizes a pipeline for *quantum-informed supervision*. From classical KD, we adopt soft-label transfer as a mechanism for conveying inductive bias beyond hard accuracy signals (Hinton, Vinyals, and Dean 2015). From QML, we draw on advances in variational quantum circuits that demonstrate expressive behavior

even under NISQ-era constraints (Benedetti et al. 2019; Mitarai et al. 2018; Havlíček et al. 2019). By unifying these ideas, we treat quantum models not as predictors competing with classical networks, but as supervisory signal generators whose output geometry shapes downstream learning.

Scalability considerations. The quantum module does not operate on raw inputs. Instead, each image is pooled into a fixed-length vector of dimension n, corresponding to the number of qubits in the VQC. Consequently, the quantum cost is independent of input resolution: high-resolution images map to the same n-dimensional quantum input as low-resolution ones. Only the classical student architecture scales with dataset complexity. This design makes the framework compatible with large-scale pipelines, with the quantum component serving as a resolution-agnostic soft-label generator suitable for near-term quantum devices.

Quantum Label Generator (QLG)

We implement a VQC as a quantum label generator that produces soft probability distributions used to supervise classical students. Given an input image (e.g., MNIST or CIFAR-10), we first apply average pooling and flattening to obtain a fixed-length vector of dimension n, where n corresponds to the number of qubits. Unless stated otherwise, we use $n = 10$ across all experiments. Each pooled value is normalized to $[0, 1]$ and mapped to a rotation angle in $[0, \pi]$, ensuring a bounded and smooth input embedding suitable for NISQ-era circuits.

QFT-PE. To inject global structure into the quantum input representation, we employ a positional encoding inspired by the QFT. Let S_i denote the i-th pooled segment of the input. The rotation angle applied to the i-th qubit is defined as

$$\theta_i = \pi \cdot \left(\left(\frac{1}{|S_i|} \sum_{x \in S_i} x \right) \cdot (i+1) \mod 1 \right), \qquad i = 1, \ldots, n$$

This encoding modulates input intensity by positional index, introducing structured phase variation across qubits. The modulo operation enforces bounded phase wrapping, ensuring that positional scaling induces controlled phase dispersion rather than unbounded angle growth as i increases. Unlike classical sinusoidal positional encodings, which impose fixed periodic structure, this formulation introduces index-dependent phase variation that is subsequently mixed through entanglement. While inspired by the phase structure of the QFT, we do not implement an explicit QFT; instead, we leverage its core inductive bias, i.e., global spectral mixing, to shape the geometry of the resulting soft-label distributions. The resulting embedding imposes a global spectral bias over the input dimensions, shaping the geometry of the quantum model's output distributions.

$$|\psi_{\text{enc}}\rangle = \bigotimes_{i=1}^{n} R_Y(\theta_i) \, |0\rangle$$

Variational circuit and readout. Following encoding, the qubits are processed by a depth-L variational circuit composed of parameterized single-qubit rotations and ring-structured entanglement (Fig. 1). Each layer applies

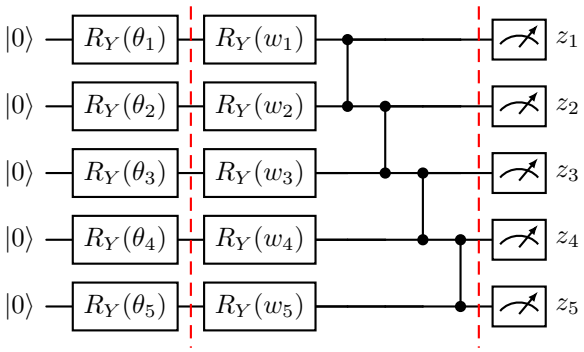


Figure 1: VQC layer: First column encodes $\{\theta_i\}$. Then one trainable layer of $R_Y(w_i)$ followed by a ring entanglement via CZ (shown for 5 qubits). Measurements yield $z_i = \langle Z_i \rangle$. For L layers, repeat the $\{R_Y(w_i)\}$+CZ block between red dash-lines for L times.

$$U(\mathbf{w}) = \prod_{i=1}^{n} R_Y(w_i) \prod_{i=1}^{n} \text{CZ}(i, (i+1) \bmod n)$$

with shared parameters $\mathbf{w}$ across layers to limit circuit capacity and stabilize training. The circuit outputs expectation values of Pauli-Z operators,

$$z_i = \langle Z_i \rangle, \qquad i = 1, \ldots, n$$

which are stacked into a feature vector $\mathbf{z} \in \mathbb{R}^n$. A linear classification head maps $\mathbf{z}$ to logits over C classes ($\mathbf{y} = W\mathbf{z} + \mathbf{b}$), trained using categorical cross-entropy. The trained quantum model is then used offline to generate soft-label distributions for distillation (see Fig. 2 for workflow).

Inductive bias and supervision geometry. The QFT-PE encoding acts as an explicit inductive bias, emphasizing global spectral structure rather than local feature correspondence. Although the resulting quantum teacher is a weak classifier in isolation, its soft-label outputs encode structured uncertainty and class relationships. These properties shape the supervision geometry presented to the student during distillation and are central to the robustness gains observed in downstream classical models.

NISQ compatibility and scalability. The QLG uses shallow circuits with $n \leq 10$ qubits, ring entanglement, and single-qubit rotations, remaining well within current NISQ depth and connectivity limits. The quantum model operates exclusively offline, and inference is fully classical. Because the quantum module processes fixed-length pooled representations rather than raw inputs, its computational cost is independent of image resolution; only the classical student scales with dataset complexity. The robustness effects studied here arise from encoding-induced geometry and entanglement-driven mixing, rather than from exploiting large Hilbert-space dimensionality.

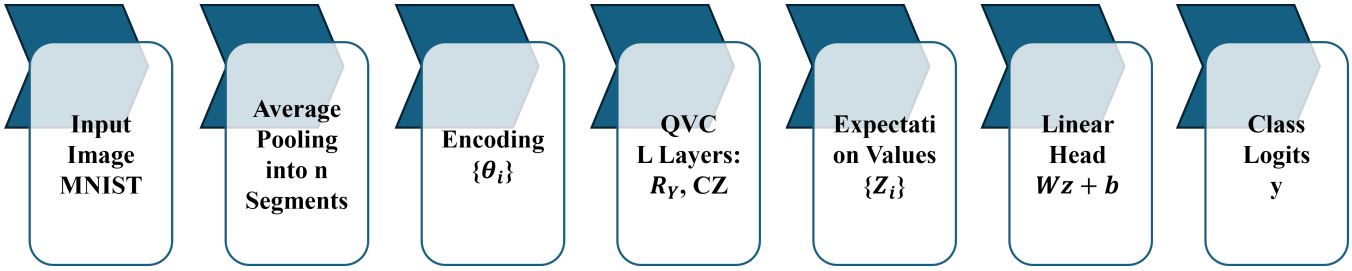


Figure 2: Pipeline of the QLG. Input images are pooled into n segments, encoded, processed through a VQC, and mapped by a linear head to produce class logits

Distributed Soft-Label Generation

To generate quantum-derived soft labels at scale, we apply the trained QLG to the full dataset in an offline inference stage. The dataset is partitioned across multiple workers, each processing a disjoint subset of inputs and producing temperature-scaled probability distributions. For an input $\mathbf{x}^{(j)}$, the QLG outputs logits $\mathbf{y}^{(j)} \in \mathbb{R}^C$, which are converted into soft labels using a temperature-scaled softmax,

$$p_c^{(j)} = \frac{\exp(\mathbf{y}_c^{(j)}/T)}{\sum_{k=1}^{C} \exp(\mathbf{y}_k^{(j)}/T)}, \qquad c = 1, \ldots, C \qquad (1)$$

where $T > 0$ controls the smoothness of the resulting distribution.

Each worker stores its locally generated soft-label shard together with the corresponding dataset indices. After all workers complete inference, these shards are merged to reconstruct an ordered soft-label tensor aligned with the original dataset splits (train/test). This procedure enables efficient generation of soft supervision for large datasets while keeping the quantum model strictly offline.

We use analytic expectation values for quantum measurements and evaluation-only inference. The per-example quantum cost scales as $O(nL)$ for n qubits and circuit depth L, while overall wall-clock time scales approximately linearly with the number of workers. As a consistency check, we verify that the merged soft labels reproduce the quantum teacher's top-1 accuracy on each dataset split.

Classical Student Training via KD

We train a compact classical student using quantum-derived soft labels generated offline by the QLG. The student operates entirely in the classical domain and is evaluated under standard supervised learning protocols. We consider lightweight architectures suitable for deployment, including a shallow multilayer perceptron (MLP) and a small convolutional neural network (CNN). Data augmentation, when used, is applied only during training to assess robustness under distributional shifts.

Distillation objective. Student training minimizes a hybrid loss that combines supervision from hard labels with soft-label matching against the quantum teacher. Let $\mathbf{z}_s \in \mathbb{R}^C$ denote the student logits for an input, and let $\mathbf{p}_{\text{teacher}} \in \mathbb{R}^C$ be the corresponding quantum-derived probability distribution. The hard-label target with optional label smoothing ε is

$$\tilde{\mathbf{y}} = (1 - \varepsilon)\,\mathsf{onehot}(y) + \frac{\varepsilon}{C}\,\mathbf{1}$$

The cross-entropy loss is

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} \tilde{\mathbf{y}}_c \log \mathsf{softmax}(\mathbf{z}_s)_c$$

and the distillation term matches the student's temperature-scaled output to the teacher distribution,

$$\mathcal{L}_{\text{KD}} = T^2\,\mathrm{KL}\big(\mathbf{p}_{\text{teacher}} \,\big\|\, \mathsf{softmax}(\mathbf{z}_s/T)\big)$$

where $T > 0$ is the distillation temperature.

$$\mathcal{L} = \alpha\,\mathcal{L}_{\text{CE}} + (1 - \alpha)\,\mathcal{L}_{\text{KD}}$$

represents the overall objective, with $\alpha \in [0, 1]$ controlling the balance between hard and soft supervision. During evaluation, only the hard-label term is used.

Training protocol. Students are optimized with standard stochastic gradient methods. The computational cost per batch scales with the number of student parameters, while distillation introduces only a minor overhead proportional to the number of classes. Because the quantum teacher is used exclusively offline, student training and inference remain fully classical.

Results and Discussions

MNIST is used intentionally as a controlled diagnostic testbed rather than as a benchmark of representation learning difficulty. By minimizing high-frequency texture and semantic complexity, MNIST allows us to isolate how supervision geometry and uncertainty structure, rather than raw feature extraction capacity, transfer from quantum teachers to classical students. Extending this analysis to more complex datasets is an important direction for future work but is orthogonal to the mechanism studied here.

Calibration and Robustness of the Quantum Teacher

Before distillation, we assess whether the QLG produces non-degenerate, well-behaved soft labels suitable for supervision. On the `test` split, the quantum teacher achieves

	Acc (%)	NLL ↓	ECE ↓	Brier ↓
Teacher	43.27	1.625	0.070	0.0718

Table 1: Calibration of the quantum teacher on `test`

Corruption	Rotation	Noise	Translation	Contrast
AUC-Acc	0.282	0.188	0.205	0.413

Table 2: Teacher robustness measured by AUC-Acc

43.27% accuracy with negative log-likelihood (NLL) of 1.625, expected calibration error (ECE) of 0.070 (15 bins), and Brier score of 0.0718 (Table 1). These values indicate mildly under-confident but stable probabilistic predictions, appropriate for use as soft supervision rather than as a standalone classifier.

We further evaluate robustness under four corruption families: rotation, Gaussian noise, translation, and contrast, each applied at increasing severity. Performance is summarized using the area under the accuracy–severity curve (AUC-Acc − mean accuracy integrated over corruption severity), where higher values indicate greater robustness. As shown in Table 2, the teacher degrades most under noise and translation, moderately under rotation, and least under contrast. The same is confirmed by the plots in Fig. 3. This monotonic and structured degradation confirms that the teacher's outputs vary smoothly with input perturbations, a desirable property for distillation.

As a control, freezing all quantum circuit parameters and training only the linear head yields 34.6% test accuracy, while training with shuffled labels collapses to near-chance performance. These checks confirm that the quantum embeddings encode meaningful signal and that no label leakage occurs. Although the quantum teacher is weak in isolation, its calibrated and structured soft-label distributions make it a viable supervisory signal for classical students.

Quantum Teacher Analysis and Validation

To ensure the quantum teacher provides meaningful supervisory signals, we conducted several validation checks. First, we verified that the generated soft-label distributions are well-formed: probability rows sum to ≈ 1 with no NaN/Inf values, and the encoding produces physically plausible rotation angles (mean ≈ 1.16 rad, std ≈ 0.99 rad).

Beyond distribution integrity, we confirmed that the quantum model learns non-trivial features through two control experiments: (1) *Head-only training* (frozen quantum circuit) achieves 34.6% test accuracy, which is above chance but well below the full model's 43.3%, indicating that the quantum embedding captures useful structure. (2) *Label-shuffled training* collapses to near-chance accuracy (14.0%), confirming no label leakage or implementation artifacts.

We further analyzed the teacher's behavior across four corruption families (rotation, Gaussian noise, translation, contrast). The teacher degrades most under noise (AUC-Acc 0.188) and translation (0.205), moderately under rotation (0.282), and least under contrast (0.413). This structured, monotonic degradation suggests the teacher's outputs vary smoothly with input perturbations, which is a desirable property for distillation.

Quantum Soft-Label Geometry

The quantum teacher's soft labels encode structured uncertainty that classical students can exploit. Figure 4 shows representative examples: low-entropy predictions correspond to clear cases (e.g., a confident "7"), while high-entropy cases capture legitimate ambiguity (e.g., between "4" and "9"). This meaningful uncertainty structure, rather than uniform confusion, forms the geometric basis for robustness transfer.

To further characterize this geometry, we visualize the soft-label distributions using t-SNE (Fig. 5). The embedding reveals class-coherent clusters with plausible confusion patterns (e.g., "2"/"3" proximity), indicating that the quantum teacher organizes probability space in a structured, semantically meaningful way. This geometry, shaped by the QFT-inspired encoding and entanglement, provides the supervisory structure that distilled students inherit.

Quantum Teacher to Classical Student KD

We first verify that a weak quantum teacher can nevertheless transfer useful supervisory information to a compact classical student. Table 3 compares the QFT-based quantum teacher and its distilled student on MNIST. Although the quantum teacher attains only modest accuracy, distillation yields a high-performing student with strong calibration, demonstrating effective knowledge transfer from NISQ-era quantum models.

Robustness to Gaussian Noise and Rotations

We evaluate the distilled *student* under combined Gaussian noise and in-plane rotations on MNIST. Table 4 reports accuracy over a grid of noise levels and rotation angles. The QFT-distilled student exhibits strong robustness, with only a modest drop under increasing Gaussian noise at zero rotation (from $\sim$96.5% to $\sim$93.6% as σ increases to 0.9), while larger degradations occur primarily under severe rotations.

Accuracy − Robustness Retention

To summarize the trade-off between clean accuracy and robustness, we report mean performance over the full noise–rotation grid and the corresponding retention ratio (mean divided by clean accuracy). As shown in Table 5, the QFT-distilled student retains a substantial fraction of its clean performance under corruption, indicating that robustness gains do not come at the expense of excessive accuracy loss.

Qubit-count Ablation under NISQ Constraints

We assess the sensitivity of quantum teacher performance to circuit width by training VQC teachers with $\{8, 9, 10\}$ qubits under identical settings (MNIST 10k, 20 epochs, 5-layer circuit). Results in Table 6 reveal a characteristic non-monotonic trend in the NISQ regime: while 8-qubit models lack expressivity, the 10-qubit model exhibits a perfor-

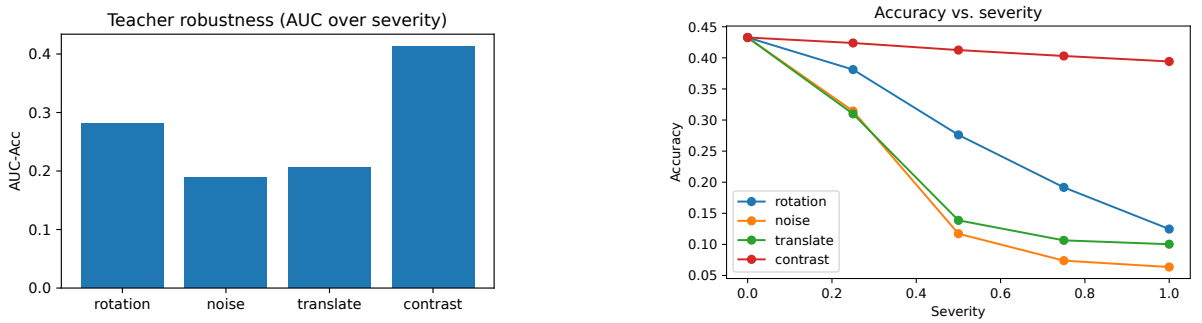


Figure 3: Teacher robustness. Left: AUC-Acc per corruption family (higher is better). Right: Accuracy vs. severity curves for each family, confirming smooth, monotonic degradation suitable for distillation

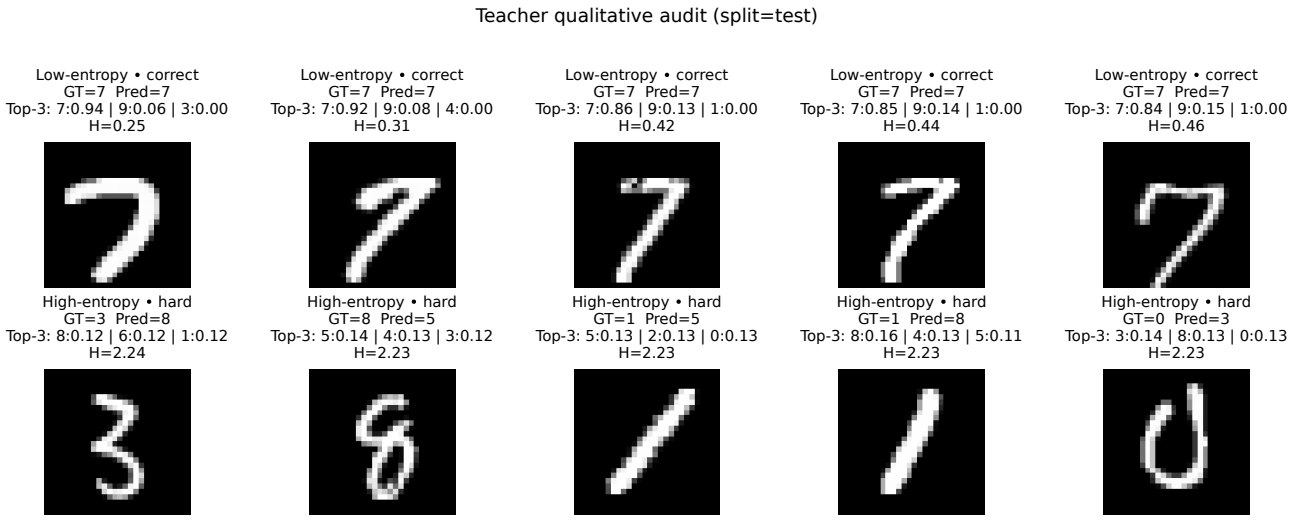


Figure 4: Qualitative audit of quantum teacher predictions. Each panel shows ground truth (GT), predicted class, and top-3 probabilities. Low-entropy (confident) cases align with human intuition, while high-entropy (ambiguous) cases reflect meaningful uncertainty

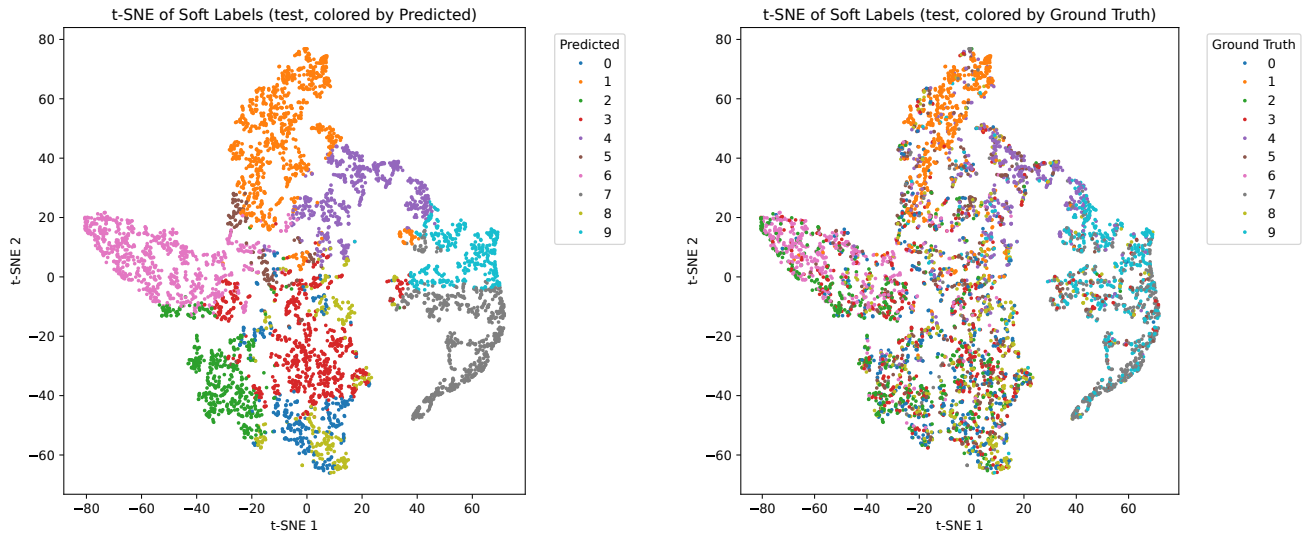
mance drop compared to the 9-qubit variant. This behavior aligns with known optimization challenges in VQCs, where increasing qubit count can lead to barren plateaus and heightened sensitivity to noise (McClean et al. 2018; Cerezo et al. 2021). The 9-qubit configuration, offering the best balance of expressivity and trainability within our experimental setup, was therefore selected for all distillation experiments. While our current implementation operates within NISQ constraints, the framework itself is agnostic to circuit width and could naturally incorporate advances in quantum hardware as they emerge.

Robustness Against Classical Baselines

We compare the robustness of the QFT-distilled student against strong classical baselines trained with identical architectures and data: Classical KD and Mixup, both using a LeNet backbone. All models are evaluated on MNIST under a combined grid of Gaussian noise and in-plane rotations. Importantly, the goal of this comparison is not to surpass classical baselines in clean accuracy, but to evaluate robustness retention under structured perturbations.

Classical KD achieves the highest clean accuracy (98.73%), followed by the QFT-distilled student (96.50%) and Mixup (89.47%). However, under geometric perturbations, the QFT-distilled student exhibits substantially stronger robustness retention, particularly under moderate to severe rotations. While the detailed robustness results for Classical KD and Mixup models under Gaussian noise and rotations are provided in Tables 7 and 8 respectively, Table 9 summarizes clean accuracy, mean accuracy over the corruption grid, and retention (mean divided by clean accuracy).

Notably, the QFT-distilled student achieves the highest retention ratio (0.774), retaining ≈78.0% of its clean accuracy under combined noise and rotation. This surpasses Classical KD (73.2%) by 4.8 percentage points and Mixup (53.5%) by 24.5 percentage points, demonstrating that quantum-informed supervision provides a robustness advantage not captured by classical distillation or augmentation alone.



(a) Colored by predicted class

(b) Colored by ground truth

Figure 5: t-SNE visualization of quantum teacher soft-label geometry on MNIST test set. (a) Points colored by teacher's argmax prediction show class-coherent clusters. (b) Ground-truth coloring reveals semantically meaningful confusion patterns (e.g., "2"/"3" proximity)

Model	Acc (%) ↑	NLL ↓	ECE ↓	Brier ↓
Teacher (QFT)	43.27	1.625	0.070	0.0718
Student (QFT KD, T=2)	96.50 ± 0.14	0.42 ± 0.00	0.28 ± 0.00	0.17 ± 0.00

Table 3: Teacher vs. student comparison on MNIST

σ	0°	15°	30°	45°	60°
0.0	96.50	95.64	92.98	76.10	50.08
0.1	93.49	90.98	83.00	62.84	37.80
0.3	93.42	90.83	83.11	62.71	37.87
0.5	93.43	90.79	83.07	63.18	37.91
0.7	93.23	90.84	83.48	63.98	38.54
0.9	93.57	91.09	84.18	65.12	39.51

Table 4: Robustness of the QFT-distilled student on MNIST under Gaussian noise (σ) and rotation

Model	**Clean** (%)	**Mean grid** (%)	**Retention**
QFT KD (student)	96.50	74.70	0.774

Table 5: Clean accuracy, mean robustness over the noise–rotation grid, and retention on MNIST

Qubits	8	9	10
Test Accuracy (%)	41.54	43.56	40.02

Table 6: Qubit-count ablation under NISQ constraints. Performance peaks at 9 qubits, reflecting the trade-off between expressivity and trainability in shallow variational circuits

The observed robustness gains under noise and rotation, coupled with the student's maintained calibration and the structured entropy patterns in quantum soft labels, suggest that the benefits of quantum-informed supervision stem not from weak regularization or reduced confidence, but from the transfer of a geometrically structured uncertainty landscape that helps the student generalize under perturbation.

CIFAR-10 Stress Test

To examine the limits of encoding-induced robustness transfer, we evaluate the same QFT-based quantum–classical distillation framework on CIFAR-10. Unlike MNIST, CIFAR-10 contains high-frequency texture, local spatial structure, and semantic variability, placing weaker emphasis on global spectral structure. As expected, robustness retention is substantially lower than on MNIST (Table 10), indicating that QFT-inspired supervision does not confer dataset-agnostic robustness. These results confirm that the proposed framework transfers encoding-dependent inductive bias rather than universal invariance.

σ	0°	15°	30°	45°	60°
0.0	98.73	96.63	84.25	56.71	33.99
0.1	98.32	95.57	82.96	55.82	33.59
0.3	98.31	95.44	82.79	55.61	33.27
0.5	98.14	95.17	82.14	55.23	33.00
0.7	98.01	94.90	81.89	54.44	32.65
0.9	97.62	94.58	81.31	53.89	32.27

Table 7: Robustness of the Classical KD student on MNIST under Gaussian noise (σ) and rotation

σ	0°	15°	30°	45°	60°
0.0	89.47	80.13	63.35	42.32	29.33
0.1	68.57	62.23	48.66	32.74	22.08
0.3	67.76	61.26	48.27	32.98	21.49
0.5	66.40	60.26	46.95	31.67	21.23
0.7	64.48	59.12	46.02	31.27	20.81
0.9	63.35	57.49	45.00	30.33	20.61

Table 8: Robustness of the Mixup student on MNIST under Gaussian noise (σ) and rotation

Limitations of QFT KD Robustness

While QFT-based quantum distillation yields clear robustness gains under Gaussian noise and in-plane rotations, these gains are not uniform across all corruption types. In particular, the improvements are encoding-specific rather than indicative of universal invariance. The QFT-inspired positional encoding induces a global, low-frequency bias that aligns naturally with rotational perturbations and structured noise, but does not confer robustness to spatial translations or severe contrast shifts, which require local spatial consistency.

This behavior reflects the inductive bias introduced by the quantum encoding rather than a deficiency of the distillation framework itself. Different encodings emphasize different invariances: for example, encodings that preserve local phase or angle information may be better suited to translation or contrast robustness. As a result, the framework should be viewed as a controllable mechanism for transferring encoding-dependent robustness, rather than as a method that guarantees robustness to arbitrary distribution shifts.

Extending this approach to broader perturbation families, such as spatial shifts, contrast changes, or adversarial perturbations, may require combining multiple quantum encodings or hybridizing quantum supervision with classical augmentation strategies. Exploring such combinations, as well as evaluating robustness under more diverse distribution shifts, remains an important direction for future work but lies beyond the scope of the present study.

The CIFAR-10 results further demonstrate that robustness gains are encoding- and data-regime dependent; QFT-based supervision emphasizes global structure and is therefore less effective for datasets dominated by local texture and fine-grained spatial features.

Method	**Clean** (%)	**Mean grid** (%)	**Retention**
Classical KD (LeNet)	98.73	72.24	0.732
Mixup (LeNet)	89.47	47.85	0.535
QFT KD (student)	96.50	74.70	0.774

Table 9: Robustness summary on MNIST under Gaussian noise and rotations

Dataset	**Clean (%)**	**Mean grid (%)**	**Retention**
CIFAR-10	65.00	24.20	0.372

Table 10: Clean accuracy, mean robustness (mean over $\sigma \in \{0, \cdots, 0.9\}$ and rotations $\{0°, 15°, 30°, 45°, 60°\}$), and retention (mean/clean) for QFT-distilled students

Conclusion

We propose a practical quantum-classical knowledge distillation framework that redefines near-term quantum models as supervisory signal generators rather than standalone classifiers. By leveraging a variational quantum circuit with a QFT-inspired positional encoding, we generate structured soft-label distributions that encode global spectral information and meaningful uncertainty. These quantum-derived labels are distilled into compact classical students via a hybrid loss, enabling quantum-informed learning while retaining classical deployment. Our experiments on MNIST demonstrate that students trained with quantum soft labels exhibit consistent and statistically meaningful robustness improvements to Gaussian noise and in-plane rotations compared to classical distillation and augmentation baselines, while maintaining strong clean accuracy and calibration. The robustness gains stem from geometric structure induced by the quantum encoding, which classical students inherit through soft-label matching. Importantly, this work does not claim universal quantum advantage but rather highlights a pragmatic pathway for integrating NISQ-era quantum models into classical machine learning pipelines. The framework is scalable, as the quantum component operates only offline on fixed-length representations, and is adaptable to future quantum hardware advances.

Limitations include the encoding-specific nature of robustness, where improvements are pronounced for perturbations aligned with the QFT encoding's inductive bias (e.g., rotations), but not for all corruption types. Future work may explore hybrid encodings, multi-teacher distillation, and evaluation under broader distribution shifts, including adversarial perturbations and real-world domain shifts.

In summary, we propose a shift in perspective: quantum models need not compete with classical networks on accuracy alone, but can instead serve as structured supervisors that enhance robustness and reliability in classical deep learning systems. This approach bridges the gap between quantum expressivity and classical practicality, offering a viable role for quantum computation in the near-term machine learning ecosystem.

References

Benedetti, M.; Lloyd, E.; Sack, S.; and Fiorentini, M. 2019. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4): 043001.

Broughton, M.; Verdon, G.; McCourt, T.; Martinez, A. J.; Yoo, J. H.; Isakov, S. V.; Massey, P.; Halavati, R.; Niu, M. Y.; Zlokapa, A.; et al. 2020. Tensorflow quantum: A software framework for quantum machine learning. *arXiv preprint arXiv:2003.02989*.

Cerezo, M.; Sone, A.; Volkoff, T.; Cincio, L.; and Coles, P. J. 2021. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature communications*, 12(1): 1791.

Coppersmith, D. 2002. An approximate Fourier transform useful in quantum factoring. *arXiv preprint quant-ph/0201067*.

Furlanello, T.; Lipton, Z. C.; Tschannen, M.; Itti, L.; and Anandkumar, A. 2018. Born Again Neural Networks. In *International Conference on Machine Learning (ICML)*.

Goldblum, M.; Fowl, L.; Feizi, S.; and Goldstein, T. 2020. Adversarially Robust Distillation. In *AAAI Conference on Artificial Intelligence*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, 70: 1321–1330.

Han, S.; Mao, H.; and Dally, W. J. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*.

Havlíček, V.; Córcoles, A. D.; Temme, K.; Harrow, A. W.; Kandala, A.; Chow, J. M.; and Gambetta, J. M. 2019. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747): 209–212.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Jerbi, S.; Fiderer, L. J.; Poulsen Nautrup, H.; Kübler, J. M.; Briegel, H. J.; and Dunjko, V. 2023. Quantum machine learning beyond kernel methods. *Nature Communications*, 14(1): 517.

McClean, J. R.; Boixo, S.; Smelyanskiy, V. N.; Babbush, R.; and Neven, H. 2018. Barren plateaus in quantum neural network training landscapes. *Nature communications*, 9(1): 4812.

Mitarai, K.; Negoro, M.; Kitagawa, M.; and Fujii, K. 2018. Quantum circuit learning. *Physical Review A*, 98(3): 032309.

Paleyes, A.; Urma, R.-G.; and Lawrence, N. D. 2022. Challenges in deploying machine learning: a survey of case studies. *ACM computing surveys*, 55(6): 1–29.

Schuld, M.; and Killoran, N. 2019. Quantum machine learning in feature Hilbert spaces. *Physical Review Letters*, 122(4): 040504.

Schuld, M.; and Petruccione, F. 2018. Supervised learning with quantum computers. *Quantum science and technology*, 17.

Wang, C. 2023. Calibration in deep learning: A survey of the state-of-the-art. *arXiv preprint arXiv:2308.01222*.

Yuan, L.; Tay, F. E. H.; Li, G.; Wang, T.; and Feng, J. 2020. Revisiting Knowledge Distillation via Label Smoothing Regularization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.