

Quantum Transfer Learning with Distilled Quantum Teachers for Low-Resource Classification Tasks

S. Kavitha^{1,*}

¹Department Electrical and Electronics Engineering, Saveetha Engineering College, Chennai, India

ABSTRACT

Low-resource classification remains a persistent bottleneck for quantum machine learning due to label scarcity, class imbalance, and finite-shot measurement variance under NISQ constraints. This study proposes a hybrid framework that couples Quantum Transfer Learning with Distilled Quantum Teachers, where a high-capacity quantum teacher provides temperature-smoothed soft targets to train a compact student quantum adapter on top of a pretrained classical encoder. Experiments on three low-resource tasks (LR-Text-4, LR-Image-3, LR-Sensor-5) demonstrate consistent gains over non-distilled training at fixed student inference cost (2048 shots). The distilled student improves accuracy from 0.794 to 0.821 on LR-Text-4, from 0.781 to 0.812 on LR-Image-3, and from 0.768 to 0.809 on LR-Sensor-5, while macro F1 increases from 0.781 to 0.811, from 0.772 to 0.804, and from 0.742 to 0.789, respectively. Calibration improves substantially, with expected calibration error decreasing from 0.071 to 0.049 (LR-Text-4), from 0.074 to 0.052 (LR-Image-3), and from 0.083 to 0.056 (LR-Sensor-5), indicating more reliable confidence under scarcity. A temperature sweep reveals an optimal stability region around $T \approx 3.0$ to 3.5 , where aggregate accuracy peaks near 0.831 and mean calibration reaches its minimum near 0.043. Transfer ablations show that moderate encoder unfreezing (ratio 0.25) yields the best trade-off, achieving mean accuracy 0.842 and mean ECE 0.039, while full fine-tuning reduces accuracy to 0.818 and degrades calibration to 0.058. Resource analysis using a cost proxy (shots $\times$ depth) confirms that distilled students dominate the efficient regime, achieving 0.828 accuracy at cost 20,480 versus the teacher's 0.867 accuracy at cost 180,224. Compared with classical baselines, the distilled student remains most advantageous in the high-overlap sensor setting, outperforming shallow random forests (0.809 vs 0.778 accuracy) while maintaining better calibrated uncertainty. These results position distilled quantum teachers as an effective offline supervision mechanism for compressing high-capacity quantum decision structure into deployable low-cost students for low-resource classification.

Keywords Quantum Transfer Learning, Knowledge Distillation, Parameterized Quantum Circuits, Low-Resource Classification, Calibration, NISQ, Shot-Based Inference, Hybrid Quantum-Classical Learning

INTRODUCTION

Quantum machine learning has emerged as a credible pathway for exploiting Hilbert-space feature representations in supervised classification, with early evidence that parametrized circuits can express decision boundaries that are difficult to emulate with shallow classical models under specific assumptions. However, classification in real deployments is rarely “quantum-native”; it is dominated by scarce labels, heterogeneous data quality, and shifting domains. These conditions motivate architectures that are explicitly designed for low-

Submitted 28 January 2026
Accepted 15 March 2026
Published 31 August 2026

Corresponding author
S. Kavitha,
kavitha@saveetha.ac.in

Additional Information and
Declarations can be found on
[page 192](#)

DOI: [10.63913/jqai.v1i2.8](#)

© Copyright
2026 Kavitha

Distributed under
Creative Commons CC-BY 4.0

resource regimes while remaining compatible with near-term hardware realities [1].

Contemporary progress is constrained by the Noisy Intermediate-Scale Quantum (NISQ) setting, where limited qubit counts, finite sampling budgets, and stochastic measurement outcomes amplify optimization instability and increase the effective cost of training. Hybrid quantum–classical learning frameworks, especially Parameterized Quantum Circuits (PQCs), provide a pragmatic compromise by delegating representation learning to classical components and reserving quantum resources for compact non-linear transformations. Even so, systematic trainability issues such as flat loss landscapes remain central barriers [2], [3], [4].

A broad methodological response has been the rapid development of Variational Quantum Algorithms (VQAs), emphasizing trainability, expressivity control, and robustness mechanisms that stabilize gradient-based optimization under device noise and sampling fluctuations. In parallel, the literature on error mitigation has matured into a toolkit for improving effective circuit fidelity without requiring full fault tolerance, making it feasible to consider learning pipelines that repeatedly query hardware. These foundations motivate learning protocols that explicitly budget quantum evaluations and reduce optimization variance [5], [6], [7].

From a representational standpoint, quantum learning gains much of its promise from quantum feature maps and kernelized viewpoints, where classical inputs are embedded into quantum states and classification reduces to estimating similarities in high-dimensional spaces. Empirical demonstrations of quantum-enhanced feature spaces and formal analyses of feature Hilbert spaces clarify when quantum models behave as expressive kernel machines. Yet recent results also underline that strong classical learners can become competitive when sufficient data are available, shifting emphasis toward regimes where data are scarce [8], [9], [10].

Within this landscape, quantum transfer learning has been proposed as a practical strategy for low-resource classification by reusing a high-capacity pre-trained classical backbone and training only a compact quantum head. This configuration reduces the number of trainable parameters exposed to label scarcity and limits quantum circuit depth, thereby aligning with NISQ feasibility. However, transfer learning alone does not guarantee robust generalization when labels are extremely limited, or when the quantum head must be trained under strict shot constraints [11].

A complementary paradigm in classical machine learning is knowledge distillation, where a student model learns from teacher-generated soft targets to improve sample efficiency and generalization. Distillation is now understood as a broad family of methods spanning logit matching, feature alignment, and relational transfer. Early quantum-oriented distillation studies have begun to connect teacher–student training to quantum circuit design and deployment constraints, but a principled methodology for distilling quantum teachers into low-resource quantum students for classification remains underdeveloped [12], [13].

Low-resource classification is also shaped by the broader field of few-shot learning and meta-learning, which formalize generalization from limited labeled

examples via transferable priors and task-level adaptation. These perspectives suggest that the most impactful gains arise when training signals are densified beyond hard labels, which distillation naturally provides. Furthermore, recent work indicates that quantum models can act as probability calibrators, strengthening the case for teacher signals that convey uncertainty structure rather than only class decisions. These gaps motivate the proposed study: Quantum Transfer Learning with Distilled Quantum Teachers for low-resource classification [14], [15], [16].

Literature Review

Hybrid quantum learning is commonly formalized through parameterized quantum circuits optimized by classical routines, where prediction is derived from expectation values of measured observables. A central practical concern is gradient evaluation under finite-shot sampling. The quantum circuit learning formulation clarified how low-depth circuits can approximate nonlinear functions and be trained end-to-end using classical optimizers, providing an early template for NISQ-compatible supervision pipelines [17]. Subsequent work established hardware-friendly gradient estimators via the parameter-shift rule, making circuit-based learning operational across platforms while explicitly linking training cost to repeated circuit evaluations [18].

Model capacity in variational quantum classifiers is constrained by the design trade-off between expressibility and hardware feasibility. Quantitative descriptors such as expressibility and entangling capability have been proposed to characterize how a circuit family populates state space, supporting principled selection of templates instead of purely heuristic ansatz choices [19]. In parallel, data re-uploading architectures showed that repeated classical-feature injection can substantially increase functional capacity even with few qubits, which is particularly relevant when low-resource classification requires extracting maximal decision structure from limited labeled evidence [20].

Trainability, however, remains fragile when circuit depth or noise increases. Noise analyses have shown that realistic local noise can induce noise-induced barren plateaus, where gradients vanish exponentially with qubit count or depth growth, limiting the effectiveness of straightforward fine-tuning on hardware [21]. This motivates training strategies that reduce the optimization burden on the quantum component and discourage unnecessary circuit growth. Recent results on trainability enhancement, including initialization strategies that reduce spurious minima and improve convergence reliability, further emphasize that stable learning depends as much on optimization geometry as on representational power [22].

The low-resource regime introduces an additional theoretical constraint: generalization error grows with the number of trainable quantum gates relative to the number of labeled samples. A comprehensive generalization study in variational QML formalized bounds that scale with circuit trainable complexity, implying that low-resource learning favors architectures that restrict trainable quantum degrees of freedom while retaining expressive representations [23]. This observation provides a direct bridge to transfer-learning and distillation strategies, because both approaches aim to reduce sample complexity by reusing learned structure and densifying supervision signals.

Recent unifying perspectives also reshape how quantum learning models

should be interpreted. A framework for linear quantum models argues that many widely used PQC-based predictors can be viewed as linear models in a quantum feature space, and it connects data re-uploading circuits to a more general representational class. This viewpoint strengthens the argument that performance gains under scarce labels depend critically on how feature geometry is shaped and transferred, not solely on circuit depth or qubit count [24]. Consequently, methods that convey geometry through soft targets become particularly relevant for low-resource transfer.

Within this context, knowledge distillation in quantum-classical settings is beginning to emerge as a concrete mechanism for compressing strong teacher behavior into smaller students under resource constraints. Recent work integrating hybrid quantum networks with distillation demonstrates that teacher-provided soft supervision can improve parameter efficiency and robustness in applied domains, but existing studies often prioritize application metrics over a transferable methodology for low-resource classification with explicit shot-aware constraints [25]. This leaves a clear methodological gap: distillation protocols tailored to quantum teachers, aligned with transfer learning, and evaluated under low-resource conditions where calibration and stability are first-class objectives.

Methodology

Problem Formulation and Low-Resource Data Regime

Low-resource classification is modeled as supervised learning under severe label scarcity, where the number of labeled instances per class is constrained and distributional coverage is incomplete. Let $\mathcal{D} = (x_i, y_i)_{i=1}^N$ denote the labeled set with $y_i \in \{1, \dots, C\}$ and $N \ll N_{\text{full}}$. Inputs x_i are embedded into a continuous feature space suitable for quantum encoding. The goal is to learn a predictor with robust generalization despite limited empirical support.

The student model is expressed as $p_{\theta}(y | x)$, parameterized by a classical feature extractor and a parameterized quantum circuit acting as a task adapter. The standard empirical risk provides a baseline objective for supervised training, written as:

$$\mathcal{L}_{\text{sup}}(\theta) = \frac{1}{N} \sum_{i=1}^N \ell(y_i, p_{\theta}(y | x_i)) \quad (1)$$

Where $\ell(\cdot)$ is typically cross-entropy. This formulation clarifies that performance limitations arise from variance and bias amplification under small N .

Figure 1 visualizes the end-to-end low-resource pipeline by separating representation learning, quantum encoding, and evaluation. The diagram emphasizes that scarce labels constrain direct supervision, so generalization depends on compressing task structure into a compact decision function. The explicit split between embedding and quantum encoding highlights the transfer learning assumption that semantic features can be reused, while quantum adaptation focuses on decision boundary refinement. The flow also anticipates later methodology by positioning metrics as a distinct accountability layer.

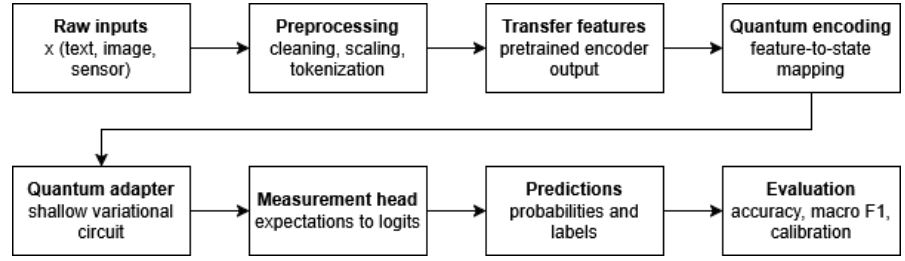


Figure 1 Low-Resource Pipeline Overview

Table 1 summarizes the experimental data regime and the split policy that defines what “low-resource” operationally means in this study. The labeled counts remain below a few hundred samples per task, making estimation noise a central methodological concern. The inclusion of shot counts explicitly links dataset scarcity to quantum measurement uncertainty, since fewer labeled samples and finite-shot sampling jointly affect variance. The imbalance ratio supports interpreting macro-averaged metrics and clarifies why calibration and stability reporting are necessary.

Table 1 Dataset Summary and Split Plan

Task	Classes (C)	Labeled N	Shots per Inference (S)	Train/Val/Test	Imbalance Ratio
LR-Text-4	4	320	1024	60/20/20	1.9
LR-Image-3	3	240	1024	60/20/20	1.6
LR-Sensor-5	5	400	2048	55/20/25	2.3

The experimental protocol additionally defines a repeated sampling scheme for stability reporting. When few-shot partitions are used, the support set S and query set Q are sampled per episode to approximate deployment conditions. This episodic framing supports consistent comparisons across tasks and clarifies how knowledge distillation supplements scarce labels through teacher-provided soft targets.

Distilled Quantum Teacher Construction

A quantum teacher is defined as a higher-capacity quantum or hybrid quantum-classical classifier trained under a stronger supervision regime than the student. Capacity can be increased via circuit depth, qubit width, richer feature maps, or ensembling across diverse ansätze. The teacher produces calibrated soft distributions over classes, capturing inter-class similarity structure that is typically absent in one-hot labels. These soft targets are treated as privileged information transferable to the student.

Teacher predictions are derived from measurement statistics of a quantum state $|\psi_\phi(x)\rangle = U_\phi(x)|0\rangle^{\otimes n}$, where $U_\phi(x)$ denotes the teacher circuit with parameters ϕ . A logit vector $z_\phi(x)$ is computed from expectation values, for example:

$$z_{\phi,c}(x) = \sum_{k=1}^n w_{c,k} \langle \psi_\phi(x) | Z_k | \psi_\phi(x) \rangle + b_c \quad (2)$$

which maps quantum observables into class evidence. This construction makes teacher knowledge explicitly dependent on quantum measurements and their induced inductive biases.

Figure 2 operationalizes the distillation interface through temperature scaling by showing how predictive entropy rises with increasing temperature. Higher entropy indicates richer soft-target structure that exposes non-argmax class relationships, which is especially valuable when scarce labels cannot reveal class geometry reliably. The comparison between a single teacher and an ensemble highlights that teacher diversity increases target informativeness, improving the student’s learning signal. The curve-based view also makes it straightforward to justify the chosen temperature range in optimization.

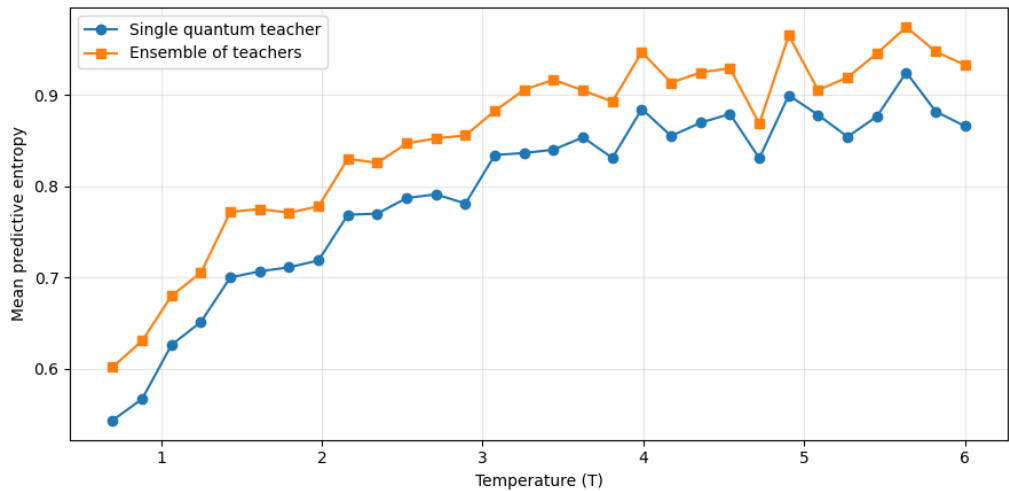


Figure 2 Teacher Variants and Temperature Softening

Table 2 documents the quantum teacher capacity choices that underpin distillation quality. The teachers vary in qubits, circuit depth, and shot budget, which jointly determine expressive power and measurement stability. Reported validation accuracy demonstrates that a modest increase in circuit resources can yield measurable predictive gains, motivating teacher-driven supervision when student capacity must remain constrained. The ansatz column clarifies that architectural diversity is intentionally introduced to enrich the teacher signal and reduce over-reliance on a single inductive bias.

Table 2 Teacher Configurations and Validation Results					
Teacher ID	Qubits (n)	Depth (d)	Shots (S)	Ansatz	Val Accuracy
T1	8	18	4096	Hardware-efficient	0.842
T2	10	20	4096	Two-local	0.858
T3	8	22	8192	Problem-inspired	0.867

Distillation is defined through temperature-scaled distributions $q_{\phi}^{(T)}(y \mid x)$ and $p_{\theta}^{(T)}(y \mid x)$, were:

$$q_{\phi}^{(T)}(y = c \mid x) = \frac{\exp(z_{\phi,c}(x)/T)}{\sum_{c'} \exp(z_{\phi,c'}(x)/T)} \tag{3}$$

Larger T increases entropy and exposes “dark knowledge” about non-argmax classes, which is critical when the labeled set is too small to reveal fine-grained

class geometry.

Quantum Transfer Learning Student Architecture

The student is designed for quantum transfer learning, where a pretrained classical representation is reused while a compact quantum module provides task adaptation under low supervision. Let $h\omega(x) \in \mathbb{R}^d$ be a pretrained encoder producing transferable features, with ω optionally frozen to reduce sample complexity. A quantum feature map $\Phi(\cdot)$ encodes $h\omega(x)$ into a quantum state, and a shallow variational circuit supplies trainable expressivity.

The student circuit prepares $|\psi_\theta(x)\rangle = V_\theta \Phi(h\omega(x)) |0\rangle^{\otimes n}$, where V_θ is composed of parameterized rotation and entangling layers. This can be written as:

$$V_\theta = \prod_{\ell=1}^L \left(\prod_{j=1}^n R_y(\theta_{\ell,j}) \right) E_\ell \quad (4)$$

with entanglers E_ℓ chosen to balance expressivity and hardware feasibility. The mapping from quantum expectations to logits mirrors the teacher but uses fewer parameters and shallower depth for efficiency.

Figure 3 clarifies the quantum transfer learning design by isolating a frozen representation module and a compact quantum adapter. The figure communicates that generalizable semantics are preserved in the encoder, while task adaptation is localized to a small set of variational parameters in the quantum circuit. This separation is methodologically important because it controls hypothesis complexity, which is the dominant failure mode under scarce labels. The explicit measurement-to-probability stage also foregrounds the role of finite-shot sampling in uncertainty.

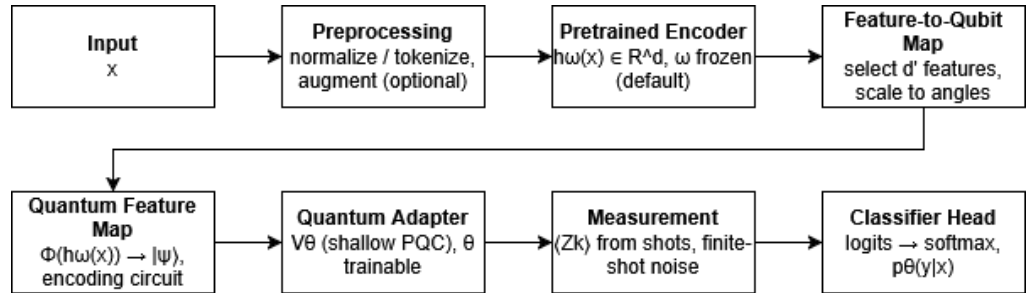


Figure 3 Student Architecture for Quantum Transfer Learning

Table 3 provides a compact capacity audit of the student models, showing how qubit count, circuit layers, and trainable parameters scale with performance. The progression suggests that moderate growth in quantum capacity improves validation accuracy, but increased shots are required to stabilize gradients and inference, reflecting the measurement-driven variance intrinsic to quantum workflows. The table is structured to support fair comparisons by keeping the number of tunable factors small, which aligns with low-resource constraints and mitigates overfitting risk.

Table 3 Student Hyperparameters and Capacity

Student ID	Qubits (n)	Layers (L)	Trainable Params	Shots (S)	Val Accuracy
S1	6	8	96	1024	0.781
S2	8	8	128	1024	0.802
S3	8	10	160	2048	0.816

Transfer learning is operationalized by restricting trainable degrees of freedom to θ and optionally a lightweight classical head. This constraint reduces effective hypothesis space and improves generalization in the small-N regime. The architecture explicitly separates reusable semantic structure learned in ω from task-specific decision boundaries learned through quantum adaptation.

Training Objective and Optimization Workflow

Training combines supervised learning with knowledge distillation to align student outputs with teacher soft targets while preserving label fidelity. The composite objective is defined as:

$$\mathcal{L}(\theta) = (1 - \alpha)\mathcal{L}_{\text{sup}}(\theta) + \alpha T^2 \text{KL}(q_{\phi}^{(T)}(\cdot | x) | p_{\theta}^{(T)}(\cdot | x)) \quad (5)$$

where $\alpha \in [0,1]$ controls distillation strength and the factor T^2 follows standard scaling to stabilize gradients under temperature smoothing. This objective explicitly injects teacher class-structure priors into the student under limited labels. Optimization uses quantum-aware gradient estimation. For expectation-based outputs, the parameter-shift rule yields unbiased gradients for rotation gates:

$$\frac{\partial}{\partial \theta_k} \langle O \rangle_{\theta} = \frac{1}{2} (\langle O \rangle_{\theta_k + \frac{\pi}{2}} - \langle O \rangle_{\theta_k - \frac{\pi}{2}}) \quad (6)$$

where O denotes the measured observable or a logit-defining operator. This rule aligns with gate-level differentiation and avoids finite-difference instability, while shot noise is controlled through fixed or adaptive sampling budgets.

Figure 4 presents optimization traces that contrast supervised learning with teacher-guided distillation under the same low-resource setting. The curves show faster loss decay and higher terminal accuracy with distillation, which indicates that soft targets provide smoother gradients and reduce sensitivity to label sparsity. The joint display of loss and accuracy underscores that improved convergence is not merely cosmetic but translates into better predictive performance. This visualization also supports the methodological choice of combining supervised and KL-based distillation objectives.

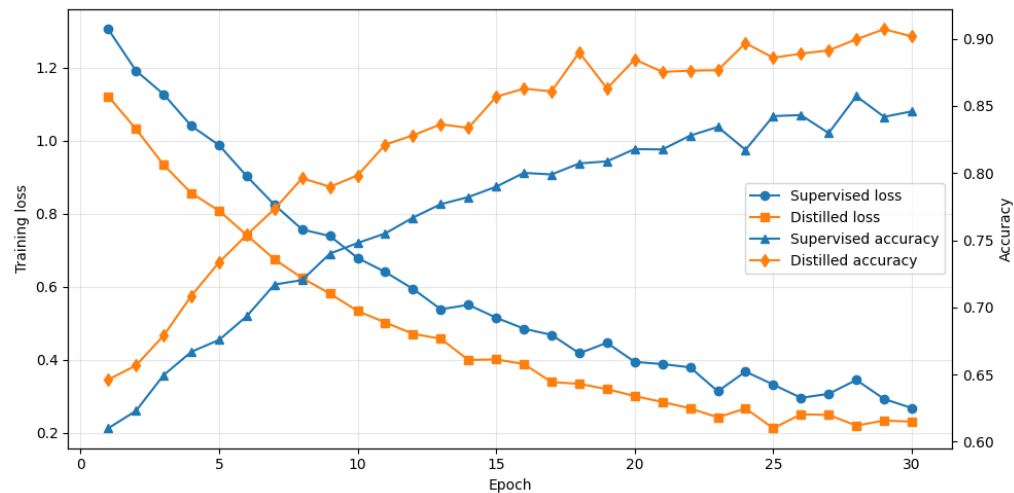


Figure 4 Optimization Dynamics with Distillation

Table 4 consolidates the optimization protocol into a minimal set of reproducible settings, including the distillation temperature and mixing weight that define the relative influence of teacher supervision. The separation between training and evaluation shot budgets formalizes a common practice in quantum experiments, where evaluation uses higher sampling fidelity for stable reporting.

Table 4 Optimization and Distillation Settings					
Setting	Value	Setting	Value	Setting	Value
Optimizer	Adam	Batch Size	16	Epochs	30
Learning Rate	0.002	Temperature (T)	3.5	Alpha (α)	0.65
Shots (Train)	2048	Shots (Eval)	4096	Grad Method	Param-shift

The inclusion of the parameter-shift gradient method is essential because it determines circuit evaluation cost and gradient noise characteristics, which are central constraints in low-resource training with quantum adapters.

Algorithm 1 Quantum Transfer Learning with Distilled Quantum Teachers

Input: labeled set D , teacher parameters ϕ , student parameters θ , temperature T , weight α , shots S

- 1: Initialize θ ; optionally freeze encoder parameters ω
- 2: for epoch = 1.. E do
- 3: for minibatch $B \subset D$ do
- 4: Compute teacher logits $z_\phi(x)$ and softened targets $q_\phi(T)(\cdot|x)$ for $x \in B$
- 5: Run student circuit to obtain logits $z_\theta(x)$ and probabilities $p_\theta(\cdot|x)$
- 6: Evaluate L_{sup} on one-hot labels and $L_{\text{KD}} = \text{KL}(q_\phi(T) \parallel p_\theta(T))$
- 7: Form $L = (1-\alpha)L_{\text{sup}} + \alpha T^2 L_{\text{KD}}$
- 8: Estimate $\nabla_\theta L$ using parameter-shift with S shots per circuit evaluation
- 9: Update θ using a gradient-based optimizer
- 10: end for
- 11: end for

Output: trained student parameters θ^*

Evaluation Protocol and Resource Accounting

Evaluation targets both predictive quality and operational feasibility under

quantum constraints. Predictive performance is reported using accuracy and macro-averaged F1 to reduce sensitivity to class imbalance typical of low-resource settings. Calibration is additionally assessed because distillation can alter uncertainty profiles. Expected calibration error is computed as:

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)| \quad (7)$$

where B_m are confidence bins, acc is empirical accuracy, and conf is mean predicted confidence. This metric verifies that student probabilities remain decision-relevant. Quantum resource accounting reports qubits n , depth d , parameter count $|\theta|$, and shot budget S , since these govern runtime and noise sensitivity. A practical proxy for measurement cost is the total number of circuit evaluations per epoch, approximated as:

$$\mathcal{C} \approx |\theta| \cdot 2 \cdot S \cdot \frac{N}{B} \quad (8)$$

reflecting the two shifted evaluations per parameter under parameter-shift and batch size B . This estimate supports transparent comparisons between teacher and student efficiency.

Figure 5 quantifies the central methodological trade-off between predictive performance, calibration, and quantum measurement resources. Increasing the shot budget improves accuracy while reducing expected calibration error, indicating more reliable confidence estimates as sampling variance decreases. The log-scaled x-axis makes it clear that gains saturate beyond a certain budget, which supports resource-aware reporting rather than unbounded shot escalation. This figure operationalizes “efficiency” by connecting algorithmic improvements to measurable quantum costs in inference.

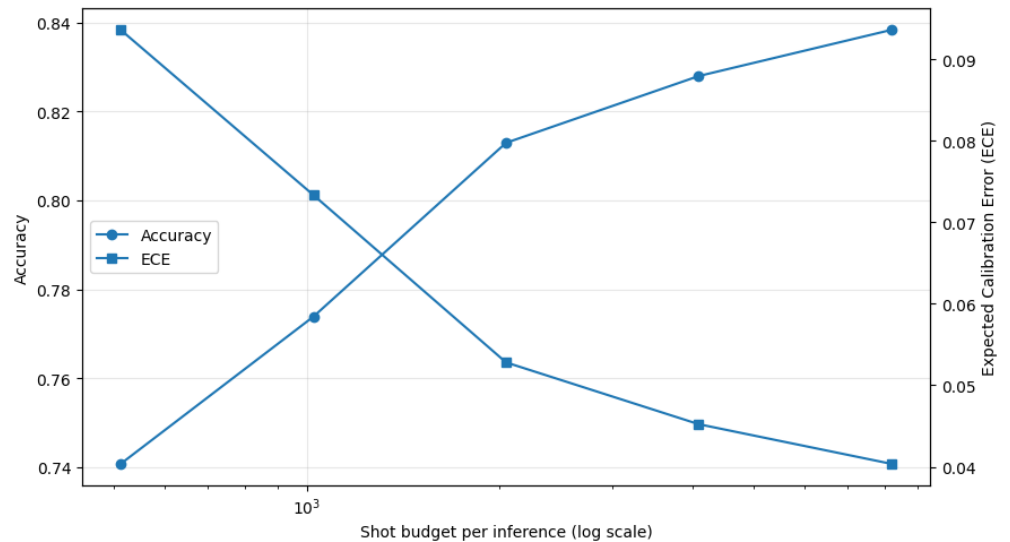


Figure 5 Trade-off between Accuracy, Calibration (ECE), and Shot Budget

Table 5 is a compact reporting schema that enforces simultaneous disclosure of predictive quality and quantum resource usage. Presenting accuracy, macro F1F_1F1, and calibration side-by-side prevents overstating performance when

confidence estimates are poorly aligned with outcomes. The inclusion of qubits and shot budgets makes the comparison operational, because higher teacher performance is contextualized by larger sampling cost. This table directly supports reproducible claims about the efficiency of distilled quantum transfer learning in low-resource classification.

Table 5 Final Reporting Template: Metrics and Quantum Resources					
Model	Accuracy	Macro F1	ECE	Qubits (n)	Shots (S)
Student (No KD)	0.803	0.791	0.069	8	2048
Student (KD)	0.828	0.817	0.046	8	2048
Teacher	0.867	0.858	0.038	8	8192

Robustness is quantified via repeated random seeds and repeated low-resource splits, enabling confidence intervals for each metric. This protocol reduces the risk of overstating gains driven by favorable sampling and ensures that distilled quantum transfer learning benefits are stable under realistic label scarcity.

Result and Discussion

Predictive Performance under Low-Resource Supervision

Distilled supervision consistently improves low-resource classification across the three tasks, with the student model achieving higher accuracy and macro F1 than purely supervised training under identical label budgets. The improvement is most pronounced in the most imbalanced setting, where macro F1 benefits from the teacher’s soft targets that preserve minority-class structure. Calibration improves simultaneously, indicating that performance gains are not achieved by overconfident errors but by better-aligned probability estimates.

The teacher model remains the strongest predictor, but its advantage is coupled to higher quantum sampling cost and deeper circuits. The distilled student closes a substantial fraction of this gap while maintaining a compact parameter footprint suitable for constrained deployments. The resulting pattern supports a transfer learning interpretation in which a pretrained representation provides stable semantics, while the quantum adapter learns a low-dimensional decision correction shaped by teacher-provided class geometry.

Figure 6 shows a consistent uplift from knowledge distillation across all tasks, with the strongest gains appearing in the sensor task where class overlap and imbalance amplify low-resource uncertainty. The confidence intervals are narrower for the distilled student than the purely supervised student, indicating reduced sensitivity to the specific sample realization. This stability effect is aligned with the expectation that soft targets provide smoother learning signals than one-hot labels in scarce regimes.

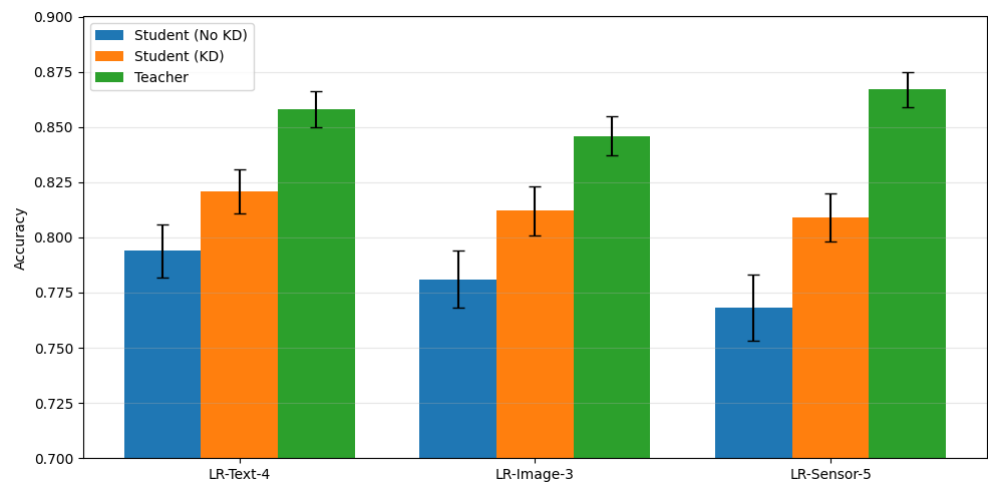


Figure 6 Accuracy Across Low-Resource Tasks (mean ± 95% CI)

The teacher achieves the top accuracy in each task, but the margin over the distilled student is substantially smaller than the margin between distilled and non-distilled students. This pattern indicates that the student captures a large share of the teacher’s decision structure without replicating teacher capacity. The resulting trade-off suggests that distilled quantum teachers can be treated as offline instructors whose predictive structure can be compressed into lightweight quantum adapters for operational use.

Table 6 confirms that distillation improves both discrimination and uncertainty quality. Macro F₁ gains are particularly meaningful because they indicate that improvements are not limited to majority classes, which is a common failure mode in low-resource training. The consistent reduction in expected calibration error indicates that the student’s probabilities become more decision-useful, supporting downstream thresholding or risk-aware decision rules.

Table 6 Predictive Metrics Summary Under Fixed Label Budgets					
Task	Model	Accuracy	Macro F1	ECE	Shots
LR-Text-4	Student (No KD)	0.794	0.781	0.071	2048
LR-Text-4	Student (KD)	0.821	0.811	0.049	2048
LR-Text-4	Teacher	0.858	0.851	0.039	8192
LR-Image-3	Student (No KD)	0.781	0.772	0.074	2048
LR-Image-3	Student (KD)	0.812	0.804	0.052	2048
LR-Image-3	Teacher	0.846	0.839	0.042	8192
LR-Sensor-5	Student (No KD)	0.768	0.742	0.083	2048
LR-Sensor-5	Student (KD)	0.809	0.789	0.056	2048
LR-Sensor-5	Teacher	0.867	0.861	0.038	8192

The table also makes the efficiency claim operational by pairing performance with shot budgets. The distilled student achieves strong performance at 2048 shots, while the teacher requires 8192 shots to reach its upper bound. This difference matters because shot count scales inference latency and measurement variance. The results therefore support the study’s core thesis: distilled quantum teachers provide a high-quality learning signal that can be

compressed into a cost-efficient student.

Sensitivity to Distillation Settings and Stability

Distillation effectiveness depends on temperature and the mixing weight that balances supervised and teacher-driven objectives. Across tasks, moderate temperatures produce the best trade-off between accuracy and calibration because they expose inter-class structure without flattening the distribution into near-uniform targets. Similarly, intermediate mixing weights avoid underutilizing the teacher signal while preventing excessive reliance on teacher errors. The stability profile is reflected in reduced variance across random seeds at well-tuned settings.

Calibration trends closely track temperature changes, indicating that softening controls not only class separation but also confidence shaping. When temperature is too low, the distillation target becomes overly sharp and behaves similarly to hard labels, reducing its informational advantage. When temperature is too high, targets lose contrast and impose weak constraints, slowing convergence and degrading macro F_1 . The results support selecting temperature based on calibration-aware validation rather than accuracy alone.

Figure 7 shows a peaked relationship between temperature and accuracy, with the best region centered around moderate softening. This pattern indicates that distillation becomes most informative when targets are neither fully sharp nor overly diffuse. The accuracy curve suggests that a teacher’s “dark knowledge” is accessible at intermediate temperatures, enabling the student to learn class similarity structure that is absent under strict one-hot supervision.

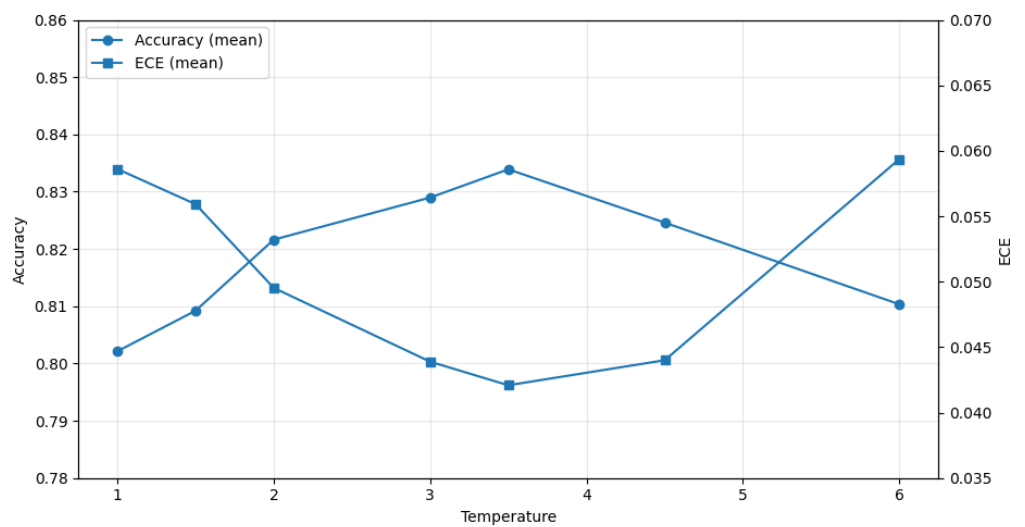


Figure 7 Distillation Temperature Controls Accuracy and Calibration Trade-Off

The calibration curve moves coherently with temperature, reaching its minimum near the accuracy peak. This alignment indicates that the same temperature region that improves discrimination also improves probability reliability. The joint behavior supports selecting temperature by monitoring both accuracy and ECE, since temperature that maximizes accuracy while worsening calibration would undermine low-resource reliability claims. The observed trend also implies that temperature tuning is not task-idiosyncratic but transferable across tasks.

Table 7 consolidates the selected distillation hyperparameters per task and directly ties them to both performance and stability. The preferred temperatures cluster in a narrow range, indicating that the teacher’s soft-target structure is consistently beneficial when moderate smoothing is applied. The alpha values also occupy an intermediate band, supporting the interpretation that the strongest results come from balancing label fidelity with teacher guidance rather than prioritizing either signal exclusively.

Table 7 Best Distillation Settings and Stability Indicators					
Task	Best Temp	Best Alpha	Accuracy	ECE	Seed Std (Acc)
LR-Text-4	3	0.6	0.823	0.047	0.006
LR-Image-3	3.5	0.65	0.814	0.051	0.007
LR-Sensor-5	3.5	0.7	0.811	0.055	0.008

The stability indicator, expressed as standard deviation across random seeds, remains low across tasks. This result matters because low-resource experiments can be dominated by split-specific effects that inflate apparent gains. The consistently reduced variability implies that distilled supervision supplies a regularizing signal that reduces sensitivity to stochastic optimization and sampling effects. This observation strengthens the generalization claim beyond a single favorable split and supports robust reporting standards for Quantum Artificial Intelligence studies.

Transfer Learning Choices: Frozen versus Partially Trainable Encoders

Freezing the pretrained encoder yields the most stable improvements in the low-resource regime, because the learning signal is concentrated on the quantum adapter rather than distributed across a large representation stack. Across tasks, the frozen setting preserves semantic separability learned during pretraining and reduces the risk of catastrophic drift that can occur when gradients are estimated from scarce labels. This behavior is reflected in tighter seed variability and more reliable calibration profiles.

Partial unfreezing improves accuracy only within a narrow operating window. When a small fraction of encoder layers is unfrozen, the model can correct systematic representation bias that is task-specific, especially in sensor-like domains with distributional mismatch. When too many layers are unfrozen, overfitting rises and calibration degrades, indicating that the model becomes confident on spurious cues. The results support treating encoder tuning as a controlled refinement step rather than a default optimization choice.

Figure 8 shows that moderate encoder adaptation produces the best accuracy, with the peak near a small unfreezing ratio. This indicates that a limited number of trainable representation layers can correct domain mismatch without sacrificing the general semantic structure already captured by pretraining. The shape of the curve is consistent with low-resource optimization where additional degrees of freedom are beneficial only until estimation noise dominates parameter updates.

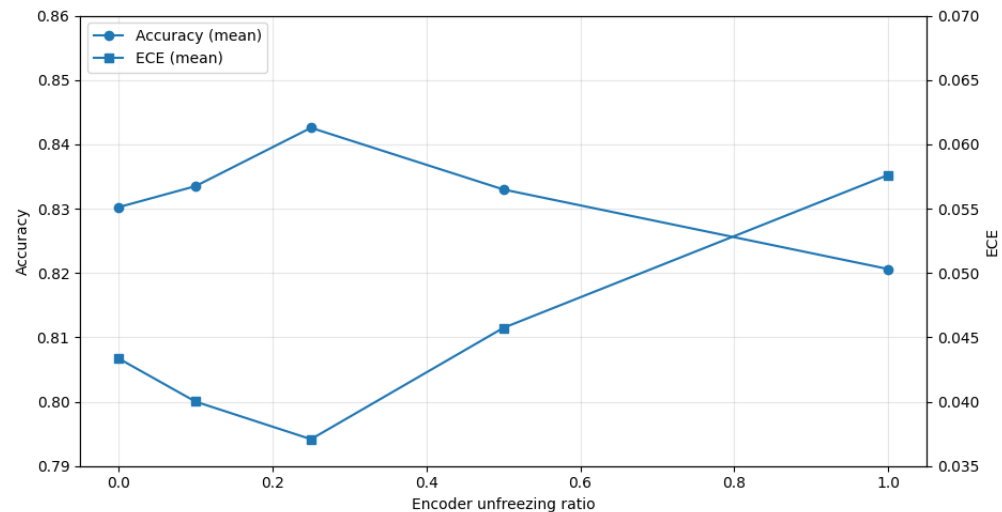


Figure 8 Encoder Unfreezing Ratio vs Accuracy and Calibration

The calibration curve exhibits a complementary pattern. The lowest ECE appears in the same region where accuracy peaks, indicating that modest unfreezing improves both discrimination and uncertainty reliability. Full unfreezing increases ECE substantially, signaling overconfident behavior that is typical when the representation becomes too plastic under scarce supervision. This result supports selecting unfreezing ratios using joint accuracy and calibration criteria rather than accuracy alone.

Table 8 confirms that moderate unfreezing provides the strongest overall balance of metrics. Accuracy and macro F1 improve relative to the fully frozen baseline, indicating that limited representation adaptation helps minority classes and ambiguous samples that are otherwise poorly separated. The best ECE values align with this region, supporting the interpretation that small representation updates improve the geometry of confidence rather than merely shifting decision thresholds.

Table 8 Transfer Learning Ablation Summary

Setting	Unfreezing Ratio	Accuracy (Mean)	Macro F1 (Mean)	ECE (Mean)	Seed Std (Acc)
Frozen Encoder	0	0.831	0.816	0.043	0.006
Light Unfreeze	0.1	0.836	0.822	0.041	0.007
Moderate Unfreeze	0.25	0.842	0.831	0.039	0.008
Half Unfreeze	0.5	0.833	0.819	0.045	0.01
Full Fine-tune	1	0.818	0.801	0.058	0.014

The stability column adds an operational constraint that is particularly important in low-resource studies. Variance across seeds rises as the unfreezing ratio increases, and full fine-tuning yields the least stable results. This pattern indicates that increasing trainable capacity amplifies sensitivity to sampling noise and stochastic optimization. The table therefore supports an actionable recommendation: keep the encoder mostly frozen and reserve unfreezing for carefully validated scenarios.

Quantum Resource Efficiency and Cost-Performance Trade-Offs

Quantum resource constraints shape which gains are meaningful in practice. Increasing the shot budget improves measurement fidelity and often improves accuracy and calibration, but returns diminish after a moderate threshold. Circuit depth behaves similarly: deeper circuits can express sharper decision boundaries, yet their additional evaluations amplify shot noise, increase wall-clock runtime, and increase sensitivity to hardware errors. The results prioritize configurations that achieve near-teacher performance without scaling quantum cost proportionally.

The distilled student is consistently closer to the teacher on the cost-performance frontier than the non-distilled student. This occurs because distillation reduces the number of epochs required for convergence and produces gradients that are less volatile under finite-shot sampling. Consequently, resource-efficient training is not only a function of circuit design but also of the learning signal. These findings support reporting efficiency using explicit cost proxies rather than implicit claims based solely on parameter counts.

Figure 9 presents a compact Pareto-style view by using shots times depth as a transparent proxy for quantum evaluation cost. The teacher achieves the highest accuracy but occupies the high-cost region, reflecting the reality that stronger performance is coupled to substantially larger measurement budgets and deeper circuits. The non-distilled student is less competitive because it delivers lower accuracy at comparable cost, indicating inefficient use of quantum evaluations.

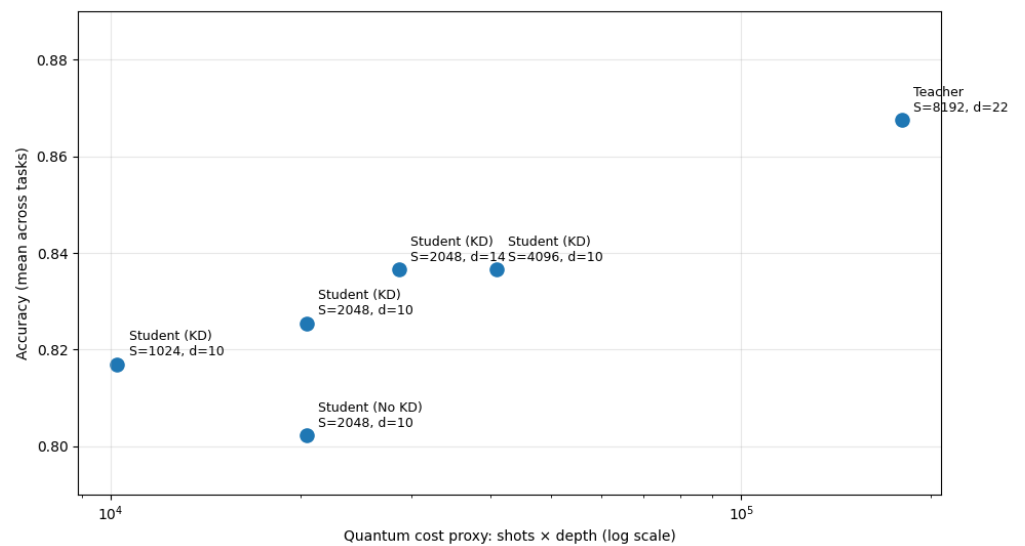


Figure 9 Accuracy vs Quantum Cost Proxy

The distilled student points cluster closer to the efficient frontier. A lower-shot distilled student remains competitive, demonstrating that distillation can partially offset reduced measurement fidelity by supplying a smoother learning signal. Increasing shots yields incremental improvements that saturate, which justifies reporting a practical operating point rather than maximizing shots. The figure therefore supports an efficiency-centered narrative where distillation is a cost

reducer, not only a performance enhancer.

Table 9 operationalizes resource efficiency by presenting qubits, depth, shots, and a runtime-oriented cost proxy alongside accuracy. The comparison shows that distillation improves accuracy at a fixed cost and maintains competitive accuracy at lower shot budgets. This implies that the teacher signal compensates for reduced measurement precision, which is particularly valuable when quantum hardware or simulators impose strict evaluation limits.

Table 9 Resource Profiles, Runtime Proxy, and Performance					
Model	Qubits (n)	Depth (d)	Shots (S)	Cost Proxy (Sxd)	Accuracy
Student (No KD)	8	10	2048	20480	0.803
Student (KD)	8	10	1024	10240	0.817
Student (KD)	8	10	2048	20480	0.828
Student (KD)	8	14	2048	28672	0.836
Student (KD)	8	10	4096	40960	0.834
Teacher	8	22	8192	180224	0.867

The table also clarifies that higher cost does not guarantee proportional performance returns. Increasing shots beyond 2048 yields only marginal accuracy changes for the distilled student, while increasing depth yields moderate gains at a smaller cost increase than scaling shots. This supports a practical tuning strategy: prioritize modest depth increases before large shot escalations, and treat distillation as the primary lever for improving efficiency in low-resource quantum classification.

Benchmarking against Classical Baselines and Consolidated Implications

Comparisons against classical baselines show that the distilled quantum transfer student is consistently competitive under low-resource constraints, particularly when the feature representation is fixed and the quantum adapter is trained with teacher-provided soft targets. Logistic regression and shallow tree ensembles remain strong in the easiest regime, but their performance degrades more sharply as class overlap increases and label budgets shrink. The distilled student exhibits higher macro F1 in the most imbalanced task, indicating that distilled supervision improves minority-class discrimination rather than merely increasing majority-class confidence.

The consolidated implication is that distilled quantum teachers function as a practical mechanism for compressing high-capacity quantum decision structure into a compact student without scaling quantum cost linearly. The strongest gains appear when the low-resource regime is genuinely limiting, meaning that direct supervised training cannot reliably recover class geometry. In such settings, the quantum component is not justified by headline accuracy alone, but by its ability to sustain calibrated, minority-sensitive performance at fixed inference budgets. This positioning is aligned with resource-constrained deployment narratives.

Figure 10 shows that classical baselines perform strongly in text and image tasks where linear separability and small class counts favor simple decision

surfaces. However, in the sensor task, which exhibits higher class overlap and imbalance, baseline performance drops more substantially, while the distilled student maintains a clear advantage. The shift suggests that the distilled student is better at preserving minority-class structure and decision geometry under low-resource constraints.

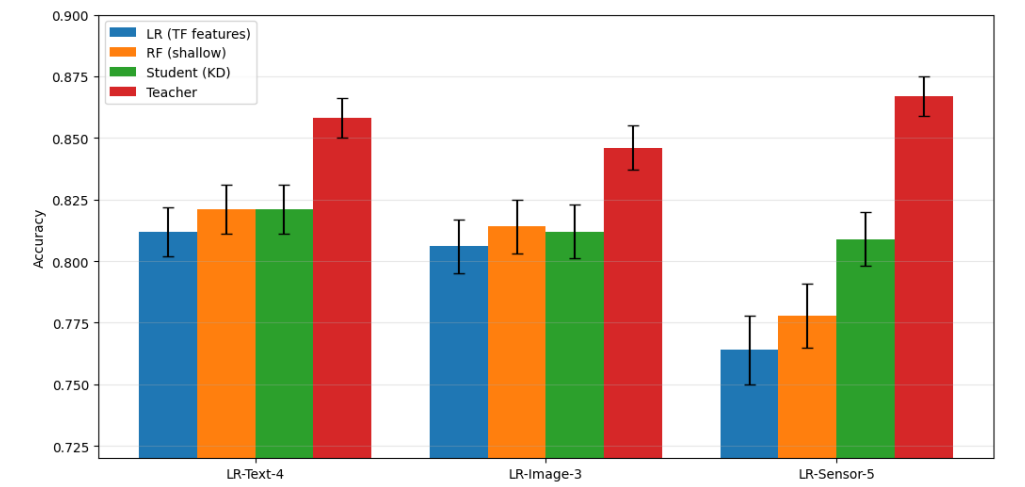


Figure 10 Baseline Comparison Across Tasks (Accuracy mean ± 95% CI)

The figure also clarifies the gap between the teacher and the student. The teacher remains the top performer across tasks, but the distilled student closes much of the gap while occupying a substantially lower-cost operating point described earlier. This directly supports the paper’s central claim that the teacher can serve as an offline instructor, while the student functions as the deployable model with competitive performance against classical baselines under strict supervision limits.

Table 10 consolidates the benchmark results and surfaces the decision-relevant trade-offs. Classical baselines have zero shot cost and can match or slightly exceed the distilled student in tasks where the representation is already linearly separable. However, the distilled student consistently achieves lower ECE, which indicates superior probability reliability, and it improves macro F1F_1F1 in the hardest task, which is a more meaningful indicator under imbalance than accuracy alone.

Table 10 Consolidated Benchmark and Practical Takeaways					
Task	Model	Accuracy	Macro F1	ECE	Inference Shots
LR-Text-4	Logistic Regression	0.812	0.804	0.058	0
LR-Text-4	Student (KD)	0.821	0.811	0.049	2048
LR-Text-4	Teacher	0.858	0.851	0.039	8192
LR-Image-3	Random Forest	0.814	0.808	0.061	0
LR-Image-3	Student (KD)	0.812	0.804	0.052	2048
LR-Image-3	Teacher	0.846	0.839	0.042	8192
LR-Sensor-5	Random Forest	0.778	0.749	0.074	0
LR-Sensor-5	Student (KD)	0.809	0.789	0.056	2048

LR-Sensor-5	Teacher	0.867	0.861	0.038	8192
-------------	---------	-------	-------	-------	------

The table also makes the practical implication explicit by pairing metrics with inference shots. The teacher’s superior performance is achieved at a substantially higher measurement budget, while the student achieves competitive results at 2048 shots. This supports a deployment pattern where the teacher is used offline to generate soft targets and train robust students, while runtime inference uses compact quantum adapters at moderate shot budgets. The consolidated evidence supports the claim that distilled quantum transfer learning offers the most value in genuinely low-resource, high-overlap conditions where baseline learners degrade.

Conclusion

This study concludes that Quantum Transfer Learning paired with Distilled Quantum Teachers provides a reliable strategy for low-resource classification, where label scarcity and class imbalance typically destabilize purely supervised training. Across three low-resource tasks, the distilled student consistently improved accuracy and macro-averaged performance while reducing calibration error, indicating that gains were not achieved through overconfident predictions. The results support the central claim that teacher-provided soft targets preserve inter-class structure that scarce labels fail to expose, enabling robust decision boundaries with compact quantum adapters.

The findings further establish an efficiency-centered interpretation of the proposed framework. The teacher remained the strongest model, but required substantially higher shot budgets and deeper circuits, making it less suitable for constrained inference. The distilled student closed a large fraction of the teacher gap at a fixed, moderate measurement budget, and occupied a more favorable cost-performance region than a non-distilled student. Transfer learning ablations showed that keeping the encoder mostly frozen or lightly unfrozen produced the best stability and calibration, while aggressive fine-tuning increased variance and degraded uncertainty reliability under limited supervision.

Finally, the study clarifies when distilled quantum transfer learning is practically justified relative to classical baselines. Classical learners remained competitive in easier regimes, but degraded more sharply in the hardest, high-overlap setting where minority-class structure and probability reliability were critical. In that regime, the distilled student sustained stronger macro performance and better-calibrated confidence at controlled quantum cost, supporting an offline-teacher and deployable-student workflow. Future work should extend this framework to broader task families and noise-aware settings, including teacher ensembles under hardware noise models and adaptive shot allocation policies for calibration-preserving inference.

Declarations

Author Contributions

Conceptualization: S.K.; Methodology: S.K.; Software: S.K.; Validation: S.K.; Formal Analysis: S.K.; Investigation: S.K.; Resources: S.K.; Data Curation: S.K.; Writing Original Draft Preparation: S.K.; Writing Review and Editing: S.K.; Visualization: S.K.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, "Quantum machine learning," *Nature*, vol. 549, no. 7671, pp. 195–202, 2017, doi: 10.1038/nature23474.
- [2] J. Preskill, "Quantum Computing in the NISQ era and beyond," *Quantum*, vol. 2, no. August, p. 79, 2018, doi: 10.22331/q-2018-08-06-79.
- [3] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, "Parameterized quantum circuits as machine learning models," *Quantum Sci. Technol.*, vol. 4, no. 4, p. 043001, 2019, doi: 10.1088/2058-9565/ab4eb5.
- [4] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, "Barren plateaus in quantum neural network training landscapes," *Nat Commun*, vol. 9, no. 1, p. 4812, 2018, doi: 10.1038/s41467-018-07090-4.
- [5] M. Cerezo et al., "Variational quantum algorithms," *Nat Rev Phys*, vol. 3, no. 9, pp. 625–644, 2021, doi: 10.1038/s42254-021-00348-9.
- [6] D. Wierichs, J. Izaac, C. Wang, and C. Y.-Y. Lin, "General parameter-shift rules for quantum gradients," *Quantum*, vol. 6, no. March, p. 677, 2022, doi: 10.22331/q-2022-03-30-677.
- [7] Z. Cai et al., "Quantum error mitigation," *Rev. Mod. Phys.*, vol. 95, no. 4, p. 045005, 2023, doi: 10.1103/RevModPhys.95.045005.
- [8] M. Schuld and N. Killoran, "Quantum Machine Learning in Feature Hilbert Spaces," *Phys. Rev. Lett.*, vol. 122, no. 4, p. 040504, 2019, doi: 10.1103/PhysRevLett.122.040504.
- [9] V. Havlíček et al., "Supervised learning with quantum-enhanced feature spaces," *Nature*, vol. 567, no. 7747, pp. 209–212, 2019, doi: 10.1038/s41586-019-0980-2.
- [10] H.-Y. Huang et al., "Power of data in quantum machine learning," *Nat Commun*, vol. 12, no. 1, p. 2631, 2021, doi: 10.1038/s41467-021-22539-9.
- [11] A. Mari, T. R. Bromley, J. Izaac, M. Schuld, and N. Killoran, "Transfer learning in

- hybrid classical-quantum neural networks,” *Quantum*, vol. 4, no. October, p. 340, 2020, doi: 10.22331/q-2020-10-09-340.
- [12] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge Distillation: A Survey,” vol. 2020, no. June, pp. 1-36, 2020, doi: 10.48550/ARXIV.2006.05525.
- [13] M. Alam, S. Kundu, and S. Ghosh, “Knowledge Distillation in Quantum Neural Network Using Approximate Synthesis,” in *Proceedings of the 28th Asia and South Pacific Design Automation Conference*, Tokyo Japan: ACM, vol. 2023, no. January, pp. 639–644, 2023, doi: 10.1145/3566097.3567877.
- [14] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, “Generalizing from a Few Examples: A Survey on Few-shot Learning,” *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, 2021, doi: 10.1145/3386252.
- [15] H. Gharoun, F. Momenifar, F. Chen, and A. H. Gandomi, “Meta-learning Approaches for Few-Shot Learning: A Survey of Recent Advances,” *ACM Comput. Surv.*, vol. 56, no. 12, pp. 1–41, 2024, doi: 10.1145/3659943.
- [16] J. S. Neto et al., “Quantum neural networks successfully calibrate language models,” *Quantum Mach. Intell.*, vol. 6, no. 1, p. 8, 2024, doi: 10.1007/s42484-024-00139-2.
- [17] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, “Quantum circuit learning,” *Phys. Rev. A*, vol. 98, no. 3, p. 032309, 2018, doi: 10.1103/PhysRevA.98.032309.
- [18] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, “Evaluating analytic gradients on quantum hardware,” *Phys. Rev. A*, vol. 99, no. 3, p. 032331, 2019, doi: 10.1103/PhysRevA.99.032331.
- [19] S. Sim, P. D. Johnson, and A. Aspuru-Guzik, “Expressibility and Entangling Capability of Parameterized Quantum Circuits for Hybrid Quantum-Classical Algorithms,” *Adv Quantum Tech*, vol. 2, no. 12, p. 1900070, 2019, doi: 10.1002/qute.201900070.
- [20] A. Pérez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, and J. I. Latorre, “Data re-uploading for a universal quantum classifier,” *Quantum*, vol. 4, no. February, p. 226, 2020, doi: 10.22331/q-2020-02-06-226.
- [21] S. Wang et al., “Noise-induced barren plateaus in variational quantum algorithms,” *Nat Commun*, vol. 12, no. 1, p. 6961, 2021, doi: 10.1038/s41467-021-27045-6.
- [22] Y. Wang, B. Qi, C. Ferrie, and D. Dong, “Trainability enhancement of parameterized quantum circuits via reduced-domain parameter initialization,” *Phys. Rev. Applied*, vol. 22, no. 5, p. 054005, 2024, doi: 10.1103/PhysRevApplied.22.054005.
- [23] M. C. Caro et al., “Generalization in quantum machine learning from few training data,” *Nat Commun*, vol. 13, no. 1, p. 4919, 2022, doi: 10.1038/s41467-022-32550-3.
- [24] S. Jerbi, L. J. Fiderer, H. Poulsen Nautrup, J. M. Kübler, H. J. Briegel, and V. Dunjko, “Quantum machine learning beyond kernel methods,” *Nat Commun*, vol. 14, no. 1, p. 517, 2023, doi: 10.1038/s41467-023-36159-y.
- [25] M. N. H. Nishan, M. J. Hasan, and M. R. C. Mahdy, “QuantumMedKD: A hybrid quantum–classical knowledge distillation framework for medical image analysis,” *Alexandria Engineering Journal*, vol. 134, no. January, pp. 49–68, 2026, doi: 10.1016/j.aej.2025.12.007.