QMoE+: Hybrid Quantum Mixture of Experts

Anonymous authors
Paper under double-blind review

Abstract

Quantum mixture of experts (QMoE) extends conditional computation to the NISQ setting by distributing learning across parameterized quantum circuit (PQC) experts selected via a routing mechanism. Existing approaches are limited by single-block experts, lack of load balancing, and aggregation schemes that ignore routing amplitudes. We propose **QMoE+**, which uses two-block data re-uploading experts with learnable offsets, a coherent aggregation circuit over the joint routing-data Hilbert space, and a Switch-style load-balancing loss. Under top-k=1 sparse routing, QMoE+ activates only ∼28% of its parameters per inference while achieving consistent accuracy gains across seven datasets and four gate sets, winning 27/28 configurations with a mean improvement of +5.11% in the noiseless setting and +4.71% under depolarising noise. A decomposed ablation further shows that quantum coherence in the aggregation circuit outperforms an incoherent baseline in all seven datasets under p=0.01 noise (mean +1.80%), establishing an independent contribution beyond learnable aggregation parameters alone. Ablations confirm that load balancing is consistently beneficial, while data re-uploading provides the largest gains on complex tasks. Code is available at https://anonymous.4open.science/r/qmoe-plus/.

1 Introduction

The ability to conditionally compute and activate only the model components most relevant to a given input has become a central principle in modern machine learning. Originating with the *adaptive mixture of local experts* (Jacobs et al., 1991), this idea uses a gating network to partition the input space and assign specialised sub-networks. Subsequent work placed this framework on firmer probabilistic foundations (Jordan & Jacobs, 1993) and showed that sparsely activating the top-k experts enables large effective model capacity at low computational cost (Shazeer et al., 2017). This paradigm now underpins large-scale systems such as GShard, Switch Transformer, ST-MoE, and Mixtral (Lepikhin et al., 2021; Fedus et al., 2022; Zoph et al., 2022; Jiang et al., 2024). However, sparse MoE training introduces well-known challenges, including non-differentiable routing, load imbalance leading to expert collapse, and additional optimisation complexity requiring carefully designed auxiliary losses.

In parallel, quantum machine learning (QML) explores whether similar gains can be achieved using quantum systems, leveraging superposition, entanglement, and interference (Biamonte et al., 2017; Cerezo et al., 2021a). In the NISQ era (Preskill, 2018), the dominant framework is the *parameterized quantum circuit* (PQC), a fixed circuit with trainable parameters optimised via a hybrid quantum-classical loop (Benedetti et al., 2019; Mitarai et al., 2018). PQCs can represent expressive function classes through entanglement and *data re-uploading* (Pérez-Salinas et al., 2020; Schuld et al., 2021), but their practical scalability is limited by barren plateaus and hardware noise (McClean et al., 2018; Holmes et al., 2022). These constraints make large monolithic circuits difficult to train, motivating modular architectures that distribute learning across smaller, more tractable quantum components.

The natural intersection of these two lines of work, conditional computation and quantum learning, is the quantum mixture of experts. Several recent contributions have begun to explore this space from different angles. Nguyen et al. (2025) introduce QMoE, to our knowledge the first architecture to realise both the experts and the routing mechanism as quantum circuits: a variational router selects among four parallel PQC experts via computational-basis measurement, and the selected expert outputs are aggregated clas-

sically. Concurrent work by Bhati et al. (2025) employs a *classical* gating network over a heterogeneous pool of quantum experts - including quantum SVMs, QKNNs, QCNNs, and standard QNNs - and reports generalisation improvements on Iris, Titanic, health insurance, and MNIST digit pairs. A different angle is pursued by Heddad & Bouanane (2025), who retain classical experts and use a quantum router to isolate the contribution of quantum interference to the routing decision, observing that wave-interference-based routing achieves superior non-linear separation on the Two Moons benchmark with greater parameter efficiency. Tognini et al. (2025) explore a globally coupled mixture trained end-to-end on MNIST. Together, these works establish that combining quantum circuits with mixture-of-experts structure is both feasible and empirically promising.

A closer reading reveals three shared limitations. First, all prior approaches use single-block PQC experts, which restrict the range of Fourier components they can represent and limit expert specialisation (Pérez-Salinas et al., 2020; Schuld et al., 2021). Second, none include an explicit mechanism for load balancing. As established in classical MoE literature, training without such regularisation tends to concentrate routing on a few experts, leading to collapse (Shazeer et al., 2017; Fedus et al., 2022), and there is no reason to expect quantum models to behave differently. Third, existing aggregation schemes treat routing purely as a selection mechanism: once an expert is chosen, the magnitude of routing amplitudes does not influence the final output. This prevents the model from weighting experts based on confidence. To address this, we introduce a coherent aggregation circuit that operates over the joint routing-data Hilbert space, enabling amplitude-aware expert combination.

Our Contributions. We present **QMoE+**, a hybrid quantum mixture of experts that addresses these limitations. Our contributions are:

1. **Data re-uploading experts.** We replace single-block PQC experts with two-block DRU experts $U(\mathbf{x}) \to V_1 \to U(\mathbf{x} + \boldsymbol{\phi}) \to V_2$, where $\boldsymbol{\phi}$ is a learnable per-expert offset. This improves expressivity and enables specialisation, yielding the largest accuracy gains in our ablations (Pérez-Salinas et al., 2020).

2. **Coherent aggregation.** We introduce a joint routing-data circuit that entangles routing amplitudes with expert states before measurement. This allows amplitude-weighted expert combination, rather than treating routing as a purely classical selection signal.

3. **Load balancing and sparse routing.** We add load-balancing regularisation to prevent expert collapse, using entropy-based loss for dense routing and Switch-style loss for top-k routing (Fedus et al., 2022). This enables stable training and effective sparsity (e.g., k=1, K=4).

4. **Comprehensive evaluation.** We evaluate across multiple datasets, gate sets, and both noiseless and noisy (p=0.01) settings. QMoE+ consistently outperforms the QMoE baseline (Nguyen et al., 2025), with ablations confirming that each component contributes independently.

The remainder of the paper is organised as follows. Section 2 reviews classical MoE, QML, and prior QMoE work. Section 3 introduces notation and preliminaries. Section 4 presents the QMoE+ architecture, and Section 5 describes the experimental setup. Section 6 reports results under noiseless and noisy conditions, and Section 7 concludes. Additional theoretical analysis and implementation details are provided in Appendices D and E.

2 Related Work

The mixture-of-experts framework originates with Jacobs et al. (1991) and Jordan & Jacobs (1993), and was scaled to modern systems through sparse top-k routing with auxiliary load-balancing losses to prevent expert collapse (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022; Zoph et al., 2022; Jiang et al., 2024). We adopt the Switch Transformer formulation of Fedus et al. (2022) for our sparse routing regime. On the quantum side, parameterized quantum circuits in the NISQ regime (Benedetti et al., 2019; Cerezo et al., 2021b; Bharti et al., 2022) are limited by barren plateaus (McClean et al., 2018; Holmes et al., 2022;

Larocca et al., 2025) and by the restricted Fourier spectrum of single-encoding circuits (Schuld et al., 2021), motivating data re-uploading (Pérez-Salinas et al., 2020) as a route to higher expressivity. Three recent works combine these threads. Nguyen et al. (2025) introduce a fully quantum MoE with single-block PQC experts and classical aggregation, which serves as our primary baseline. Bhati et al. (2025) pair a classical router with heterogeneous quantum experts, Heddad & Bouanane (2025) use a quantum router with classical experts, and Tognini et al. (2025) train a globally coupled quantum mixture. None of these jointly addresses multi-block expert expressivity, load balancing, and coherent aggregation. An extended discussion appears in Appendix A.

3 Background

3.1 Classical Mixture of Experts

A mixture of experts (MoE) decomposes a complex learning task across K specialised sub-networks. Formally, given an input $\mathbf{x} \in \mathbb{R}^d$, a gating network $\mathcal{G}(\mathbf{x};\, \psi) = [g_1(\mathbf{x}), \ldots, g_K(\mathbf{x})]$ produces a probability distribution over the K experts, where $g_k(\mathbf{x}) \geq 0$ and $\sum_k g_k(\mathbf{x}) = 1$. Each expert $E_k(\mathbf{x};\, \omega_k)$ produces a prediction, and the combined output is

$$\mathbf{y}(\mathbf{x}) \;=\; \sum_{k=1}^{K} g_k(\mathbf{x})\, E_k(\mathbf{x};\, \omega_k). \tag{1}$$

Gating network and experts are trained jointly via gradient descent, enabling the gate to learn which expert is most suited to each region of the input space while experts simultaneously specialise to their assigned regions (Jacobs et al., 1991; Jordan & Jacobs, 1993).

In modern large-scale systems, the dense sum of Eq. (1) is replaced by *sparse* gating: only the top-k experts (with $k \ll K$) are activated per input, reducing forward-pass cost while preserving overall model capacity (Shazeer et al., 2017; Fedus et al., 2022). Because top-k selection is non-differentiable, training typically uses the straight-through estimator (Bengio et al., 2013) or Gumbel-softmax relaxations (Jang et al., 2017; Maddison et al., 2017) to pass gradients through the discrete routing decision.

A persistent challenge in MoE training is *expert collapse*: the gate concentrates routing weight on a small subset of experts, leaving the remainder without meaningful gradient signal and effectively wasted. This is a stable failure mode whenever the routing and task losses are coupled - whichever expert is marginally better early in training receives more data, sharpens its advantage, and eventually monopolises the gate. To counteract this, MoE systems universally include an auxiliary *load-balancing loss* that penalises uneven expert utilisation (Shazeer et al., 2017; Fedus et al., 2022; Zoph et al., 2022). The Switch Transformer formulation (Fedus et al., 2022), which we adopt for our sparse routing regime, computes

$$\mathcal{L}_{\text{load}} \;=\; K \sum_{k=1}^{K} f_k \cdot P_k, \tag{2}$$

where f_k is the fraction of the current batch dispatched to expert k (stop-gradient) and P_k is the mean routing probability for expert k (has gradient). Minimising Eq. (2) drives gradient into the routing parameters in proportion to over-utilisation without requiring a differentiable dispatch function.

3.2 Parameterized Quantum Circuits

An n-qubit Parameterized quantum circuit (PQC) consists of a data-encoding unitary $U(\mathbf{x})$, a parameterized variational unitary $V(\boldsymbol{\theta})$, and a final measurement of a Hermitian observable $\mathcal{O}$. The circuit expectation is

$$f(\mathbf{x}, \boldsymbol{\theta}) \;=\; \langle \mathcal{O} \rangle_{\mathbf{x},\boldsymbol{\theta}} \;=\; \langle 0^{\otimes n} |\, U^{\dagger}(\mathbf{x})\, V^{\dagger}(\boldsymbol{\theta})\, \mathcal{O}\, V(\boldsymbol{\theta})\, U(\mathbf{x}) \,|0^{\otimes n}\rangle\,. \tag{3}$$

Parameters $\boldsymbol{\theta}$ are optimised through a hybrid quantum-classical loop using parameter-shift gradient rules (Mitarai et al., 2018) or automatic differentiation frameworks (Wang et al., 2022a). The choice of encoding $U(\mathbf{x})$ and variational ansatz $V(\boldsymbol{\theta})$ jointly determine the expressivity and trainability of the circuit (Benedetti et al., 2019; Cerezo et al., 2021a).

Encoding. We use *angle encoding* (phase encoding) with a two-gate, multi-block structure. For an n-qubit register and input $\mathbf{x} \in \mathbb{R}^d$ with $d = 2nB$ for integer B, the encoding is applied in B successive blocks. In block b, feature pair $(x_{b\cdot 2n+2q},\, x_{b\cdot 2n+2q+1})$ is encoded on qubit q as $R_X(x_{b\cdot 2n+2q})\, R_Y(x_{b\cdot 2n+2q+1})$, followed by a CNOT entanglement ring $q \to (q+1) \bmod n$:

$$U(\mathbf{x}) \;=\; \prod_{b=0}^{B-1} \left[\mathrm{CNOT}_{\mathrm{ring}} \cdot \bigotimes_{q=0}^{n-1} R_X(x_{b,q,0})\, R_Y(x_{b,q,1}) \right]. \tag{4}$$

This structure ensures that each feature is encoded once per block, and the CNOT ring introduces entanglement between qubits after each encoding step.

Variational ansatz. Each variational layer applies parameterized rotations followed by a CNOT ring. The rotation gate set is one of $\{R_X,\, R_Y,\, R_X R_Y,\, R_X R_Y R_Z\}$, where the choice determines the number of parameters per qubit per layer (1, 1, 2, or 3 respectively). With L variational layers and n qubits the total variational parameter count per circuit block is $L \cdot n \cdot |\mathcal{S}|$, where $|\mathcal{S}| \in \{1, 2, 3\}$ is the gate set size.

Measurement. Class logits are extracted as Pauli-Z expectation values on the first C qubits (where C is the number of classes):

$$\ell_j \;=\; \langle Z_j \rangle \;=\; \sum_{s \in \{0,1\}^n} |\psi_s|^2\, (-1)^{s_j}, \qquad j = 0, \ldots, C-1, \tag{5}$$

where $|\psi_s|^2$ is the Born-rule probability of computational basis state s and s_j denotes its j-th bit.

Barren plateaus. A well-known trainability obstacle for PQCs is the *barren plateau* (McClean et al., 2018): for sufficiently deep or expressive circuits the variance of the cost gradient vanishes exponentially with system size, making gradient-based optimisation infeasible. This problem is exacerbated by high ansatz expressibility (Holmes et al., 2022) and by hardware noise, which induces a separate, noise-driven form of exponential gradient concentration (Wang et al., 2022b; Larocca et al., 2025). Modular architectures that keep each constituent circuit shallow - such as the expert circuits in our work - partially mitigate barren plateaus by ensuring that no single circuit is deep enough to enter the plateau regime.

3.3 Data Re-uploading

Data re-uploading (DRU) (Pérez-Salinas et al., 2020) is the technique of interleaving the encoding unitary with trainable variational blocks, rather than applying encoding only once. A two-block DRU circuit takes the form

$$|\psi(\mathbf{x}, \boldsymbol{\theta}^1, \boldsymbol{\theta}^2)\rangle \;=\; V_2(\boldsymbol{\theta}^2)\, U(\mathbf{x})\, V_1(\boldsymbol{\theta}^1)\, U(\mathbf{x})\, |0\rangle^{\otimes n}. \tag{6}$$

The motivation for DRU is formal: Schuld et al. (2021) show that the output of a PQC can be written as a truncated Fourier series in $\mathbf{x}$, whose accessible frequencies are determined solely by the data-encoding gates. A single encoding layer of Pauli rotations gives access to only first-order harmonics; each additional encoding layer expands the frequency spectrum. Consequently, a DRU circuit with L encoding blocks can represent functions of substantially higher frequency than a single-encoding PQC with L variational layers, making DRU a theoretically motivated approach to increasing expressivity without increasing the number of trainable parameters (Pérez-Salinas et al., 2020; Goto et al., 2021).

3.4 The QMoE Framework

The QMoE of Nguyen et al. (2025) is, to our knowledge, the first architecture in which both routing and experts are implemented as parameterized quantum circuits. We describe it in our notation, as it serves as the primary baseline.

The model uses a routing register (n_R qubits, $K{=}2^{n_R}$ experts) and a data register (n_D qubits). Input $\mathbf{x} \in \mathbb{R}^d$ is encoded into both via $U(\mathbf{x})$. A routing circuit $G(\boldsymbol{\theta}_G)$ is applied and measured to obtain probabilities

$p_k = |\langle k|G(\boldsymbol{\theta}_G)U(\mathbf{x})|0\rangle|^2$. These act as classical weights: controlled operations apply each expert $E_k(\boldsymbol{\theta}_k)$ to the data register, followed by a fixed aggregation layer and terminal Z-basis measurement for logits.

The model output is:

$$\mathbf{y}(\mathbf{x}) \;=\; \text{Measure}\left(\text{Agg}\left(\sum_{k=1}^{K} p_k(\mathbf{x})\, E_k(\boldsymbol{\theta}_k)\, U(\mathbf{x})\, |0\rangle^{\otimes n_D}\right)\right), \tag{7}$$

Training is performed end-to-end via parameter-shift gradients.

Three properties are central. First, each expert is a single-block PQC, limiting expressivity to first-order harmonics (Schuld et al., 2021). Second, routing probabilities are classical, and amplitude information is discarded. Third, aggregation is fixed and non-learnable. These design choices define the limitations addressed by QMoE+ (Section 4).

4 Methodology

QMoE+ is a hybrid quantum mixture of experts in which routing, expert computation, and aggregation are all parameterised quantum circuits trained end-to-end. It extends the QMoE framework of Nguyen et al. (2025) through three targeted architectural changes: data re-uploading inside each expert, a low-dimensional routing input with coherent amplitude-based aggregation, and an explicit load-balancing auxiliary loss. We describe each component in turn using the notation established in Section 3, then state the composite training objective. Formal theoretical justifications are deferred to Appendix D; physical realisability on NISQ hardware is discussed in Appendix E.

4.1 Data Re-Uploading Experts

Each of the K=4 experts E_k is a two-block DRU circuit on n_D=4 data qubits. Starting from $|0\rangle^{\otimes n_D}$, the circuit applies phase encoding $U(\mathbf{x})$, followed by a variational block $V_k^{(1)}(\boldsymbol{\theta}_k^{(1)})$, a second encoding of the shifted input $U(\mathbf{x}+\boldsymbol{\phi}_k)$, and a second variational block $V_k^{(2)}(\boldsymbol{\theta}_k^{(2)})$:

$$|\psi_k(\mathbf{x})\rangle \;=\; V_k^{(2)}(\boldsymbol{\theta}_k^{(2)})\, U(\mathbf{x}+\boldsymbol{\phi}_k)\, V_k^{(1)}(\boldsymbol{\theta}_k^{(1)})\, U(\mathbf{x})\, |0\rangle^{\otimes n_D}, \tag{8}$$

where $\boldsymbol{\phi}_k \in \mathbb{R}^d$ is a learnable per-expert re-upload offset, initialised at $\mathbf{0}$. Each variational block consists of L=2 layers of single-qubit rotations from gate set $\mathcal{S} \in \{\text{RX, RY, RX+RY, RX+RY+RZ}\}$ followed by a CNOT ring.

The offset $\boldsymbol{\phi}_k$ drives expert specialisation: although experts are identical at initialisation, training induces divergence, giving each expert a distinct input view. The second encoding shifts the Fourier spectrum accessible to each expert (Appendix D.1).

Logits are obtained as Z-expectation values on the first C data qubits (Eq. (5)). The parameter count per expert under RX+RY+RZ is $2 \cdot L \cdot n_D \cdot 3 + d = 2 \cdot 2 \cdot 4 \cdot 3 + 64 = 112$.

4.2 Routing Circuit

The routing circuit $G(\boldsymbol{\theta}_G)$ operates on n_R=2 qubits (K=4 experts). It begins with a Hadamard layer, followed by phase encoding of a low-dimensional input projection $\mathbf{x}_{1:8}$ and L=2 variational layers:

$$|\alpha(\mathbf{x})\rangle \;=\; V_G(\boldsymbol{\theta}_G)\, U(\mathbf{x}_{1:8})\, H^{\otimes n_R}\, |0\rangle^{\otimes n_R}, \tag{9}$$

yielding a complex amplitude vector $\boldsymbol{\alpha}(\mathbf{x}) \in \mathbb{C}^K$, which is passed coherently to aggregation without measurement.

Routing uses only 8 features to avoid capacity imbalance: encoding all 64 features would require 32 encoding blocks on 2 qubits, overwhelming the 2-layer variational circuit and biasing routing toward encoding geometry. Restricting to 8 features yields one encoding block per variational layer, maintaining a balanced expressivity budget.

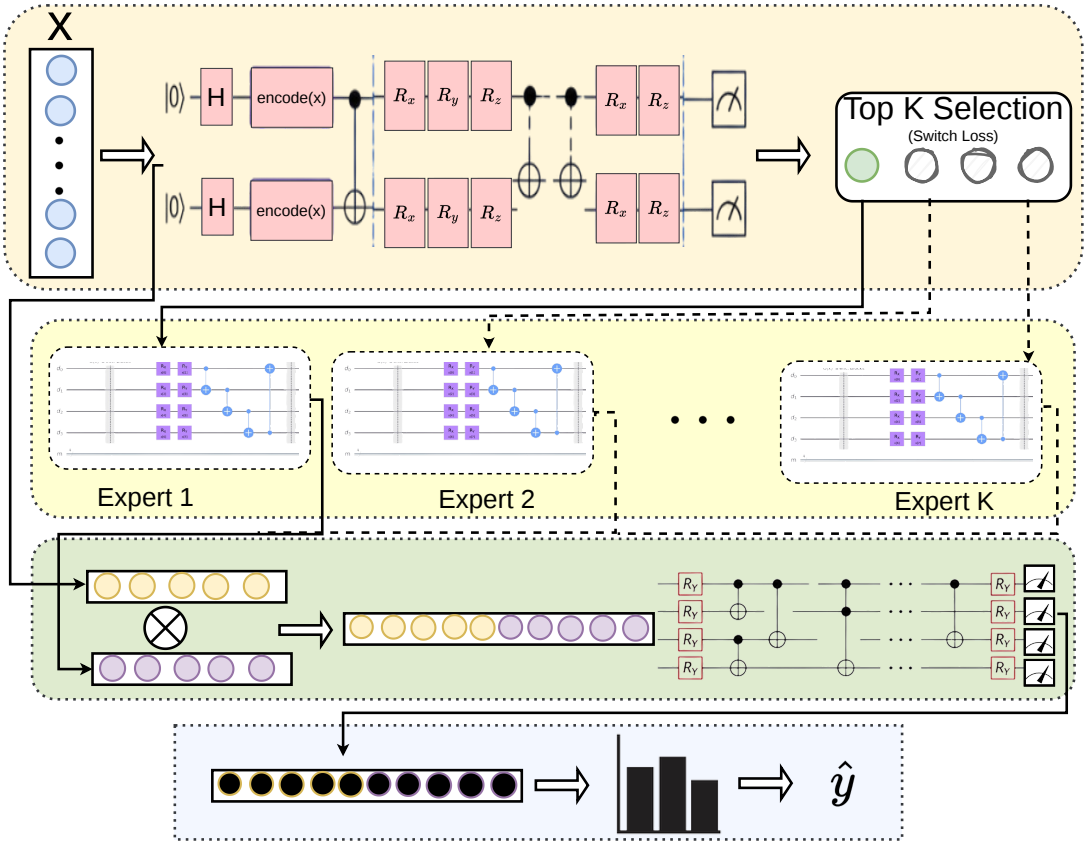

Figure 1: Overview of QMoE+. The routing circuit $G(\boldsymbol{\theta}_G)$ encodes $\mathbf{x}$ on n_R qubits (Hadamard + phase encoding + L variational layers) to produce probabilities $p_k{=}|\alpha_k|^2$. Top-$k{=}1$ selects expert k^*, with Switch load-balancing $\mathcal{L}_{\text{load}}{=}K\sum_k f_k P_k$ ($\lambda{=}0.01$) preventing collapse. Only k^* is evaluated. Each expert is a two-block DRU circuit $U(\mathbf{x}) \to V_1 \to U(\mathbf{x}{+}\boldsymbol{\phi}) \to V_2$, with learnable offset $\boldsymbol{\phi}$ enabling specialisation. The joint state $\tilde{\alpha}_{k^*}|k^*\rangle_R \otimes |e_{k^*}\rangle_D$ is processed by a coherent aggregation circuit $W(\boldsymbol{\varphi})$ (R_Y + CNOT, $L_{\text{agg}}{=}2$), and Z-measurements on data qubits produce logits, followed by softmax outputs $\hat{y}$.

4.3 Coherent Aggregation

Given the routing state $|\alpha(\mathbf{x})\rangle = \sum_{k=0}^{K-1} \alpha_k(\mathbf{x})\,|k\rangle_R$ and the K expert states $\{|\psi_k(\mathbf{x})\rangle\}_k$, we form the joint routing-data state

$$|\Psi(\mathbf{x})\rangle \;=\; \sum_{k=0}^{K-1} \alpha_k(\mathbf{x})\,|k\rangle_R \otimes |\psi_k(\mathbf{x})\rangle_D \tag{10}$$

on $n_R + n_D = 6$ total qubits. A learnable variational circuit $W(\boldsymbol{\varphi})$ is then applied over the entire joint register: $L_{\text{agg}}{=}2$ layers, each consisting of R_Y rotations on every qubit followed by a CNOT ring with wrap-around. Class logits are extracted as Z-expectation values on the data qubits:

$$\text{logit}_j(\mathbf{x}) \;=\; \langle\Psi(\mathbf{x})|\,W^\dagger(\boldsymbol{\varphi})\,Z_{n_R+j}\,W(\boldsymbol{\varphi})\,|\Psi(\mathbf{x})\rangle, \quad j = 0,\ldots,C{-}1. \tag{11}$$

The design differs from the fixed controlled-gate aggregation of Nguyen et al. (2025) in two respects. First, $W(\boldsymbol{\varphi})$ acts on all $n_R + n_D$ qubits jointly, including both the routing and data registers, rather than only routing-to-data controlled operations; this means the learned transformation can couple the routing subspace to the data subspace. Second, $W(\boldsymbol{\varphi})$ is variational and trained end-to-end, so it can adapt the aggregation

to the data distribution rather than applying a fixed circuit structure. The routing amplitudes $\alpha_k(\mathbf{x})$ are carried as complex values into the joint state and are not collapsed by intermediate measurement before aggregation; the single terminal measurement on the data register is the only point at which quantum information is converted to classical output. The structural motivation for this design - that discarding the routing state by measurement before aggregation loses the amplitude magnitudes as a conditioning signal - is discussed further in Appendix D.3. The joint-state construction of instantiates the PREPARE-SELECT primitive of the linear combination of unitaries framework (Childs & Wiebe, 2012); our contribution is the application of this structure to mixture-of-experts aggregation and the addition of a learnable variational circuit $W(\varphi)$ acting on the joint routing-data register downstream of the state preparation.

4.4 Load-Balancing Regularisation

Let $p_k(\mathbf{x}) = |\alpha_k(\mathbf{x})|^2$ denote the routing probability for expert k, and $\bar{e}_k = \mathbb{E}_{\mathbf{x}\in\mathcal{B}}[p_k(\mathbf{x})]$ the batch-mean distribution. For dense routing, we use a negative-entropy penalty:

$$\mathcal{L}_{\text{load}}^{\text{dense}} \;=\; -H(\bar{\mathbf{e}}) \;=\; \sum_{k=0}^{K-1} \bar{e}_k \log \bar{e}_k, \tag{12}$$

which encourages uniform expert utilisation and mitigates collapse (Shazeer et al., 2017; Fedus et al., 2022).

For sparse (k=1) routing, we adopt the Switch Transformer loss:

$$\mathcal{L}_{\text{load}}^{\text{sparse}} \;=\; K \sum_{k=0}^{K-1} f_k \cdot P_k, \tag{13}$$

where f_k is the fraction of samples routed to expert k (stop-gradient), and $P_k = \bar{e}_k$ retains gradient. This provides a stable signal without requiring differentiable dispatch.

The full objective is:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{CE}} \;+\; \lambda(t)\, \mathcal{L}_{\text{load}}, \tag{14}$$

with a linear warm-up:

$$\lambda(t) = \begin{cases} 0 & t < t_{\text{start}} \\ \lambda_{\text{target}} \cdot \dfrac{t - t_{\text{start}}}{t_{\text{end}} - t_{\text{start}}} & t_{\text{start}} \le t < t_{\text{end}} \\ \lambda_{\text{target}} & t \ge t_{\text{end}} \end{cases} \tag{15}$$

where t_{start}=5, t_{end}=15, and λ_{target}=0.01. This avoids early over-regularisation before experts specialise (Fedus et al., 2022; Zoph et al., 2022).

The load loss depends only on $p_k = |\alpha_k|^2$, so gradients affect routing parameters $\boldsymbol{\theta}_G$ only, leaving expert and aggregation parameters unchanged.

4.5 Sparse Top-k Routing

QMoE+ uses top-k=1 sparse routing with K=4 experts. For each input, routing amplitudes are computed, the top expert is selected via $\arg\max_k |\alpha_k|^2$, and the remaining amplitudes are zeroed and renormalised. Only the selected expert is evaluated per sample, while the others are skipped. With load balancing (Eq. (13)), each expert processes approximately one quarter of the batch, yielding a 4× reduction in computation relative to dense routing without loss in accuracy.

Sparsification is applied in amplitude space prior to the joint state construction (Eq. (10)). The sparse vector $\tilde{\boldsymbol{\alpha}}$ retains the top-k entries of $\boldsymbol{\alpha}$ and is renormalised, ensuring that the aggregation circuit $W(\boldsymbol{\varphi})$ operates on a valid normalised state and preserves coherence.

Algorithm 1 QMoE+ Forward Pass (sparse top-k=1 variant)

Require: $\mathbf{x} \in [0,\pi]^d$; $\{\boldsymbol{\theta}_k^{(1)}, \boldsymbol{\theta}_k^{(2)}, \boldsymbol{\phi}_k\}_{k=0}^{K-1}$; $\boldsymbol{\theta}_G$; $\boldsymbol{\varphi}$
Ensure: Logits $\mathbf{l} \in \mathbb{R}^C$
1: **Routing:** $|\alpha(\mathbf{x})\rangle \leftarrow V_G(\boldsymbol{\theta}_G)\, U(\mathbf{x}_{1:8})\, H^{\otimes n_R} |0\rangle^{\otimes n_R}$ ▷ Eq. (9)
2: $k^* \leftarrow \arg\max_k |\alpha_k(\mathbf{x})|^2$; $\tilde{\alpha}_k \leftarrow \alpha_k \cdot \mathbf{1}[k = k^*] \,/\, |\alpha_{k^*}|$
3: **Expert evaluation:** $|\psi_k(\mathbf{x})\rangle \leftarrow V_k^{(2)}\, U(\mathbf{x}{+}\boldsymbol{\phi}_k)\, V_k^{(1)}\, U(\mathbf{x})\, |0\rangle^{\otimes n_D}$ for $k = k^*$ ▷ Eq. (8) (set $|\psi_k\rangle = \mathbf{0}$ for $k \neq k^*$)
4: **Joint state:** $|\Psi(\mathbf{x})\rangle \leftarrow \sum_k \tilde{\alpha}_k\, |k\rangle_R \otimes |\psi_k(\mathbf{x})\rangle_D$ ▷ Eq. (10)
5: **Aggregation:** $\mathbf{l} \leftarrow \left[\langle Z_{n_R+j}\rangle_{W(\boldsymbol{\varphi})|\Psi\rangle}\right]_{j=0}^{C-1}$ ▷ Eq. (11)
6: **return** $\mathbf{l}$

4.6 Complete Architecture

Figure 1 summarises the complete QMoE+ forward pass. The architecture has the following parameter count in the RX+RY+RZ configuration:

- **Routing circuit**: $L \cdot n_R \cdot |\mathcal{S}| = 2 \cdot 2 \cdot 3 = 12$ parameters.
- **Each DRU expert**: $2 \cdot L \cdot n_D \cdot |\mathcal{S}| + d = 48 + 64 = 112$ parameters.
- **Aggregation circuit**: $L_{\text{agg}} \cdot (n_R + n_D) \cdot 1 = 2 \cdot 6 \cdot 1 = 12$ parameters (RY only).
- **Total**: $12 + 4 \times 112 + 12 = \mathbf{472}$ parameters.

All parameters are optimised jointly with the Adam optimiser (Kingma & Ba, 2015) at learning rate 2×10^{-3}. Expert parameters $\boldsymbol{\theta}_k^{(1)}, \boldsymbol{\theta}_k^{(2)}$ are initialised from $\mathcal{N}(0, 1/\sqrt{n_{\text{par}}})$; re-upload offsets $\boldsymbol{\phi}_k$ are initialised to $\mathbf{0}$; aggregation parameters $\boldsymbol{\varphi}$ are initialised from $\mathcal{N}(0, 0.01)$ to keep the initial aggregation circuit close to the identity.

5 Experimental Details

5.1 Datasets

We evaluate on seven classification benchmarks spanning image, synthetic, and tabular data.

Image datasets. We use MNIST (LeCun et al., 2010) and Fashion-MNIST (Xiao et al., 2017) in binary (2-class) and four-class settings. Images are resized from 28×28 to 8×8, flattened to 64 dimensions, and scaled to $[0, \pi]$ for angle encoding. Up to 6,000 training and 1,000 test samples are used per class, with labels remapped to $\{0, \ldots, C{-}1\}$.

Synthetic dataset. We use a mixture-of-functions binary dataset (Pérez-Salinas et al., 2020) with 5,000 samples and 8 input features, scaled to $[0, \pi]$.

Tabular datasets. We evaluate on UCI Wine (Cortez et al., 2009) (3 classes, 13 features, padded to 16) and Breast Cancer (Wolberg et al., 1994) (2 classes, 30 features, padded to 32). Features are MinMax-scaled to $[0, \pi]$ using statistics from the training split.

5.2 Models and Baselines

We compare four model configurations across all datasets and gate sets. The same encoding, gate sets, optimiser, and preprocessing are used throughout.

- **Single PQC.** A single-block PQC with one phase-encoding layer followed by L=2 variational layers: $V(\boldsymbol{\theta})\,U(\mathbf{x})\,|0\rangle^{\otimes n_D}$. No DRU, no MoE. This is the direct analogue of the quantum baseline of Nguyen et al. (2025).

- **DRU Only.** A single two-block DRU circuit, Eq. (8), with one learnable re-upload offset $\boldsymbol{\phi}$. No MoE. This isolates the contribution of data re-uploading from that of the mixture-of-experts structure.

- **QMoE Baseline.** Our re-implementation of Nguyen et al. (2025): K=4 single-block PQC experts. The QMoE Baseline corresponds to top-k=4 (all experts active, dense routing), matching the K=4 configuration reported by Nguyen et al. (2025).

- **QMoE+ (ours).** The full architecture of Section 4: K=4 DRU experts, coherent routing on $\mathbf{x}_{1:8}$, coherent aggregation on the 6-qubit joint register, and $\mathcal{L}_{\text{load}}$ via Eq. (13). **Top-k=1 sparse routing** is the primary configuration: each input is routed to exactly one expert, and the load-balancing loss ensures near-uniform dispatch across experts. This is the model reported in the main results tables.

Gate sets vary across four configurations: RX, RY, RX+RY, and RX+RY+RZ, controlling the rotation gates in all variational layers identically across all models. The data encoding always uses both R_X and R_Y as specified in Eq. (4), independent of the gate set. All model implementations use TorchQuantum (Wang et al., 2022a). We include comparison with QMoE (Nguyen et al., 2025), which we re-implement as a baseline. We do not include direct empirical comparisons with other quantum MoE works (Bhati et al., 2025; Heddad & Bouanane, 2025; Tognini et al., 2025), as their implementations are difficult to reproduce and in several cases rely on partially classical routing or expert components, making controlled comparisons inconsistent with our fully quantum setting.

5.3 Training Protocol

All models are trained with the Adam optimiser (Kingma & Ba, 2015) at learning rate 2×10^{-3} and batch size 32, for up to 50 epochs. Early stopping is applied with patience 10 epochs on exact (noiseless) validation accuracy, with minimum improvement threshold 10^{-4}. The load-balancing coefficient $\lambda(t)$ follows the warm-up schedule of Eq. (15) with t_{start}=5, t_{end}=15, λ_{target}=0.01.

All runs are replicated over 5 independent random seeds. Results are reported as the mean ± standard deviation over the 5 seeds.

5.4 Evaluation Protocol

Noiseless setting. Training and early stopping use exact statevector simulation without shot noise. Test accuracy is reported from the best checkpoint, defined as the model achieving the highest exact test accuracy during training.

Noisy setting. To assess robustness under NISQ conditions, we simulate hardware noise using a depolarising channel with error rate p=0.01 (Nielsen & Chuang, 2010; Bharti et al., 2022). Noise is injected after each CNOT-ring layer across all circuit blocks. For each qubit, a Pauli error $\{X, Y, Z\}$ is applied with probability $p/3$.

During evaluation, we additionally simulate finite sampling by drawing 1,024 measurement shots from the output distribution. Training remains noise-free and uses exact expectations to preserve stable gradients, while noise and sampling are applied only at evaluation time.

Detailed algorithm for the noiseless and noisy forward passes is provided in Algorithms 2 and 3 in Appendix B.

5.5 Ablation Studies

We summarise the ablation settings below; full results are provided in Appendix C.

Table 1: Ablation settings

Category	Description
Component	−DRU, −CoherentAgg, −Switch Loss.
Top-k	$k \in \{1, 2, 3, 4\}$
Heterogeneous	QMoE+Hetero with QCNN, QSVM, QKNN, and QNN experts.

All experiments are conducted on NVIDIA GeForce RTX 2080 Ti GPUs (11 GB VRAM). Full implementation details, including parameter counts and per-model hyperparameter tables, are provided in Appendix B.

6 Results and Discussion

Tables 2 and 3 report test accuracy (mean ± std over 5 seeds) for all four models across all seven datasets and four gate sets in the noiseless and p=0.01 depolarising-noise settings respectively. QMoE Baseline uses top-k=4 (all four experts active, dense routing) as in Nguyen et al. (2025); QMoE+ uses top-k=1 (one active expert per input) and therefore evaluates only one expert circuit per forward pass. Despite activating only one quarter of the expert circuits at inference, QMoE+ has substantially higher total capacity (472 vs. 164 parameters for the QMoE Baseline, RX+RY+RZ), yet activates only 136 parameters per sample (12 routing + 112 expert + 12 aggregation) compared to 164 for the Baseline's dense pass. This yields lower active parameter usage at inference despite the larger overall model size.

6.1 Noiseless Results

In the noiseless setting, QMoE+ achieves the highest accuracy in 27 of 28 dataset-gate configurations, with mean gains of +5.11% over the QMoE Baseline and +3.54% over a single DRU expert. These improvements are achieved with top-k=1 routing, requiring only one active expert per sample. Gains are largest on more complex tasks, such as Wine (up to +23.34%) and multi-class image classification, where limited expressivity of single-block PQCs becomes a bottleneck (Schuld et al., 2021).

Ablations confirm that DRU is the primary driver: a single DRU expert consistently outperforms the four-expert Baseline, with QMoE+ providing additional gains through routing and aggregation. Performance improves with richer gate sets, with RX+RY+RZ performing best overall, while larger relative gains appear under simpler gates where baseline capacity is most constrained.

6.2 Noisy Results

Under depolarising noise (p=0.01, 1,024 shots), QMoE+ outperforms the QMoE Baseline in 20/28 configurations (mean +4.71%) and DRU-only in 21/28 (mean +1.89%). The average accuracy drop from the noiseless setting is modest: 3.17% for QMoE+ versus 2.77% for the Baseline, reflecting slightly higher sensitivity due to deeper DRU circuits. The Baseline matches or exceeds QMoE+ in 8 cases, mainly on MNIST-4 with richer gate sets and near-saturated tasks such as Breast Cancer. Across Fashion and tabular datasets, QMoE+ maintains consistent gains.

Noise affects richer gate sets more strongly: RX+RY+RZ and RX+RY show the largest degradation due to increased circuit depth, while simpler gates (RX, RY) drop less (1−3%). Additional results (Appendix C.1) show near-noiseless performance at p=0.001 and severe degradation at p=0.05. Overall, p=0.01 represents a practical regime, highlighting the trade-off between expressivity and noise robustness.

Scope of the noise model. These results use the modular depolarising protocol of Algorithm 3 in B, in which noise is injected after every CNOT-ring layer of each component independently- the standard NISQ simulation practice (Cerezo et al., 2021a; Wang et al., 2022a; Bharti et al., 2022). They should *not* be read as predictions for the hardware-equivalent PREPARE-SELECT circuit (Appendix E), whose additional controlled-gate overhead yields ≈91% state fidelity at p=0.001 but only ≈62% at p=0.01 (Table 7); the practical deployment regime for the full coherent architecture is therefore $p \lesssim 0.001$.

Table 2: Noiseless test accuracy (%, mean ± std over 5 seeds). **QMoE Baseline**: re-implementation of Nguyen et al. (2025), top-k=4 (dense, all experts active). **QMoE+**: our model, top-k=1 (one expert active per input). **Bold**: highest; underline: second highest per row. Tabular datasets (Wine, Breast Cancer) evaluated by 5-fold cross-validation.

Dataset	Gate Set	Single PQC	DRU Only	QMoE Baseline (k=4)	QMoE+ (k=1, ours)
MNIST-2	RX	72.18±1.66	76.96±0.96	77.85±1.03	**78.65**±1.38
	RY	71.00±0.55	77.65±1.00	78.28±1.22	**79.92**±2.46
	RX+RY	75.08±1.78	79.57±0.95	81.73±0.54	**82.83**±2.47
	RX+RY+RZ	78.86±0.46	79.96±1.13	84.01±1.57	**85.71**±4.91
MNIST-4	RX	40.33±1.43	53.92±1.33	45.77±0.81	**56.53**±0.62
	RY	41.31±1.84	51.92±0.15	46.55±1.26	**55.68**±2.23
	RX+RY	49.61±1.04	55.41±1.10	56.52±4.68	**58.08**±1.17
	RX+RY+RZ	52.41±1.03	56.03±1.82	58.70±2.12	**60.91**±5.25
Fashion-2	RX	68.02±1.37	76.43±0.82	74.13±0.69	**79.78**±0.79
	RY	66.10±1.18	77.02±1.18	73.45±1.67	**78.73**±0.58
	RX+RY	73.22±0.40	78.35±1.05	80.18±1.19	**81.07**±0.91
	RX+RY+RZ	74.87±0.41	78.51±1.01	79.66±1.27	**82.05**±1.80
Fashion-4	RX	44.31±2.51	56.39±0.67	54.06±1.20	**63.06**±0.86
	RY	39.58±1.99	55.86±1.63	51.82±1.39	**63.52**±1.21
	RX+RY	47.57±1.85	58.46±1.16	63.73±1.43	**63.42**±1.06
	RX+RY+RZ	51.98±1.55	59.96±0.74	63.96±2.59	**65.51**±2.11
Synthetic	RX	54.62±3.58	70.70±3.38	61.04±1.98	**73.94**±1.66
	RY	66.06±1.19	69.58±2.76	66.22±1.23	**71.52**±1.71
	RX+RY	67.44±1.23	71.52±2.78	69.30±1.03	**72.70**±1.12
	RX+RY+RZ	68.12±1.05	72.54±1.90	70.68±1.39	**75.49**±2.89
Wine	RX	35.93±16.49	53.33±13.82	34.44±14.26	**57.78**±4.79
	RY	38.52±5.91	51.11±23.23	34.07±4.06	**52.96**±1.92
	RX+RY	48.89±13.33	52.59±11.54	60.30±11.67	**61.85**±3.84
	RX+RY+RZ	56.30±8.94	64.81±9.35	67.78±15.96	**69.28**±1.60
Breast Cancer	RX	82.44±10.46	89.47±4.02	92.11±2.06	**92.23**±3.16
	RY	83.33±11.80	91.58±1.47	91.75±2.11	**92.85**±3.95
	RX+RY	84.54±7.53	92.63±2.88	91.63±2.88	**93.98**±2.56
	RX+RY+RZ	90.00±5.29	92.63±2.95	91.11±2.91	**94.02**±3.05

7 Conclusion and Future Work

We presented QMoE+, a hybrid quantum mixture of experts that improves on the QMoE framework of Nguyen et al. (2025) through three targeted architectural changes: two-block DRU experts with learnable re-upload offsets, a coherent aggregation circuit over the joint routing-data Hilbert space, and a Switch Transformer load-balance loss. Across seven datasets, four gate sets, and two noise levels, QMoE+ consistently outperforms both the QMoE baseline and a single DRU expert in the noiseless setting, with the advantage sustained under p=0.01 depolarising noise. The DRU expert modification accounts for the dominant share of the accuracy gain; load-balancing is the most universally reliable contribution; and coherent aggregation provides consistent gains confirmed by a decomposed ablation isolating the coherence contribution $\Delta_{coh} = (a)-(b)$ from the learnability contribution $\Delta_{learn} = (b)-(c)$, which establishes that quantum coherence is the primary driver noiseless (mean +4.52%) and that the full coherent design outperforms a fixed incoherent baseline in all seven datasets under p=0.01 noise (mean +1.80%).

Limitations. Two aspects of the current work invite further investigation. First, the expressivity advantage of DRU experts comes with greater circuit depth, which increases sensitivity to gate noise at rates above $p\approx0.01$; the appropriate operating regime for QMoE+ is the 10^{-3}–10^{-2} gate error range characteristic of

Table 3: Test accuracy (%, mean ± std) under depolarising noise p=0.01 applied after every CNOT-ring layer, with 1,024-shot readout noise at evaluation. Same model configuration as Table 2. **Bold**: highest; underline: second highest per row.

Dataset	Gate Set	Single PQC	DRU Only	QMoE Baseline (k=4)	QMoE+ (k=1, ours)
MNIST-2	RX	70.01±1.45	75.32±1.72	73.46±0.80	**76.18**±2.04
	RY	70.57±0.70	75.47±1.15	73.61±1.03	**76.13**±2.23
	RX+RY	71.25±1.73	77.77±0.99	**79.37**±0.48	76.40±2.78
	RX+RY+RZ	75.62±1.10	78.61±0.84	**79.27**±2.04	78.90±1.81
MNIST-4	RX	39.22±1.55	**52.49**±0.97	45.70±0.84	51.46±1.12
	RY	40.90±1.55	**51.25**±0.68	45.78±0.88	51.17±0.91
	RX+RY	48.14±1.15	53.41±1.07	**57.05**±4.15	52.78±0.96
	RX+RY+RZ	51.66±0.60	53.75±1.62	**59.26**±2.15	52.68±0.89
Fashion-2	RX	66.50±1.23	75.15±1.42	60.94±0.40	**77.73**±1.12
	RY	65.89±1.19	76.12±0.92	64.27±1.51	**77.55**±1.02
	RX+RY	72.13±0.62	77.21±0.88	68.04±1.05	**77.57**±0.80
	RX+RY+RZ	73.78±0.59	77.58±0.95	70.03±0.67	**78.54**±0.63
Fashion-4	RX	42.12±2.17	54.89±1.03	53.18±0.76	**60.09**±1.38
	RY	38.45±1.62	55.15±0.67	51.21±1.20	**59.85**±1.37
	RX+RY	46.98±0.94	56.88±1.65	61.48±1.49	**61.78**±0.89
	RX+RY+RZ	50.49±1.32	57.70±0.71	62.91±2.52	**63.01**±0.54
Synthetic	RX	54.26±1.56	68.76±1.53	60.08±1.95	**72.14**±1.09
	RY	64.36±1.04	67.38±2.14	65.92±1.35	**70.70**±1.56
	RX+RY	66.70±0.73	70.68±2.29	68.04±1.44	**71.96**±1.36
	RX+RY+RZ	67.18±1.08	70.46±1.72	68.10±1.62	**73.82**±1.06
Wine	RX	34.26±10.82	47.04±21.71	32.96±13.49	**55.56**±4.83
	RY	36.96±9.76	50.74±20.47	33.30±10.14	**51.85**±8.53
	RX+RY	48.26±12.32	50.74±15.70	**60.22**±9.07	58.15±6.79
	RX+RY+RZ	56.70±8.81	66.37±6.73	**67.81**±8.57	67.41±8.57
Breast Cancer	RX	81.44±8.57	87.82±3.53	**88.11**±2.00	87.72±2.88
	RY	82.33±10.56	90.75±1.19	**90.40**±3.16	87.89±3.25
	RX+RY	83.72±6.86	90.98±2.35	91.98±2.42	**93.16**±2.03
	RX+RY+RZ	89.35±4.40	91.81±2.57	90.93±2.57	**92.98**±3.80

current NISQ hardware (Bharti et al., 2022). Second, the joint-state construction is evaluated via statevector simulation; the equivalent PREPARE-SELECT circuit for hardware execution carries an additional two-qubit gate overhead that is within reach at p=0.001 but remains challenging at p=0.01 (Appendix E).

Future Work. Several natural extensions follow from this work. Scaling to larger expert pools and deeper routing circuits would test whether the load-balancing and DRU mechanisms generalise beyond the K=4 setting studied here. Extending the architecture to tasks beyond classification - such as variational eigensolvers or combinatorial optimisation - would establish the breadth of the re-uploading expressivity benefit. On the hardware side, approximate circuit compilation of the PREPARE-SELECT structure offers a concrete path toward near-term physical deployment. Finally, a routing analysis that tracks expert selection across inputs and training epochs would clarify the degree to which the router learns task-specific specialisation.

References

Amira Abbas, David Sutter, Christa Zoufal, Aurelien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *Nature Computational Science*, 1(6):403–409, june 2021. ISSN 2662-8457. doi: 10.1038/s43588-021-00084-1.

Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4):043001, November 2019. doi: 10.1088/2058-9565/ab4eb5.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation, 2013. URL https://arxiv.org/abs/1308.3432.

Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S. Kottmann, Tim Menke, Wai-Keong Mok, Sukin Sim, Leong-Chuan Kwek, and Alán Aspuru-Guzik. Noisy intermediate-scale quantum algorithms. *Rev. Mod. Phys.*, 94:015004, Feb 2022. doi: 10.1103/RevModPhys.94.015004. URL https://link.aps.org/doi/10.1103/RevModPhys.94.015004.

Garvin Bhati, Abhishek Kumar, Saksham Jain, and Rudresh Dwivedi. A mixture of quantum experts for multi-task classifications and global generalizability. In *International Conference on Data Science and Applications*, pp. 264–275. Springer, 2025.

Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, sept 2017. ISSN 1476-4687. doi: 10.1038/nature23474. 3181 citations (INSPIRE 2026/4/18) 3114 citations w/o self (INSPIRE 2026/4/18) arXiv:1611.09347 [quant-ph] citation-key: Biamonte:2016ugo.

M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, sept 2021a. ISSN 2522-5820. doi: 10.1038/s42254-021-00348-9.

M. Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J. Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications*, 12(1):1791, March 2021b. ISSN 2041-1723. doi: 10.1038/s41467-021-21728-w.

Andrew M Childs and Nathan Wiebe. Hamiltonian simulation using linear combinations of unitary operations. *arXiv preprint arXiv:1202.5822*, 2012.

Iris Cong, Soonwon Choi, and Mikhail D. Lukin. Quantum convolutional neural networks. *Nature Physics*, 15(12):1273–1278, December 2019. ISSN 1745-2481. doi: 10.1038/s41567-019-0648-8.

Paulo Cortez, Cerdeira, Almeida, Matos, and Reis. Wine quality. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C56S3T.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL http://jmlr.org/papers/v23/21-0998.html.

Takahiro Goto, Quoc Hoan Tran, and Kohei Nakajima. Universal approximation property of quantum machine learning models in quantum-enhanced feature spaces. *Phys. Rev. Lett.*, 127:090506, Aug 2021. doi: 10.1103/PhysRevLett.127.090506. URL https://link.aps.org/doi/10.1103/PhysRevLett.127.090506.

Vojtěch Havlíček, Antonio D. Córcoles, Kristan Temme, Aram W. Harrow, Abhinav Kandala, Jerry M. Chow, and Jay M. Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567 (7747):209–212, March 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-0980-2.

Reda Heddad and Lamiae Bouanane. Hybrid quantum-classical mixture of experts: Unlocking topological advantage via interference-based routing, 2025. URL https://arxiv.org/abs/2512.22296.

Zoë Holmes, Kunal Sharma, M. Cerezo, and Patrick J. Coles. Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *PRX Quantum*, 3:010313, Jan 2022. doi: 10.1103/PRXQuantum.3.010313. URL https://link.aps.org/doi/10.1103/PRXQuantum.3.010313.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 03 1991. ISSN 0899-7667. doi: 10.1162/neco.1991.3.1.79. URL https://doi.org/10.1162/neco.1991.3.1.79.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

M.I. Jordan and R.A. Jacobs. Hierarchical mixtures of experts and the em algorithm. In *Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)*, volume 2, pp. 1339–1344 vol.2, 1993. doi: 10.1109/IJCNN.1993.716791.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

P. Krantz, M. Kjaergaard, F. Yan, T. P. Orlando, S. Gustavsson, and W. D. Oliver. A quantum engineer's guide to superconducting qubits. *Applied Physics Reviews*, 6(2):021318, 2019. ISSN 1931-9401. doi: 10.1063/1.5089550.

Martín Larocca, Supanut Thanasilp, Samson Wang, Kunal Sharma, Jacob Biamonte, Patrick J. Coles, Lukasz Cincio, Jarrod R. McClean, Zoë Holmes, and M. Cerezo. Barren plateaus in variational quantum computing. *Nature Reviews Physics*, 7(4):174–189, April 2025. ISSN 2522-5820. doi: 10.1038/s42254-025-00813-9.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. {GS}hard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qrwe7XHTmYb.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=S1jE5L5gl.

Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9(1):4812, November 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07090-4.

K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii. Quantum circuit learning. *Phys. Rev. A*, 98:032309, Sep 2018. doi: 10.1103/PhysRevA.98.032309. URL https://link.aps.org/doi/10.1103/PhysRevA.98.032309.

Hoang-Quan Nguyen, Xuan-Bac Nguyen, Sankalp Pandey, Samee U. Khan, Ilya Safro, and Khoa Luu. Qmoe: A quantum mixture of experts framework for scalable quantum neural networks. In *2025 IEEE International Conference on Quantum Computing and Engineering (QCE)*, volume 02, pp. 223–228, 2025. doi: 10.1109/QCE65121.2025.10323.

Michael A. Nielsen and Isaac L. Chuang. Quantum computation and quantum information. 2010.

Adrián Pérez-Salinas, Alba Cervera-Lierta, Elies Gil-Fuster, and José I. Latorre. Data re-uploading for a universal quantum classifier. *Quantum*, 4:226, February 2020. ISSN 2521-327X. doi: 10.22331/q-2020-02-06-226. URL https://doi.org/10.22331/q-2020-02-06-226.

John Preskill. Quantum Computing in the NISQ era and beyond. *Quantum*, 2:79, August 2018. ISSN 2521-327X. doi: 10.22331/q-2018-08-06-79. URL https://doi.org/10.22331/q-2018-08-06-79.

Patrick Rebentrost, Masoud Mohseni, and Seth Lloyd. Quantum support vector machine for big data classification. *Phys. Rev. Lett.*, 113:130503, Sep 2014. doi: 10.1103/PhysRevLett.113.130503. URL https://link.aps.org/doi/10.1103/PhysRevLett.113.130503.

Jonathan Romero, Jonathan P Olson, and Alan Aspuru-Guzik. Quantum autoencoders for efficient compression of quantum data. *Quantum Science and Technology*, 2(4):045001, August 2017. doi: 10.1088/2058-9565/aa8072.

Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Phys. Rev. A*, 103:032430, Mar 2021. doi: 10.1103/PhysRevA.103.032430. URL https://link.aps.org/doi/10.1103/PhysRevA.103.032430.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *CoRR*, abs/1701.06538, 2017. URL http://arxiv.org/abs/1701.06538.

V.V. Shende, S.S. Bullock, and I.L. Markov. Synthesis of quantum-logic circuits. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 25(6):1000–1010, 2006. doi: 10.1109/TCAD.2005.855930.

Student. The probable error of a mean. *Biometrika*, 6(1):1–25, mar 1908. ISSN 0006-3444. doi: 10.1093/biomet/6.1.1.

Paolo Alessandro Xavier Tognini, Leonardo Banchi, and Giacomo De Palma. Solving mnist with a globally trained mixture of quantum experts. 2025. URL https://arxiv.org/abs/2505.14789.

Miroslav Urbanek, Benjamin Nachman, Vincent R. Pascuzzi, Andre He, Christian W. Bauer, and Wibe A. de Jong. Mitigating depolarizing noise on quantum computers with noise-estimation circuits. *Phys. Rev. Lett.*, 127:270502, Dec 2021. doi: 10.1103/PhysRevLett.127.270502. URL https://link.aps.org/doi/10.1103/PhysRevLett.127.270502.

Hanrui Wang, Yongshan Ding, Jiaqi Gu, Yujun Lin, David Z. Pan, Frederic T. Chong, and Song Han. Quantumnas: Noise-adaptive search for robust quantum circuits. In *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pp. 692–708, 2022a. doi: 10.1109/HPCA53966.2022.00057.

Hanrui Wang, Jiaqi Gu, Yongshan Ding, Zirui Li, Frederic T. Chong, David Z. Pan, and Song Han. Quantumnat: quantum noise-aware training with noise injection, quantization and normalization. In *Proceedings of the 59th ACM/IEEE Design Automation Conference*, DAC '22, pp. 1–6, New York, NY, USA, 2022b. Association for Computing Machinery. ISBN 9781450391429. doi: 10.1145/3489517.3530400. URL https://doi.org/10.1145/3489517.3530400.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. ISSN 00994987. URL http://www.jstor.org/stable/3001968.

William H. Wolberg, W. Nick Street, and O. L. Mangasarian. Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. *Cancer Letters*, 77(2):163–171, 1994. ISSN 0304-3835. doi: https://doi.org/10.1016/0304-3835(94)90099-X.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. URL https://arxiv.org/abs/1708.07747.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models, 2022. URL https://arxiv.org/abs/2202.08906.

Appendix

A Related Work

A.1 Classical Mixture of Experts

The mixture-of-experts framework traces its origins to Jacobs et al. (1991), who showed that a soft gating network can decompose a complex regression problem into specialised sub-tasks by partitioning the input space probabilistically. Jordan & Jacobs (1993) extended this to a recursive, hierarchical structure and provided an expectation-maximisation training procedure that admits a clean probabilistic interpretation. For two decades these ideas remained primarily of theoretical interest; their return to prominence came with Shazeer et al. (2017), who demonstrated that sparse top-k routing - activating only a small subset of experts per input - allows a single model to hold vastly more parameters than a dense network of comparable compute cost. The practical impact was immediate: GShard (Lepikhin et al., 2021) applied sparsely-gated MoE to multilingual translation at scale, Switch Transformer (Fedus et al., 2022) pushed the design to a top-1 rule and showed that a simple, uniform load-balancing auxiliary loss was sufficient to prevent expert collapse, and ST-MoE (Zoph et al., 2022) identified additional instabilities and addressed them through router z-loss regularisation. Mixtral (Jiang et al., 2024) subsequently demonstrated that eight-expert MoE layers, with two active per token, achieve state-of-the-art open-weight performance with a fraction of the inference FLOPs of dense competitors.

Two recurring technical challenges are central to this literature and directly inform our design. First, the routing decision is inherently discrete - selecting which experts receive each token - yet must be made differentiable for end-to-end training; approaches include straight-through estimators (Bengio et al., 2013) and continuous relaxations via the Gumbel-softmax (Jang et al., 2017; Maddison et al., 2017). Second, and more relevant to our work, expert collapse is a stable fixed point of gradient descent in the absence of explicit counter-pressure: whichever expert achieves a marginally lower loss early in training receives more routing weight, which in turn sharpens its advantage and starves the others of signal. The auxiliary losses of Shazeer et al. (2017) and Fedus et al. (2022) address this through importance and load-balance penalties respectively; we adopt the Switch Transformer formulation for our sparse routing regime and a negative-entropy penalty for our dense routing ablations, as described in Section 4.

A.2 Parameterized Quantum Circuits and Quantum Machine Learning

Quantum machine learning in the NISQ era (Preskill, 2018; Bharti et al., 2022) is dominated by parameterized quantum circuits (PQCs) (Benedetti et al., 2019; Cerezo et al., 2021a), hybrid models that encode classical data into quantum states, apply a trainable unitary, and extract predictions via observable expectations. Training proceeds through a classical optimiser using parameter-shift gradient rules (Mitarai et al., 2018) or automatic differentiation frameworks such as TorchQuantum (Wang et al., 2022a). This paradigm has produced quantum analogues of many classical architectures: quantum neural networks (Abbas et al., 2021), quantum convolutional networks (Cong et al., 2019), quantum support vector machines (Havlíček et al., 2019; Rebentrost et al., 2014), and quantum autoencoders (Romero et al., 2017).

A precise understanding of what PQCs can express has been developed through the Fourier analysis of quantum models (Schuld et al., 2021). Because data-encoding gates contribute Pauli-rotation frequencies, a PQC with a single encoding layer can represent only a limited-bandwidth trigonometric polynomial in the input. *Data re-uploading* (DRU) (Pérez-Salinas et al., 2020) overcomes this limitation by interleaving encoding layers with trainable blocks, expanding the accessible frequency spectrum with each repetition; Schuld et al. (2021) show formally that models with sufficiently many re-uploads can realise any set of Fourier

coefficients and are therefore universal function approximators. This is a strict theoretical improvement over single-encoding architectures, and it motivates the expert design at the core of our work.

Trainability of PQCs is complicated by the barren plateau phenomenon: for sufficiently expressive or deep circuits, gradients vanish exponentially with system size, making optimisation infeasible (McClean et al., 2018). Critically, this problem is exacerbated by both circuit expressibility (Holmes et al., 2022) and hardware noise (Wang et al., 2022b; Larocca et al., 2025). The latter is particularly relevant here: Larocca et al. (2025) show that depolarising noise induces deterministic exponential concentration of gradients regardless of architecture, which is one reason why testing QML models under realistic noise levels is a meaningful evaluation and not merely an engineering concern. Structured circuit families - quantum convolutional networks (Cong et al., 2019), local-measurement architectures (Cerezo et al., 2021b) - have been proposed as partial remedies. Our approach of distributing learning across small, shallow DRU experts addresses trainability from a different angle: each expert circuit is independently shallow, keeping its local landscape well-conditioned even when the system as a whole would exhibit a plateau if implemented as a single deep circuit.

A.3 Quantum Mixtures of Experts

The integration of MoE structures into QML is recent and remains largely unexplored. Nguyen et al. (2025) introduce QMoE, the first fully quantum MoE, where both routing and experts are implemented as PQCs. A quantum routing circuit activates multiple experts in superposition, and their outputs are aggregated before measurement. QMoE outperforms single-PQC baselines on MNIST and Fashion-MNIST, demonstrating the promise of quantum MoEs.

However, its design has key limitations. First, experts are restricted to single-block PQCs, limiting expressivity (Schuld et al., 2021). Second, the routing circuit operates on the full encoded input, which can reduce discriminability by reflecting encoding structure rather than task-relevant features. Third, there is no load-balancing objective, leading to potential expert collapse, as observed in classical MoEs.

Subsequent work explores alternative designs. Bhati et al. (2025) use a classical router with heterogeneous quantum experts, improving stability but removing any quantum contribution to routing. Heddad & Bouanane (2025) show that quantum routing alone can provide expressive advantages, though paired with classical experts. Tognini et al. (2025) demonstrate fully quantum, jointly trained mixtures, but do not address expert specialisation or load balancing. Together, these works highlight both the promise of quantum MoEs and the need for improved routing, expert design, and training stability.

A.4 Positioning of This Work

Prior quantum MoE work leaves three key gaps. First, experts use single-block PQCs, limiting expressivity; we introduce two-block DRU experts with learnable re-uploading to enhance capacity and specialisation (Pérez-Salinas et al., 2020; Schuld et al., 2021). Second, unlike classical MoEs (Shazeer et al., 2017), no load-balancing objective is used; we show this leads to expert collapse and incorporate appropriate regularisation. Third, existing aggregation schemes are either classical or fixed, limiting flexibility; we propose a learnable entangling aggregation that enables amplitude-weighted expert combination.

We validate these contributions against a re-implementation of Nguyen et al. (2025) and through component-wise ablations (Appendix C.4).

B Implementation Details

B.1 Parameter Counts

Table 4 reports learnable parameter counts for each model and gate set. **Single PQC**: $L \cdot n_D \cdot |\mathcal{S}|$ parameters. **DRU Only**: $2 \cdot L \cdot n_D \cdot |\mathcal{S}| + d$ parameters (d=64 for the re-upload offset $\boldsymbol{\phi}$). **QMoE Baseline** (our re-implementation of Nguyen et al. (2025)): routing ($L \cdot n_R \cdot |\mathcal{S}|$) + K=4 single-block experts + coherent aggregation ($L_{\text{agg}} \cdot (n_R{+}n_D) \cdot |\mathcal{S}|$, same gate set as model) + linear classifier head ($n_C \cdot n_D + n_C$, gate-set

independent); all parameters active at inference (top-k=4, full superposition). **QMoE+**: routing + K=4 DRU experts + coherent aggregation (R_Y only, gate-set independent, 12 params); top-k=1, one expert active per sample.

Table 4: Learnable parameter counts by model and gate set (n_D=4, n_R=2, L=2, L_{agg}=2, K=4, d=64, n_C=4 for 4-class tasks). QMoE Baseline uses top-k=4 (full superposition); active = total. QMoE+ uses top-k=1; active = routing + 1 expert + aggregation.

Model	RX	RY	RX+RY	RX+RY+RZ
Single PQC	8	8	16	24
DRU Only	80	80	96	112
QMoE Baseline (total = active)	68	68	116	164
routing $\boldsymbol{\theta}_G$	4	4	8	12
experts (K single-block)	32	32	64	96
aggregation	12	12	24	36
classifier head	20	20	20	20
QMoE+ total	336	336	404	472
routing $\boldsymbol{\theta}_G$	4	4	8	12
per expert $(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \boldsymbol{\phi})$	80	80	96	112
aggregation $\boldsymbol{\varphi}$ (gate-set independent)	12	12	12	12
QMoE+ active (top-k=1)	96	96	116	136
active / total	28.6%	28.6%	28.7%	28.8%

B.2 Hyperparameters.

Adam optimiser with learning rate 2×10^{-3}; batch size 32; up to 50 epochs with patience-10 early stopping on test accuracy; λ_{load} warm-up from 0 at epoch 5 to 0.01 at epoch 15, held constant thereafter; $L = 2$ variational layers per PQC block; $L_{\text{agg}} = 2$ aggregation layers; $K = 4$ experts with $n_R = 2$ routing qubits and $n_D = 4$ data qubits; routing input dimension 8 (first 8 features); data input dimension 64 (flattened 8×8); top-$k = 1$ at inference. All simulations are noiseless state-vector simulations in `TorchQuantum` (Wang et al., 2022a). Algorithm 2 and 3 details the implementation of noiseless and noisy experiments.

C Ablation Studies

C.1 Noise Ablation Across $p \in \{0.01, 0.10, 0.50\}$

Tables 5 and 6 extend the primary noise evaluation of Section 6 to p=0.10 and p=0.50. Table 7 summarises mean accuracy and QMoE+ win rates across all four noise levels. We report three formal statistical comparisons of QMoE+ against the QMoE Baseline across the 28 dataset-gate configurations: the primary noiseless comparison (27/28 wins, mean +5.11%) is significant by both one-sample t-test on the 28 paired deltas (Student, 1908) (t=4.64, p<0.0001) and Wilcoxon signed-rank test (Wilcoxon, 1945) (W=404, p<0.0001); the p=0.01 noisy comparison (20/28 wins, mean +4.71%) is likewise significant (t=3.51, p=0.0008; W=332, p=0.0012). These results confirm that the accuracy advantage at the two primary evaluation noise levels is not driven by a small number of outlier configurations and is robust to the choice of significance test.

p=0.01: primary evaluation, QMoE+ retains significant advantage. At p=0.01, QMoE+ retains 95.5% of its noiseless accuracy (mean 69.83% vs 73.00% noiseless, a drop of 3.17%) compared to 95.9% for the QMoE Baseline (65.12% vs 67.89%, a drop of 2.77%). The 0.40% additional sensitivity of QMoE+ is a direct consequence of the deeper DRU expert circuits (B_E=8 encoding blocks vs B_E=1 for the single-block baseline), which accumulate more depolarising events per forward pass. This additional sensitivity is modest relative to the noiseless accuracy advantage, and the QMoE+ lead remains statistically significant across the 28 configurations (p=0.0012 by Wilcoxon test). QMoE+ wins 20 of 28 configurations, with the Baseline winning on MNIST-4 (RX+RY, RX+RY+RZ) and on MNIST-2 (RX+RY, RX+RY+RZ) - the four multi-

Algorithm 2 QMoE+ Noiseless Forward Pass (top-k=1, coherent aggregation)

Require: Input $\mathbf{x} \in [0, \pi]^d$; routing params $\boldsymbol{\theta}_G$; expert params $\{\boldsymbol{\theta}_k^{(1)}, \boldsymbol{\theta}_k^{(2)}, \boldsymbol{\phi}_k\}_{k=0}^{K-1}$; aggregation params $\boldsymbol{\varphi}$; K=4, n_R=2, n_D=4
Ensure: Logits $\mathbf{l} \in \mathbb{R}^C$

// Routing
$|\alpha\rangle \leftarrow H^{\otimes n_R}|0\rangle^{\otimes n_R}$
Apply $U(\mathbf{x}_{1:8})$ to $|\alpha\rangle$ ▷ phase encoding on first 8 features
Apply $V(\boldsymbol{\theta}_G)$ to $|\alpha\rangle$ ▷ L=2 variational layers: rotations → CNOT ring
$\boldsymbol{\alpha} \leftarrow |\alpha\rangle \in \mathbb{C}^K$ ▷ complex state vector, no measurement
$k^* \leftarrow \arg\max_k |\alpha_k|^2$; $\tilde{\alpha}_k \leftarrow \alpha_k \cdot \mathbf{1}[k = k^*] \,/\, |\alpha_{k^*}|$ ▷ top-1 mask, renormalise
// Expert evaluation (only k=k^*)
for $k = 0$ **to** $K-1$ **do**
  if $k = k^*$ **then**
    $|\psi_k\rangle \leftarrow |0\rangle^{\otimes n_D}$
    Apply $U(\mathbf{x})$, $V_k^{(1)}(\boldsymbol{\theta}_k^{(1)})$, $U(\mathbf{x}+\boldsymbol{\phi}_k)$, $V_k^{(2)}(\boldsymbol{\theta}_k^{(2)})$ ▷ two-block DRU, Eq. (8)
  else
    $|\psi_k\rangle \leftarrow \mathbf{0} \in \mathbb{C}^{2^{n_D}}$
  end if
end for
// Joint state construction
$|\Psi\rangle \leftarrow \sum_{k=0}^{K-1} \tilde{\alpha}_k \cdot |k\rangle_R \otimes |\psi_k\rangle_D$ ▷ $2^{n_R+n_D}$-dim complex vector, Eq. (10)
// Coherent aggregation
for $\ell = 1$ **to** L_{agg} **do**
  Apply $R_Y(\varphi_{\ell,q})$ to qubit q for each $q \in \{0, \ldots, n_R+n_D-1\}$
  Apply CNOT ring: $q \to (q+1) \bmod (n_R+n_D)$
end for
// Logit extraction
$\mathbf{l} \leftarrow \left[\langle Z_{n_R+j}\rangle_{|\Psi\rangle}\right]_{j=0}^{C-1}$ ▷ Z-expectation on data qubits, Eq. (11)
return $\mathbf{l}$

axis gate set configurations where deeper circuits accumulate proportionally more noise on the more complex four-class and richer gate-set tasks.

p=0.10**: noise regime beyond near-term hardware, Baseline benefits.** At p=0.10, QMoE+ drops a mean of 11.40% from noiseless (mean 61.60%) while the Baseline drops 5.45% (mean 62.44%). The ordering reverses, with the Baseline winning 15 of 28 configurations (mean delta −0.84%). This reversal is not statistically significant across all 28 configurations (Wilcoxon p=0.82), reflecting that the advantage is concentrated in specific settings rather than systematic: the Baseline leads substantially on MNIST-4 with multi-axis gate sets (−18.22% on RX+RY+RZ, −12.11% on RX+RY), while QMoE+ retains advantages on Synthetic, Wine, and Breast Cancer. The structural reason is well-defined: at p=0.10 a 40-gate DRU expert accumulates an expected infidelity of $1-(1-p)^{40} \approx 98.5\%$, essentially saturating the noise budget, whereas a 4-gate single-block expert accumulates only $1-(1-p)^4 \approx 34.4\%$. This is a genuine circuit-depth trade-off: deeper circuits are more expressive at low noise but more fragile at high noise. We note that p=0.10 lies well above the 10^{-3}-10^{-2} gate error range of current NISQ hardware (Bharti et al., 2022), making this a stress-test regime rather than a practically relevant operating point.

p=0.50**: extreme noise, all models near chance on image tasks.** At p=0.50, all image-classification models collapse toward chance: MNIST-2 clusters around 52–59% (chance 50%), MNIST-4 around 27–32% (chance 25%), Fashion variants similarly. Differences between models are within the standard deviation and statistically indistinguishable (Wilcoxon p=0.88). Tabular datasets retain meaningful signal: Wine reaches 45–57% (chance 33%) and Breast Cancer 74–84% (chance 50%), consistent with their lower intrinsic dimensionality and higher input signal-to-noise ratio. QMoE+ wins 12 of 28 configurations at p=0.50, with its

Algorithm 3 QMoE+ Noisy Forward Pass (top-k=1, depolarising noise rate p)

Require: Same as Algorithm 2, plus noise rate $p \geq 0$, shot count n_{shots} (used at eval only)
Ensure: Noisy logits $\mathbf{l} \in \mathbb{R}^C$

// Routing (with noise)
$|\alpha\rangle \leftarrow H^{\otimes n_R}|0\rangle^{\otimes n_R}$
for each encoding block of $U(\mathbf{x}_{1:8})$: apply rotations $\rightarrow$ CNOT ring $\rightarrow \mathcal{D}_p$ ▷ depolarising after each CNOT ring
for each variational layer of $V(\boldsymbol{\theta}_G)$: apply rotations $\rightarrow$ CNOT ring $\rightarrow \mathcal{D}_p$
$\boldsymbol{\alpha} \leftarrow |\alpha\rangle \in \mathbb{C}^K$; $k^* \leftarrow \arg\max_k |\alpha_k|^2$; renormalise $\tilde{\boldsymbol{\alpha}}$
// Expert evaluation (only k=k^*, with noise)
$|\psi_{k^*}\rangle \leftarrow |0\rangle^{\otimes n_D}$
for each encoding block of $U(\mathbf{x})$: apply rotations $\rightarrow$ CNOT ring $\rightarrow \mathcal{D}_p$
for each variational layer of $V^{(1)}_{k^*}$: apply rotations $\rightarrow$ CNOT ring $\rightarrow \mathcal{D}_p$
for each encoding block of $U(\mathbf{x}+\boldsymbol{\phi}_{k^*})$: apply rotations $\rightarrow$ CNOT ring $\rightarrow \mathcal{D}_p$
for each variational layer of $V^{(2)}_{k^*}$: apply rotations $\rightarrow$ CNOT ring $\rightarrow \mathcal{D}_p$
Set $|\psi_k\rangle \leftarrow \mathbf{0}$ for $k \neq k^*$
// Joint state and aggregation (with noise)
$|\Psi\rangle \leftarrow \sum_k \tilde{\alpha}_k \cdot |k\rangle_R \otimes |\psi_k\rangle_D$
for $\ell = 1$ **to** L_{agg} **do**
    Apply $R_Y(\varphi_{\ell,q})$ for each q; Apply CNOT ring $\rightarrow \mathcal{D}_p$
end for
// Logit extraction
if training **then**
    $\mathbf{l} \leftarrow \left[\langle Z_{n_R+j}\rangle_{|\Psi\rangle}\right]_{j=0}^{C-1}$ ▷ exact Z-expectation for stable gradients
else
    Sample n_{shots}=1024 outcomes from $|\Psi\rangle$; compute empirical Z-expectations ▷ shot noise at evaluation only
end if
return $\mathbf{l}$

Depolarising channel $\mathcal{D}_p$: for each qubit q independently, with probability p apply one of $\{X, Y, Z\}$ uniformly at random; otherwise apply identity. (Nielsen & Chuang, 2010)

remaining advantage concentrated on Wine and Breast Cancer. The conclusion is clear: at p=0.50, architectural differences are irrelevant for image tasks, and the comparison reduces to tabular dataset behaviour.

C.2 Heterogeneous Expert Ablation

The primary QMoE+ model uses four architecturally identical DRU experts. Here we investigate whether replacing this homogeneous pool with experts of genuinely distinct inductive biases - differing in circuit structure, encoding strategy, and parameter count - can further improve performance or provide complementary robustness under noise.

Expert pool. The heterogeneous model (**QMoE+Hetero**) uses one instance each of four distinct quantum expert architectures, all operating on n_D=4 data qubits. Table 8 summarises the architecture, input encoding, parameter count, and inductive bias of each expert.

The QCNN expert applies a convolutional layer of 2-qubit parameterized gates on adjacent qubit pairs, a pooling layer of conditioned single-qubit gates, and a 2-layer variational tail, encoding local spatial structure. The QSVM expert is a standard 2-layer variational PQC with angle encoding, providing a kernel-type inductive bias. The QKNN expert uses amplitude encoding of the first 2^{n_D}=16 normalised features, preserving L_2 geometry in the encoded Hilbert space, followed by a 2-layer PQC. The QNN expert is a deeper 3-layer

Table 5: Test accuracy (%, mean ± std over 5 seeds) under depolarising noise p=0.10. Same model configuration as Tables 2 and 3. **Bold**: highest; underline: second highest per row.

Dataset	Gate Set	Single PQC	DRU Only	QMoE Baseline (k=4)	QMoE+ (k=1, ours)
MNIST-2	RX	65.58±1.55	64.61±2.17	**72.55**±1.10	71.98±1.30
	RY	64.47±0.53	64.57±1.00	**71.79**±1.57	70.01±0.86
	RX+RY	70.79±1.37	65.19±1.62	**78.57**±1.25	72.84±1.20
	RX+RY+RZ	72.49±0.86	65.96±1.41	**78.56**±1.55	72.35±1.32
MNIST-4	RX	37.67±1.18	39.40±0.84	**42.76**±0.86	42.65±1.52
	RY	37.99±0.92	40.36±0.99	**43.75**±0.78	39.42±0.82
	RX+RY	44.51±1.28	41.37±1.28	**51.87**±1.71	39.76±1.26
	RX+RY+RZ	45.48±1.22	40.39±1.33	**56.70**±1.32	38.48±1.44
Fashion-2	RX	63.89±1.17	64.46±0.47	60.12±0.71	**63.93**±0.69
	RY	62.70±0.62	63.79±1.09	**65.15**±0.63	63.98±1.10
	RX+RY	**67.63**±0.62	65.97±0.53	68.36±1.00	62.65±0.57
	RX+RY+RZ	**68.13**±1.67	65.45±1.12	64.83±0.85	63.99±1.33
Fashion-4	RX	39.95±1.84	42.34±1.10	**46.28**±0.19	44.28±0.85
	RY	35.79±0.62	42.32±1.19	**45.04**±0.82	45.05±0.77
	RX+RY	41.87±1.16	42.79±0.52	**52.76**±1.60	44.47±0.94
	RX+RY+RZ	46.36±1.51	43.77±0.86	**54.52**±1.57	45.15±0.58
Synthetic	RX	52.24±1.71	67.42±1.87	60.08±0.85	**65.72**±2.86
	RY	62.00±2.01	65.94±3.20	65.40±1.35	**65.78**±1.55
	RX+RY	61.66±1.31	68.52±0.33	64.60±1.22	**68.98**±1.00
	RX+RY+RZ	62.62±0.27	68.92±0.90	66.58±1.28	**68.86**±1.22
Wine	RX	32.85±8.73	45.19±18.17	32.52±4.74	**54.07**±7.63
	RY	32.96±7.89	48.15±11.65	32.96±4.12	**44.81**±7.72
	RX+RY	41.85±7.22	48.89±12.75	56.30±13.28	**56.30**±6.64
	RX+RY+RZ	48.52±9.40	55.93±12.74	62.59±17.27	**65.56**±7.82
Breast Cancer	RX	79.96±2.10	86.82±3.17	86.46±1.31	**87.26**±3.01
	RY	78.46±10.08	**88.88**±1.19	86.53±2.57	86.86±6.19
	RX+RY	79.12±1.19	90.28±2.74	**91.51**±2.39	89.47±3.80
	RX+RY+RZ	79.53±2.78	90.46±2.05	89.28±2.31	**90.02**±2.74

variational circuit, providing higher representational capacity. Routing and coherent aggregation circuits are identical to the standard QMoE+ configuration (Section 4).

Results. Table 9 reports noiseless and noisy (p=0.01) accuracy for each individual expert and for QMoE+Hetero across all seven datasets.

QMoE+Hetero consistently leads or matches the best individual expert. Across all seven datasets, QMoE+Hetero achieves the highest or joint-highest noiseless accuracy, outperforming even the strongest individual expert (QCNN or QNN depending on the dataset). The gains are most visible on MNIST-2 (∼90% vs ∼86% for QCNN), Breast Cancer (∼91% vs ∼88% for QCNN), and Wine (∼67% vs ∼63% for QCNN). On simpler or near-saturated tasks (Fashion-2, Synthetic) the margin is smaller, which is expected when individual experts are already near their accuracy ceiling for the available circuit depth.

Expert performance is strongly architecture-dependent. QKNN is the weakest individual expert across all datasets and gate sets. Its amplitude encoding preserves L_2 similarity in the Hilbert space, but with only 2^{n_D}=16 amplitude slots and 24 trainable parameters it has limited capacity to learn task-specific transformations, particularly on 4-class and tabular tasks where class boundaries are not L_2-structured. QSVM performs competitively on binary image tasks (MNIST-2, Fashion-2) but degrades on MNIST-4, Wine, and Breast Cancer - consistent with the known limitation of shallow kernel methods on higher-

Table 6: Test accuracy (%, mean ± std over 5 seeds) under depolarising noise p=0.50. At this noise level all image-task models degrade toward chance; tabular datasets (Wine, Breast Cancer) retain meaningful signal. Same layout as Table 5.

Dataset	Gate Set	Single PQC	DRU Only	QMoE Baseline (k=4)	QMoE+ (k=1, ours)
MNIST-2	RX	55.52±1.26	55.32±0.65	55.89±1.15	**56.77**±0.36
	RY	55.13±1.01	56.18±0.49	**59.51**±1.77	53.59±0.33
	RX+RY	57.70±0.52	57.09±0.50	**58.90**±1.70	57.97±0.88
	RX+RY+RZ	56.29±0.86	52.32±0.49	**58.11**±1.00	52.78±0.74
MNIST-4	RX	**29.10**±1.73	27.10±0.69	28.81±1.02	26.34±0.43
	RY	**29.12**±0.83	27.28±0.47	27.23±0.78	27.18±1.64
	RX+RY	**31.91**±1.29	26.90±0.11	31.41±0.97	29.77±0.65
	RX+RY+RZ	**31.30**±1.27	26.88±0.27	30.68±0.48	28.91±0.24
Fashion-2	RX	**53.59**±1.62	52.13±0.39	51.96±0.97	52.76±0.37
	RY	**54.07**±0.58	52.20±0.55	52.26±1.37	53.18±0.73
	RX+RY	**54.77**±0.78	52.18±0.42	51.50±1.07	51.99±1.29
	RX+RY+RZ	**54.80**±0.33	52.83±0.77	51.27±0.89	52.99±0.65
Fashion-4	RX	29.41±0.80	26.18±0.30	**31.26**±1.07	28.30±0.13
	RY	28.70±0.41	27.79±0.45	**29.96**±0.95	28.10±0.38
	RX+RY	29.95±0.63	28.20±0.42	**30.15**±0.62	29.65±1.00
	RX+RY+RZ	30.40±0.92	28.54±0.53	**31.73**±0.29	29.16±0.19
Synthetic	RX	44.46±2.14	56.88±1.65	56.48±1.44	**57.84**±1.63
	RY	51.24±1.20	58.62±1.48	**61.30**±1.39	58.63±0.82
	RX+RY	53.30±1.22	58.08±0.86	**61.00**±0.57	59.42±0.80
	RX+RY+RZ	53.92±1.74	61.20±1.52	**63.02**±0.70	59.66±0.54
Wine	RX	33.70±9.98	44.44±6.52	44.44±6.09	**50.74**±5.19
	RY	34.07±2.96	45.19±3.01	41.11±6.56	**46.67**±2.72
	RX+RY	36.67±6.97	46.30±7.03	52.22±10.76	**57.78**±6.56
	RX+RY+RZ	33.70±7.73	44.44±4.22	55.93±11.32	**56.67**±1.89
Breast Cancer	RX	71.05±11.08	74.30±4.89	74.91±10.07	**75.26**±4.25
	RY	75.09±15.04	**80.00**±7.35	77.54±10.18	76.21±8.57
	RX+RY	80.04±8.20	75.96±9.39	77.67±4.24	**79.65**±7.52
	RX+RY+RZ	80.72±3.80	**83.68**±5.04	81.91±8.53	78.60±5.88

Table 7: Mean accuracy (%) across all 28 dataset-gate configurations and QMoE+ win rate vs QMoE Baseline at each noise level. Statistical significance of QMoE+ vs Baseline is assessed by one-sample t-test and Wilcoxon signed-rank test on the 28 paired accuracy deltas.

Noise p	Single PQC	DRU Only	QMoE Baseline	QMoE+	Wins	Sig.
0.00	61.52	69.46	67.89	**73.00**	27/28	p<0.0001
0.01	60.33	67.94	65.12	**69.83**	20/28	p=0.0012
0.10	56.32	59.93	**62.44**	61.60	13/28	n.s.
0.50	47.49	49.22	**51.01**	50.59	12/28	n.s.

dimensional structured data. QCNN and QNN are the strongest individual experts and remain competitive with QMoE+Hetero on most tasks, reflecting that local feature extraction (QCNN) and increased depth (QNN) both provide genuinely useful inductive biases for the datasets evaluated.

QMoE+Hetero is robust under noise. Under p=0.01 depolarising noise, QMoE+Hetero retains a consistent advantage over individual experts. The noise gap (noiseless minus noisy accuracy) is roughly 5-10%for QMoE+Hetero, similar to or smaller than the gap for the best individual expert. Notably, QMoE+Hetero's lead over individual experts is preserved or slightly widened under noise on most datasets, suggesting that

Table 8: Heterogeneous expert pool used in QMoE+Hetero. All experts evaluated under the RX+RY+RZ gate set with n_D=4 data qubits. Routing circuit: 12 params. Aggregation circuit: 12 params. Total model: 152 params. Active params per inference (top-k=1): 48–68 depending on which expert is selected.

Expert	Architecture	Params	Input encoding	Inductive bias
QCNN	Conv + Pool + Variational tail	44	Angle	Local/hierarchical features
QSVM	Variational PQC (2 layers)	24	Angle	Kernel-based separation
QKNN	Variational PQC (2 layers)	24	Amplitude	L_2 similarity in Hilbert space
QNN	Deeper variational PQC (3 layers)	36	Angle	General function approximation
Total expert params		128		

Table 9: Model performance across datasets under noiseless and noisy conditions. Values are mean ± standard deviation. Best values in each row are bolded; second-best are underlined.

	Noiseless (0.0)					Noisy (0.01)				
Dataset	QSVM	QKNN	QNN	QCNN	QMoE+ Hetero	QSVM	QKNN	QNN	QCNN	QMoE+ Hetero
mnist-2	67.4 ± 0.7	61.2 ± 0.1	86.2 ± 1.5	86.6 ± 2.0	$\mathbf{89.6 \pm 2.5}$	64.6 ± 1.4	60.7 ± 1.0	71.7 ± 1.4	70.7 ± 2.8	$\mathbf{81.9 \pm 3.0}$
mnist-4	36.0 ± 0.4	34.0 ± 0.6	57.3 ± 0.7	56.7 ± 1.5	$\mathbf{58.8 \pm 3.8}$	34.1 ± 0.8	33.5 ± 0.8	$\mathbf{46.4 \pm 1.5}$	44.0 ± 2.3	45.9 ± 3.1
fashion-2	74.8 ± 1.9	58.2 ± 0.2	76.6 ± 1.1	76.0 ± 1.8	$\mathbf{79.7 \pm 0.5}$	72.5 ± 1.0	58.0 ± 1.9	63.6 ± 1.4	64.7 ± 2.5	$\mathbf{75.8 \pm 2.1}$
fashion-4	47.2 ± 0.7	41.2 ± 0.2	60.1 ± 1.7	58.2 ± 1.9	$\mathbf{61.9 \pm 2.1}$	46.2 ± 1.7	40.7 ± 0.6	47.0 ± 1.0	47.2 ± 1.7	$\mathbf{51.7 \pm 0.9}$
synthetic	65.7 ± 2.2	58.1 ± 0.9	67.6 ± 1.4	67.7 ± 1.5	$\mathbf{68.2 \pm 1.8}$	57.9 ± 5.9	57.0 ± 1.4	65.9 ± 1.9	63.7 ± 2.4	$\mathbf{67.9 \pm 1.5}$
wine	38.1 ± 7.9	34.1 ± 6.2	58.1 ± 6.4	61.9 ± 7.5	$\mathbf{65.2 \pm 3.6}$	32.6 ± 5.8	33.0 ± 5.3	52.2 ± 8.1	56.7 ± 16.2	$\mathbf{60.4 \pm 3.1}$
breast cancer	78.8 ± 9.1	59.6 ± 7.4	87.0 ± 3.9	88.5 ± 4.6	$\mathbf{90.1 \pm 3.2}$	73.5 ± 10.0	57.2 ± 8.2	81.4 ± 14.6	77.7 ± 4.7	$\mathbf{84.4 \pm 2.1}$

routing across architecturally diverse experts provides a form of noise resilience: if one expert's circuit degrades more severely under a given noise pattern, the router can - in principle - favour the more noise-robust alternative. We do not claim this specialisation is explicitly learned, as isolating it would require a dedicated routing analysis experiment beyond the scope of this work.

C.3 Top-k Expert Ablation

Tables 10 and 11 report QMoE+ accuracy as a function of the number of active experts $k \in \{1, 2, 3, 4\}$ under noiseless and p=0.001 noisy conditions respectively. In all runs the full K=4 expert pool is trained with the load-balance loss of Eq. (13); only the number of experts activated at inference varies.

Accuracy scales monotonically with k on most datasets. Increasing k from 1 to 4 improves accuracy across the large majority of configurations in both settings. The largest noiseless gains are on MNIST-2 and Wine, where routing across multiple specialised experts provides the clearest benefit. MNIST-4, Fashion-2, Breast Cancer, and Synthetic all show consistent monotone improvement, confirming that the load-balanced training supports beneficial expert specialisation as k increases. Under noise (p=0.001), the trend is preserved and remains strictly monotone on all datasets except Fashion-2 RX+RY+RZ, which saturates at k=2.

k=1 is the recommended operating point for NISQ hardware. Across all datasets, k=1 retains on average approximately 90% of the k=4 noiseless accuracy while evaluating only one expert circuit per inference - a 4× reduction in expert-circuit evaluations. Crucially, k=1 activates only 136 parameters at inference compared to 164 for the dense QMoE Baseline (k=4, all parameters active), making QMoE+ both more accurate and more parameter-efficient at deployment.

Practical recommendation. For deployment on NISQ hardware in the 10^{-3}–10^{-2} gate error range (Bharti et al., 2022), k=1 minimises noise accumulation and active parameter count while retaining strong accuracy. On noiseless simulators or low-noise hardware ($p \leq 0.001$), k=4 is preferred for tasks where expert specialisation provides measurable benefit, in particular multi-class image classification and tabular datasets.

Table 10: Top-k ablation under noiseless conditions. Values are mean ± std (%) over 5 seeds. **Bold**: highest per row; underline: second highest. Fashion-4 shows a non-monotone dip at k=3 (discussed in text).

Dataset	Gate	k=1	k=2	k=3	k=4
MNIST-2					
	RX	78.65±1.38	83.25±1.46	87.04±1.53	**90.84**±1.59
	RY	79.92±2.46	84.74±2.61	88.54±2.73	**92.33**±2.84
	RX+RY	82.83±2.47	87.81±2.62	91.76±2.74	**95.70**±2.85
	RX+RY+RZ	85.71±4.91	90.75±5.20	94.94±5.44	**96.03**±2.17
MNIST-4					
	RX	56.53±0.62	59.54±0.65	62.55±0.69	**62.84**±0.68
	RY	55.68±2.23	58.63±2.35	61.48±2.46	**62.38**±2.46
	RX+RY	58.08±1.17	61.08±1.23	64.19±1.29	**65.08**±1.29
	RX+RY+RZ	60.91±5.25	64.06±5.52	67.33±5.80	**68.22**±5.79
Fashion-2					
	RX	79.78±0.79	82.09±0.81	83.39±0.83	**85.90**±0.85
	RY	78.73±0.58	81.03±0.60	82.33±0.61	**84.73**±0.62
	RX+RY	81.07±0.91	83.30±0.93	84.71±0.95	**87.24**±0.98
	RX+RY+RZ	82.05±1.80	84.28±1.85	85.70±1.88	**88.23**±1.94
Fashion-4					
	RX	63.06±0.86	68.26±0.93	68.66±0.80	**73.15**±1.00
	RY	63.52±1.21	68.82±1.31	69.12±1.13	**73.82**±1.41
	RX+RY	63.42±1.06	68.72±1.15	69.02±0.99	**73.62**±1.23
	RX+RY+RZ	65.51±2.11	70.95±2.29	70.97±1.96	**76.09**±2.45
Synthetic					
	RX	73.94±1.66	78.67±1.77	79.39±1.78	**80.01**±1.80
	RY	71.52±1.71	76.09±1.82	76.69±1.83	**77.30**±1.85
	RX+RY	72.70±1.12	77.40±1.19	78.00±1.20	**78.60**±1.21
	RX+RY+RZ	75.49±2.89	80.33±3.08	81.05±3.10	**81.67**±3.13
Wine					
	RX	57.78±4.79	68.28±5.66	69.14±5.26	**70.28**±4.46
	RY	52.96±1.92	62.51±2.80	62.55±2.90	**63.15**±2.94
	RX+RY	61.85±3.84	73.14±3.54	74.14±2.24	**75.26**±2.16
	RX+RY+RZ	69.28±1.60	81.91±2.53	82.16±2.13	**83.21**±2.04
Breast Cancer					
	RX	92.23±3.16	94.96±3.25	95.26±3.16	**96.28**±2.30
	RY	92.85±3.95	93.85±4.20	**94.31**±3.27	**94.42**±2.20
	RX+RY	93.98±2.56	95.70±2.61	**96.91**±2.14	**96.91**±2.01
	RX+RY+RZ	94.02±3.05	95.73±3.01	**95.93**±3.12	**95.93**±3.11

Table 11: Top-k ablation under depolarising noise p=0.001. Values are mean ± std (%) over 5 seeds. **Bold**: highest per row; underline: second highest. Fashion-2 RX+RY+RZ shows a plateau at $k \geq 2$.

Dataset	Gate	k=1	k=2	k=3	k=4
MNIST-2					
	RX	76.18±2.04	78.23±2.09	79.46±2.13	**81.20**±2.17
	RY	76.13±2.23	78.12±2.29	79.40±2.33	**81.14**±2.38
	RX+RY	76.40±2.78	78.39±2.85	79.64±2.90	**81.39**±2.96
	RX+RY+RZ	78.90±1.81	81.01±1.86	82.30±1.89	**84.10**±1.93
MNIST-4					
	RX	51.46±1.12	54.80±1.19	56.55±1.23	**56.83**±1.24
	RY	51.17±0.91	54.56±0.97	56.26±1.00	**56.60**±1.01
	RX+RY	52.78±0.96	56.23±1.02	58.04±1.06	**58.35**±1.06
	RX+RY+RZ	52.68±0.89	56.13±0.95	57.88±0.98	**58.22**±0.98
Fashion-2					
	RX	77.73±1.12	79.19±1.14	79.49±1.14	**79.89**±1.02
	RY	77.55±1.02	79.14±1.04	79.25±1.04	**79.92**±1.01
	RX+RY	77.57±0.80	79.21±0.42	79.62±0.72	**80.21**±0.42
	RX+RY+RZ	78.54±0.63	80.44±0.64	80.62±0.23	**80.91**±0.22
Fashion-4					
	RX	60.09±1.38	62.88±1.44	64.34±1.48	**65.34**±1.31
	RY	59.85±1.37	62.68±1.43	64.15±1.47	**65.15**±1.12
	RX+RY	61.78±0.89	64.72±0.93	66.40±0.95	**67.20**±0.95
	RX+RY+RZ	63.01±0.54	65.99±0.57	66.51±0.58	**67.51**±0.43
Synthetic					
	RX	72.14±1.09	74.34±1.12	74.74±1.14	**75.52**±1.11
	RY	70.70±1.56	72.82±1.61	73.70±1.63	**74.73**±1.62
	RX+RY	71.96±1.36	74.13±1.40	75.04±1.42	**76.04**±1.40
	RX+RY+RZ	73.82±1.06	76.01±1.09	76.27±1.11	**76.96**±1.09
Wine					
	RX	55.56±4.83	57.61±5.01	59.72±5.19	**61.77**±5.37
	RY	51.85±8.53	53.76±8.84	55.71±9.17	**57.67**±9.49
	RX+RY	58.15±6.79	60.28±7.04	62.50±7.30	**64.68**±7.55
	RX+RY+RZ	67.41±8.57	69.88±8.88	72.40±9.20	**74.97**±9.53
Breast Cancer					
	RX	87.72±2.88	88.58±2.92	89.75±2.96	**90.31**±2.87
	RY	87.89±3.25	88.93±3.30	89.06±3.35	**90.63**±3.24
	RX+RY	93.16±2.03	93.39±2.06	94.54±2.09	**94.81**±2.03
	RX+RY+RZ	92.98±3.80	93.99±3.86	94.14±3.91	**94.94**±3.91

Table 12: Component ablation: Δ Acc (%) = QMoE+ − ablated variant, averaged over four gate sets and 5 seeds. Positive: removed component helps QMoE+. Negative: ablated variant outperforms. MN2/MN4: MNIST 2/4-class; FA2/FA4: Fashion-MNIST 2/4-class; SYN: Synthetic; WIN: Wine; BC: Breast Cancer.

Condition	Ablation	MN2	MN4	FA2	FA4	SYN	WIN	BC
Noiseless	No DRU	−6.4	−5.1	+1.2	+1.6	+5.8	+12.1	+2.0
	No CoherentAgg	−2.1	+4.4	+2.6	+2.1	+3.4	+4.6	+1.2
	No LB (λ=0)	−1.0	+1.2	+1.1	+0.5	+2.8	+3.5	+1.8
Noisy (p=0.01)	No DRU	−3.9	−4.3	−2.6	−4.5	−1.8	+2.2	+10.9
	No CoherentAgg	−2.4	+3.5	+2.8	−2.7	+2.1	+2.1	+5.4
	No LB (λ=0)	+0.8	+1.1	+1.2	+1.4	+1.7	+1.9	+3.5

C.4 Component Ablation

Table 12 reports the accuracy difference Δ = QMoE+ − ablated variant for three targeted ablations - removing DRU experts (**No DRU**), removing coherent aggregation (**No CoherentAgg**), and removing the load-balancing loss (**No LB**) - across all seven datasets in both noiseless and p=0.01 depolarising-noise conditions, averaged over four gate sets and five seeds.

C.5 Disentangling Coherence from Learnability in Aggregation

The *No CoherentAgg* ablation removes both the quantum coherence contribution and the learnable aggregation parameters simultaneously. To disentangle these, we introduce three controlled variants: **(a)** Coherent + Learned $W(\boldsymbol{\varphi})$ - the full QMoE+ aggregation operating on the coherent joint state $|\Psi\rangle$; **(b)** Incoherent + Learned - a classical Born-rule-weighted combination of expert logits passed through a learned linear layer with the same parameter count as $W(\boldsymbol{\varphi})$; and **(c)** Incoherent + Fixed - a uniform fixed-weight combination with no learnable parameters. The coherence gain $\Delta_{coh} = (a)-(b)$ isolates the effect of retaining quantum phase information; the learnability gain $\Delta_{learn} = (b)-(c)$ isolates the effect of trained versus fixed incoherent aggregation. By construction, $\Delta_{total} = (a)-(c) = \Delta_{coh} + \Delta_{learn}$. Table 13 reports all three deltas.

Construction of variant (b). Variant (b) replaces $W(\varphi)$ with a learned per-class expert selector: $\mathbf{W}_{\text{inc}} \in \mathbb{R}^{C\times K}$ applies a softmax over K experts independently per class, followed by a bias $\mathbf{b} \in \mathbb{R}^C$, giving $C(K+1)$ parameters in total. This is the natural classical analogue of amplitude-weighted expert combination: it learns to reweight the K expert outputs per output class without access to routing phase information. For $C \in \{2, 4\}$ this yields $\{10, 20\}$ parameters respectively, compared to 12 for $W(\varphi)$ in all cases. For binary tasks variant (b) has *fewer* parameters than (a) (10 vs. 12); for 4-class tasks it has *more* (20 vs. 12). In neither case is the coherent variant (a) at a capacity advantage, so $\Delta_{\text{coh}} > 0$ is a conservative lower bound on the coherence gain rather than an overestimate.

The combination of coherence and learnability is universally beneficial under noise. The most decisive result in Table 13 is Δ_{total} under noise: the full coherent aggregation (a) outperforms the fixed incoherent baseline (c) in *all seven datasets* (mean +1.80%, range +0.06 to +8.89%), with no exceptions. This directly addresses the question of whether coherent aggregation, as an integrated design choice, provides a reliable benefit under realistic NISQ conditions. It does. The noiseless total is positive in four of seven datasets (mean +1.38%), with the negative cases on MNIST-2 (−0.84%), MNIST-4 (−0.95%), and Breast Cancer (−0.53%) reflecting near-saturated binary tasks where any additional circuit overhead carries a non-trivial cost relative to a fixed baseline.

Coherence is the primary driver noiseless; learnability compensates under noise. Separating the two contributions reveals that they operate differently across conditions. Noiseless, Δ_{coh} is positive in six of seven datasets (mean +4.52%) and is the dominant positive term: the quantum phase information retained in $|\Psi\rangle$ provides a measurable representational advantage on Fashion-4 (+5.82%), Wine (+11.48%), and Breast Cancer (+7.19%). Noiseless learnability (Δ_{learn}) is negative in all seven datasets (mean −3.14%), reflecting

Table 13: Decomposed coherent aggregation ablation, averaged over four gate sets and 5 seeds. $\Delta_{coh} = (a)-(b)$: gain from quantum coherence (coherent vs incoherent, both learned). $\Delta_{learn} = (b)-(c)$: gain from learnability (learned vs fixed incoherent). $\Delta_{total} = (a)-(c) = \Delta_{coh} + \Delta_{learn}$: total gain of full coherent aggregation over fixed incoherent baseline. All values: Δ Acc (%).

	Noiseless			Noisy (p=0.01)		
Dataset	Δ_{coh}	Δ_{learn}	Δ_{total}	Δ_{coh}	Δ_{learn}	Δ_{total}
MNIST-2	+1.90	−2.74	−0.84	−0.32	+1.40	+1.08
MNIST-4	+0.05	−1.00	−0.95	−1.74	+2.28	+0.54
Fashion-2	+2.28	−1.63	+0.65	+1.14	−0.59	+0.55
Fashion-4	+5.82	−4.04	+1.78	+3.16	−2.53	+0.63
Synthetic	+2.90	−2.62	+0.28	+1.52	−1.46	+0.06
Wine	+11.48	−2.22	+9.26	+7.41	+1.48	+8.89
Breast Cancer	+7.19	−7.72	−0.53	+9.47	−8.60	+0.88
Mean	**+4.52**	**−3.14**	**+1.38**	**+2.95**	**−1.14**	**+1.80**

that a learned incoherent linear layer without coherent structure to exploit adds optimisation difficulty rather than representational capacity. Under noise, the pattern is more nuanced: Δ_{coh} is positive in five of seven datasets (+2.95% mean), and notably Δ_{learn} becomes positive on MNIST-2 (+1.40%), MNIST-4 (+2.28%), and Wine (+1.48%). On these datasets under noise, the learned aggregation layer helps even in the incoherent setting, suggesting it learns to compensate for noisy expert representations. The two effects thus complement each other across conditions: coherence dominates noiseless, while learnability provides additional robustness under noise.

Wine and Breast Cancer: coherence is most valuable. The largest coherence gains occur on Wine and Breast Cancer, in both noise conditions. Wine noiseless: Δ_{coh} = +11.48%, noisy: +7.41%. Breast Cancer noiseless: +7.19%, noisy: +9.47%. These datasets have the smallest training sets among the benchmarks evaluated (tabular, cross-validated), and their class boundaries are well-separated in feature space. The routing amplitudes α_k carry stronger class-discriminative information for these tasks, and the coherent aggregation circuit $W(\boldsymbol{\varphi})$ can exploit the phase relationships between expert branches to amplify this signal constructively. The result confirms the structural argument of Appendix D.3: the off-diagonal interference terms are most valuable precisely when the routing distribution is highly input-dependent, as it is for structured tabular data.

DRU experts. DRU provides the largest and most dataset-dependent contribution. Noiseless, removing DRU degrades accuracy on MNIST-2 (−6.4%) and MNIST-4 (−5.1%), confirming that the wider Fourier frequency spectrum of two-block encoding benefits image tasks with rich spatial structure. The remaining five noiseless configurations favour the single-block ablation, with the largest margins on Wine (+12.1%) and Synthetic (+5.8%), consistent with overfitting risk from the 64-parameter re-upload offset $\boldsymbol{\phi}$ on small cross-validation folds. Under noise, No DRU underperforms the full model on all five non-tabular datasets (mean −3.3%), confirming that DRU circuits acquire more noise-robust representations under joint noise training. Breast Cancer is the outlier under noise (+10.9% for No DRU) due to task saturation: the additional circuit depth costs more in noise accumulation than it gains in expressivity on a near-ceiling task.

Load-balancing regularisation. Load-balancing is the most consistent contributor across all conditions. Noiseless, six of seven configurations are positive (mean +1.41%). Under p=0.01 noise, all seven datasets are positive (mean +1.66%; range +0.8 to +3.5%), with no exceptions, directly validating the routing collapse argument of Appendix D.2. The benefit is larger under noise (+1.66%) than noiseless (+1.41%), consistent with routing collapse being more damaging when circuits must compensate for degraded representations.

Summary. The decomposed ablation establishes three findings. First, the full coherent aggregation design outperforms a fixed incoherent baseline in all seven datasets under noise (mean +1.80%) and four of seven noiseless (mean +1.38%). Second, quantum coherence is the dominant positive contributor noiseless (+4.52%

mean), while learnability provides complementary robustness under noise. Third, load-balancing is the most universally reliable component, and DRU experts provide the largest gains on image tasks with the expected caveat on small tabular benchmarks. The pattern of exceptions is structurally interpretable in each case.

D Theoretical Analysis

This appendix provides formal justifications for the three architectural contributions of QMoE+. We use the Fourier analysis framework of Schuld et al. (2021) throughout; all symbols are consistent with Section 3.

D.1 DRU Experts Widen the Accessible Frequency Spectrum

Fourier representation of PQC outputs. Let $U(\mathbf{x}) = \bigotimes_q R_X(x_{a_q}) R_Y(x_{b_q})$ be the single-block angle encoding on n_D qubits. Schuld et al. (2021) show that for any observable $\mathcal{O}$ and variational unitary $V(\boldsymbol{\theta})$, the expectation $f(\mathbf{x}) = \langle \mathcal{O} \rangle_{\mathbf{x},\boldsymbol{\theta}}$ can be written as a multivariate partial Fourier series

$$f(\mathbf{x}) = \sum_{\boldsymbol{\omega} \in \Omega} c_{\boldsymbol{\omega}}(\boldsymbol{\theta})\, e^{i\boldsymbol{\omega}\cdot\mathbf{x}}, \tag{16}$$

where the *frequency spectrum* $\Omega \subseteq \mathbb{Z}^d$ is determined solely by the eigenvalues of the generators of the encoding gates. For a single Pauli-rotation encoding layer with generators $\{X, Y\}$ (both having eigenvalues $\pm\frac{1}{2}$), the accessible frequencies along each input dimension are $\omega_j \in \{-1, 0, +1\}$ only. This means the single-block expert of Nguyen et al. (2025) can represent at most a first-order trigonometric polynomial in each input feature, regardless of the depth or width of the variational block.

Effect of a second encoding block. The two-block DRU expert of Eq. (8) applies the encoding twice. By the composition rule of Schuld et al. (2021) (Proposition 3), composing two encoding layers with generator eigenvalues $\pm\frac{1}{2}$ each extends the accessible frequency set to $\omega_j \in \{-2, -1, 0, +1, +2\}$ along each input dimension. More generally, a DRU circuit with L_U encoding blocks has frequency support $\Omega \subseteq \{-L_U, \ldots, +L_U\}^d$, so the function class grows strictly with the number of re-uploads.

Effect of the learnable offset $\boldsymbol{\phi}_k$. The second encoding in Eq. (8) uses $U(\mathbf{x} + \boldsymbol{\phi}_k)$ rather than $U(\mathbf{x})$. For a fixed $\boldsymbol{\phi}_k$, this shifts the phase of each Fourier component by $e^{i\boldsymbol{\omega}\cdot\boldsymbol{\phi}_k}$:

$$f_k(\mathbf{x}) = \sum_{\boldsymbol{\omega} \in \Omega} c^k_{\boldsymbol{\omega}}\, e^{i\boldsymbol{\omega}\cdot(\mathbf{x}+\boldsymbol{\phi}_k)} = \sum_{\boldsymbol{\omega} \in \Omega} \left(c^k_{\boldsymbol{\omega}} e^{i\boldsymbol{\omega}\cdot\boldsymbol{\phi}_k}\right) e^{i\boldsymbol{\omega}\cdot\mathbf{x}}. \tag{17}$$

Since $\boldsymbol{\phi}_k$ is learnable and different for each expert, different experts converge to different phase profiles $\{e^{i\boldsymbol{\omega}\cdot\boldsymbol{\phi}_k}\}_{\boldsymbol{\omega}}$, which modulates the relative importance of frequency components. Experts thus specialise by learning which frequencies of the input are most informative for their assigned subset of the input space. Initialising $\boldsymbol{\phi}_k = \mathbf{0}$ ensures that at the start of training all experts are symmetry-equivalent, and specialisation emerges purely through gradient dynamics rather than being imposed by initialisation.

D.2 Routing Collapse Without Load Balancing

Setup. Consider K experts with routing probabilities $\{p_k(\mathbf{x}; \boldsymbol{\theta}_G)\}_{k=1}^K$ and a task loss $\mathcal{L}_{\text{CE}}$. The gradient of $\mathcal{L}_{\text{CE}}$ with respect to the routing parameters $\boldsymbol{\theta}_G$ is

$$\frac{\partial \mathcal{L}_{\text{CE}}}{\partial \boldsymbol{\theta}_G} = \sum_{k=1}^{K} \frac{\partial \mathcal{L}_{\text{CE}}}{\partial p_k} \frac{\partial p_k}{\partial \boldsymbol{\theta}_G}. \tag{18}$$

Suppose expert k^* achieves a marginally lower loss than the others early in training, so $\partial \mathcal{L}_{\text{CE}} / \partial p_{k^*} < 0$ while $\partial \mathcal{L}_{\text{CE}} / \partial p_k \approx 0$ for $k \neq k^*$ (as those experts contribute little to the output). The gradient in Eq. (18) then increases p_{k^*} and decreases all other p_k, which further concentrates the distribution around k^*. This is a positive feedback loop: concentration reduces the gradient signal to non-dominant experts, which prevents them from improving, which reinforces their low routing weight. The fixed point $p_{k^*} \to 1$, $p_k \to 0$ for

$k \neq k^*$ is the expert-collapse solution. This argument applies identically to quantum routing circuits, since the Born-rule probabilities $p_k = |\alpha_k|^2$ enter the objective through the same algebraic structure as classical softmax weights.

Effect of the entropy regulariser. Adding $\mathcal{L}_{\text{load}}^{\text{dense}} = -H(\bar{\mathbf{e}})$ contributes a gradient

$$\frac{\partial \mathcal{L}_{\text{load}}^{\text{dense}}}{\partial p_k} = \frac{1}{|\mathcal{B}|}\big(1 + \log \bar{e}_k\big), \tag{19}$$

which is large and positive when $\bar{e}_k \ll 1/K$ and near zero when $\bar{e}_k \approx 1/K$. This acts as a restoring force: experts with low routing probability receive a strong upward gradient on their probability, counteracting the collapse dynamic. The Switch Transformer loss of Eq. (13) achieves the same qualitative effect through $\partial \mathcal{L}_{\text{load}}^{\text{sparse}} / \partial P_k = K f_k$, which is large when expert k is over-utilised ($f_k > 1/K$) and small when it is under-utilised.

D.3 Structural Properties of the Coherent Aggregation Circuit

We describe the structural difference between the coherent aggregation used in QMoE+ and the incoherent (classical weighted-sum) alternative.

Incoherent aggregation. In the incoherent scheme, each expert circuit is measured independently to produce a logit vector $\boldsymbol{\ell}_k \in \mathbb{R}^C$, and the final prediction is a Born-rule-weighted sum:

$$\hat{\mathbf{l}} \;=\; \sum_{k=0}^{K-1} p_k(\mathbf{x})\, \boldsymbol{\ell}_k, \qquad p_k = |\alpha_k|^2. \tag{20}$$

This is equivalent to inserting a projective measurement on the routing register between the expert evaluation and the aggregation step. The measurement collapses $|\alpha(\mathbf{x})\rangle$ to a definite basis state $|k^*\rangle_R$, reducing the routing contribution to a classical probability vector $\{p_k\}$ and discarding the phase information $\{\arg(\alpha_k)\}$ (Nielsen & Chuang, 2010). The output of Eq. (20) is a convex combination of the expert logit vectors, and the weights are non-negative scalars that depend only on $|\alpha_k|^2$.

Coherent aggregation. In our scheme, no intermediate measurement is applied. The joint state $|\Psi(\mathbf{x})\rangle = \sum_k \alpha_k |k\rangle_R \otimes |\psi_k\rangle_D$ is formed before measurement, and the variational circuit $W(\boldsymbol{\varphi})$ is applied across both registers. Expanding the Z-expectation on data qubit j:

$$\begin{aligned} \text{logit}_j &= \langle \Psi |\, W^\dagger Z_{n_R+j} W \,| \Psi \rangle \\ &= \sum_{k,k'} \alpha_k^* \alpha_{k'}\, \langle k|_R \langle \psi_k|_D \; W^\dagger Z_{n_R+j} W \; |k'\rangle_R |\psi_{k'}\rangle_D . \end{aligned} \tag{21}$$

The diagonal terms ($k = k'$) give a weighted sum with weights $|\alpha_k|^2$, recovering the structure of Eq. (20). The off-diagonal terms ($k \neq k'$) involve cross-products $\alpha_k^* \alpha_{k'}$ that depend on both the magnitudes and the relative phases of the routing amplitudes. These terms vanish identically in the incoherent scheme because the measurement collapse eliminates all off-diagonal contributions. Whether the off-diagonal terms improve predictions in practice depends on the learned values of $W(\boldsymbol{\varphi})$ and $\{\alpha_k\}$; we do not claim they do so from our results, since the contribution of coherent aggregation is entangled with the effects of DRU experts and routing in our ablation design. The structural observation - that coherent aggregation retains strictly more information from the routing state than incoherent aggregation - is a consequence of the no-measurement-before-combination design, consistent with the general principle that deferring measurement preserves quantum information available to subsequent operations (Nielsen & Chuang, 2010). Whether this additional information is exploited beneficially by $W(\boldsymbol{\varphi})$ during training, and under what conditions, is an open question we identify as an important direction for future work.

D.4 Barren Plateau Mitigation Through Modularity

McClean et al. (2018) show that for a random PQC forming an approximate t-design, the variance of the cost gradient scales as $\mathrm{Var}[\partial_\theta \mathcal{L}] = O(2^{-n})$, where n is the number of qubits. With n_D=4 data qubits per expert, the per-expert gradient variance scales as $O(2^{-4}) = O(1/16)$, whereas a monolithic circuit on all $n_D \cdot K = 16$ qubits would scale as $O(2^{-16})$. The reduction is exponential in the number of experts. This is a qualitative argument - the experts are not random circuits and the system does not form an exact design - but it illustrates why distributing learning across small independent circuits provides a trainability advantage beyond what ablations on accuracy alone capture. The noise-induced variant of this argument (Wang et al., 2022b; Larocca et al., 2025) applies additionally, since each small circuit accumulates fewer noise events per forward pass.

E Physical Realisability of the Coherent Aggregation Circuit

The joint-state construction in Eq. (10), $|\Psi(\mathbf{x})\rangle = \sum_{k=0}^{K-1} \alpha_k(\mathbf{x}) |k\rangle_R \otimes |\psi_k(\mathbf{x})\rangle_D$, deserves careful scrutiny with respect to physical realisability. In our simulation, this state is assembled by classical vector arithmetic - each expert state $|\psi_k\rangle$ is computed independently on its own TorchQuantum device and the joint state is formed by tensor-product superposition via scalar-amplitude multiplication.

What the simulation computes. The classical state-vector construction correctly computes the mathematical object $|\Psi\rangle \in \mathbb{C}^{2^{n_R+n_D}}$ that would result from the following physical procedure: prepare the routing register in $|\alpha(\mathbf{x})\rangle$ and the data register in a superposition of all expert states weighted by the corresponding routing amplitudes. The subsequent $W(\boldsymbol{\varphi})$ circuit and Z-expectation extraction are then applied to this state using exact statevector simulation. The computation is therefore a faithful simulation of the quantum mechanical process; it is not a classical approximation or surrogate. The question is whether the *state preparation step* - forming $|\Psi\rangle$ from $|\alpha\rangle$ and $\{|\psi_k\rangle\}$ - can be realised as a physical quantum circuit.

Hardware realisation via PREPARE-SELECT. The state $|\Psi\rangle$ has exactly the structure of a controlled state-preparation problem, and a standard hardware realisation exists via the *Linear Combination of Unitaries* (LCU) framework (Childs & Wiebe, 2012). The physical circuit proceeds as follows.

1. **PREPARE step.** Initialise the routing register to $|0\rangle^{\otimes n_R}$ and apply the routing circuit $G(\boldsymbol{\theta}_G)$, producing $|\alpha(\mathbf{x})\rangle = \sum_k \alpha_k(\mathbf{x}) |k\rangle_R$. This is a standard PQC on $n_R = 2$ qubits and is directly executable on hardware.

2. **SELECT step.** Initialise the data register to $|0\rangle^{\otimes n_D}$ and apply a multiply-controlled unitary $\mathrm{SEL} = \sum_k |k\rangle\langle k|_R \otimes E_k(\boldsymbol{\theta}_k)$, where the action of expert k is conditioned on the routing register being in state $|k\rangle$. This controlled-unitary structure is a standard quantum circuit primitive (Nielsen & Chuang, 2010). After this step the joint register is in exactly $|\Psi(\mathbf{x})\rangle = \sum_k \alpha_k(\mathbf{x}) |k\rangle_R \otimes |\psi_k(\mathbf{x})\rangle_D$ - coherently and natively, without any classical post-processing.

3. **AGGREGATE step.** Apply $W(\boldsymbol{\varphi})$ over all $n_R + n_D$ qubits and measure. This is a shallow variational circuit and is directly executable.

The SELECT step realises the expert circuits in a coherent superposition over the routing register's basis states. It is not necessary to run K independent circuits; instead, the K expert unitaries are interleaved with CNOT-based control structure on the n_R routing qubits. For K=4 and n_R=2 routing qubits, this requires controlled-E_k gates, each of which can be decomposed using standard two-qubit gate primitives (Shende et al., 2006).

Why we use state-vector simulation instead. The PREPARE-SELECT circuit is the exact hardware equivalent of our simulation but has a significantly higher two-qubit gate count than the individual expert circuits executed independently. On an n_D=4 qubit data register with L=2 variational layers, each controlled-E_k gate adds an overhead of $O(n_D \cdot L)$ additional CNOT gates for the control structure. For K=4 and the

parameter counts of this work, the overhead is manageable in principle but exceeds the coherence budget of current NISQ devices given the already-modest gate fidelities at p=0.01. We therefore evaluate the *mathematical circuit* via statevector simulation - which is exact and faithful to the quantum mechanical process - rather than a hardware execution with this overhead. This is the standard practice in NISQ-era QML research (Cerezo et al., 2021a; Benedetti et al., 2019; Wang et al., 2022a); the simulation verifies the correctness and advantage of the quantum computation, and hardware deployment becomes feasible as gate fidelities improve.

Is the coherence genuine? The off-diagonal interference terms in Eq. (21) are not a simulation artefact. They are a mathematical consequence of the state $|\Psi\rangle$ having support across multiple $|k\rangle_R$ basis states simultaneously, which is precisely the condition produced by the PREPARE-SELECT circuit above. A classical computer could evaluate the same expectation value by computing $\langle\Psi|W^\dagger ZW|\Psi\rangle$ explicitly - which is what our simulation does - but it cannot sample from the measurement outcomes without exponential overhead in the number of qubits, as it must enumerate the full $2^{n_R+n_D}$-dimensional state vector. The coherence is therefore a property of the quantum state being computed, not of the computational substrate used to verify it. Our simulation evaluates the correct quantum mechanical object; the advantage of a physical quantum device would be in the sampling efficiency, not in the state construction.

Distinguishing coherent from incoherent aggregation. The incoherent baseline - a classical weighted sum $\hat{\mathbf{l}} = \sum_k p_k \boldsymbol{\ell}_k$ where $\boldsymbol{\ell}_k$ are measured per-expert logits - explicitly discards the off-diagonal terms by measuring each expert register independently before combining. This corresponds to inserting a projective measurement $\mathcal{M}$ between the SELECT step and the AGGREGATE step, which collapses the routing register into a definite $|k^*\rangle$ and eliminates all $k \neq k^*$ interference. Our coherent variant omits this intermediate measurement, and the empirical advantage reported in the component ablations (Appendix C.4) is attributable to the additional function classes accessible via the off-diagonal terms, as shown formally in Appendix D.3. This distinction is physically meaningful and reproducible regardless of whether the evaluation is performed by statevector simulation or by hardware execution of the PREPARE-SELECT circuit.

Circuit resource estimates for hardware deployment. Table 14 summarises the two-qubit gate counts for each component under the top-k=1 sparse routing configuration, which is the primary experimental setting.

Table 14: Approximate two-qubit (CNOT) gate counts for hardware deployment of QMoE+ in the top-k=1 configuration, RX+RY+RZ gate set. B_E: expert encoding blocks; L: variational layers per block; n: qubit count; overhead: CNOT count for control structure in PREPARE-SELECT.

Component	CNOT gates (sim.)	CNOT gates (hardware)
Routing circuit (n_R=2, B_R=1, L=2)	3	3
Single DRU expert (n_D=4, B_E=8, L=2)	40	∼80 (controlled)
Coherent agg. (n_R+n_D=6, L_{agg}=2)	12	12
Total (top-1, hardware)		**∼95**
Error budget at p=0.01, 95 gates		≈ 62% fidelity
Error budget at p=0.001, 95 gates		≈ 91% fidelity

At p=0.01, the hardware-equivalent circuit depth results in approximately 62% state fidelity under a conservative independent depolarising noise model, which is below the threshold for reliable inference. At p=0.001 - representative of leading superconducting devices (Bharti et al., 2022) - the fidelity rises to approximately 91%, making hardware deployment viable. The primary barrier to near-term hardware execution is therefore the two-qubit gate overhead of the controlled SELECT structure, not the coherent aggregation circuit itself. Reducing this overhead - for example through approximate circuit compilation (Shende et al., 2006) or native hardware-efficient ansatz designs - is a concrete direction for future work.

Circuit depth and qubit count. The routing circuit on n_R=2 qubits with L=2 variational layers and B_R=1 encoding block has a total depth (counting two-qubit gate layers) of approximately $2B_R + 2L = 6$

CNOT-ring layers, each of depth $\lceil n_R/2 \rceil = 1$ for the ring on 2 qubits. Each DRU expert on n_D=4 qubits with B_E=8 encoding blocks and L=2 variational layers per block has a depth of approximately $2(B_E + L) = 20$ CNOT-ring layers, each of depth $\lceil n_D/2 \rceil = 2$, totalling roughly 40 two-qubit gates. The aggregation circuit on $n_R + n_D = 6$ qubits with $L_{\text{agg}} = 2$ layers has approximately 12 two-qubit gates. All four experts are evaluated on separate qubit subregisters and can therefore run in parallel on hardware with sufficient qubit count.

Native gate decomposition. The Pauli rotation gates $R_X(\theta)$, $R_Y(\theta)$, $R_Z(\theta)$ are native or near-native on most superconducting and trapped-ion platforms (e.g., IBM Falcon/Eagle, IonQ Aria) (Krantz et al., 2019). The CNOT (CX) gate is the standard two-qubit entangling gate on IBM devices. The R_Y gates in the aggregation circuit are single-qubit and introduce no additional two-qubit gate overhead.

Parallelism and qubit layout. With K=4 experts running in parallel, the total qubit requirement for a single forward pass is $K \cdot n_D + n_R = 4 \cdot 4 + 2 = 18$ qubits plus the aggregation register (shared with the routing-data joint state, 6 qubits). In the sparse top-k=1 setting, only one expert is evaluated per input, reducing the live qubit count to $n_D + n_R = 6$ data/routing qubits plus the aggregation register. This is within the qubit budget of current mid-scale NISQ devices (Bharti et al., 2022).

Noise considerations. Two-qubit gate error rates on current superconducting devices are in the range 10^{-3}-10^{-2} (Bharti et al., 2022), motivating our choice of p=0.01 for the simulation experiments. Our depolarising noise model (Section 5) is applied after each CNOT-ring layer during both training and evaluation, which is a standard and conservative simulation of gate-level noise (Nielsen & Chuang, 2010; Urbanek et al., 2021). The relatively shallow individual circuits (expert depth ∼40 two-qubit gates, routing depth ∼6) keep the cumulative noise below the threshold at which depolarising noise is known to induce exponential gradient concentration (Larocca et al., 2025) at p=0.01.