HYBRID QUANTUM-CLASSICAL RECURRENT NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a hybrid quantum-classical recurrent neural network (QRNN) architecture in which the recurrent core is realized as a parametrized quantum circuit (PQC) controlled by a classical feedforward network. The hidden state is the quantum state of an n-qubit PQC in an exponentially large Hilbert space $\mathbb{C}^{2^n}$, which serves as a coherent recurrent quantum memory. The PQC is unitary by construction, making the hidden-state evolution norm-preserving without external constraints. At each timestep, mid-circuit Pauli expectation-value readouts are combined with the input embedding and processed by the feedforward network, which provides explicit classical nonlinearity. The outputs parametrize the PQC, which updates the hidden state via unitary dynamics. The QRNN is compact and physically consistent, and it unifies (i) unitary recurrence as a high-capacity memory, (ii) partial observation via mid-circuit readouts, and (iii) nonlinear classical control for input-conditioned parametrization. We evaluate the model in simulation with up to 14 qubits on sentiment analysis, MNIST, permuted MNIST, copying memory, and language modeling. For sequence-to-sequence learning, we further devise a soft attention mechanism over the mid-circuit readouts and show its effectiveness for machine translation. To our knowledge, this is the first model (RNN or otherwise) grounded in quantum operations to achieve competitive performance against strong classical baselines across a broad class of sequence-learning tasks.

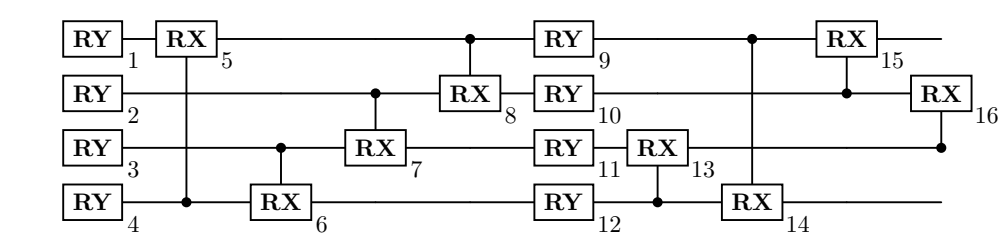
1 INTRODUCTION

Recurrent neural networks (RNNs) process sequence data by maintaining a hidden state that is updated at each timestep, which can create a bottleneck for memory and representational capacity. While vanilla RNNs have been empirically shown to retain roughly one real value of information per hidden unit, with the effective task-specific capacity linearly bounded by the number of model parameters (Collins et al., 2017), similar limitations extend to gated architectures such as LSTMs and GRUs (Hochreiter and Schmidhuber, 1997; Cho et al., 2014), despite their use of gating and explicit memory cells (Collins et al., 2017). This means that more complex sequences may exceed what the hidden state can encode, forcing the model to compress or forget.

Another challenge in training RNNs is the vanishing and exploding gradient problem (Bengio et al., 1994; Hochreiter and Schmidhuber, 1997), which arises from repeated multiplication through the recurrent Jacobian. Among various strategies to address this (Mikolov, 2012; Pascanu et al., 2013; Le et al., 2015), unitary and orthogonal RNNs (Arjovsky et al., 2016; Jing et al., 2019; Helfrich et al., 2018; Kiani et al., 2022) constrain the recurrent weights to be norm-preserving, allowing gradients to remain stable across timesteps.

The introduction of the Transformer (Vaswani et al., 2017) appeared to obviate explicit recurrence by bypassing the hidden-state bottleneck. However, recent work shows that recurrent inductive bias remains highly competitive and provides representational advantages not matched by Transformers (Gu and Dao, 2023; Orvieto et al., 2023; Bhattamishra et al., 2024; Beck et al., 2024).

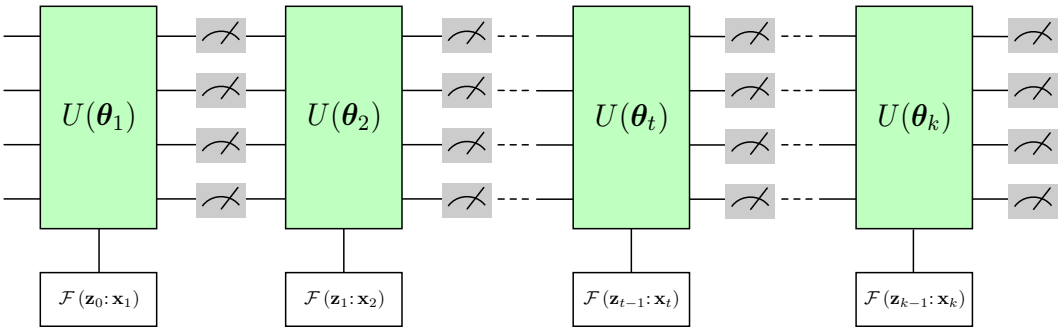
With the advancement of quantum computing (Arute et al., 2019; Acharya et al., 2024; Reichardt et al., 2024; DeCross et al., 2025), parametrized quantum circuits (PQCs), which are a core component of variational hybrid quantum-classical models, have concurrently emerged as an alternative mechanism for function approximation (Benedetti et al., 2019; Du et al., 2019; Bondesan and Welling, 2020; Pérez-Salinas et al., 2021; Schuld et al., 2021; Yu et al., 2024b). PQCs implement unitary transformations



(a)

$$\boxed{\mathbf{RX}}_i = \begin{bmatrix} \cos\left(\frac{\theta_i}{2}\right) & -i\sin\left(\frac{\theta_i}{2}\right) \\ -i\sin\left(\frac{\theta_i}{2}\right) & \cos\left(\frac{\theta_i}{2}\right) \end{bmatrix}, \quad \boxed{\mathbf{RY}}_i = \begin{bmatrix} \cos\left(\frac{\theta_i}{2}\right) & -\sin\left(\frac{\theta_i}{2}\right) \\ \sin\left(\frac{\theta_i}{2}\right) & \cos\left(\frac{\theta_i}{2}\right) \end{bmatrix}.$$

(b)



(c)

Figure 1: Hybrid QRNN. (a) Recurrent core PQC with $n = 4$ qubits (illustrative) and 16 parametrized gates, acting on a quantum state in the Hilbert space $\mathbb{C}^{2^n}$; each horizontal line corresponds to one qubit. (b) $\mathbf{RX}$ and $\mathbf{RY}$ gates.[1] Each gate in the PQC is parametrized by a rotation angle θ_i, where $1 \leq i \leq 16$. (c) QRNN unrolled for a sequence of length k. The feedforward network $\mathcal{F}$ takes as input the concatenation of the mid-circuit readout vector from the previous timestep and the current input $(\mathbf{z}_{t-1}{:}\,\mathbf{x}_t)$, and outputs $\boldsymbol{\theta}_t \in \mathbb{R}^{16}$ containing the 16 rotation angles that parametrize the PQC shown in (a) as $U(\boldsymbol{\theta}_t)$ (usually referred to as *angle encoding*),[2] which acts on the quantum state propagated from the previous step. ↗ denotes qubit expectation-value readouts.

by construction, which naturally preserve norms (§3.1). Acting on n qubits, they enable expressive transformations over quantum states in an exponentially large Hilbert space $\mathbb{C}^{2^n}$. Although such spaces are classically intractable beyond moderate n, they can be manipulated with only n qubits on quantum hardware.

In this work, we present a hybrid quantum-classical QRNN in which the recurrent core is realized as a PQC controlled by a classical feedforward network. The hidden state is the quantum state of the n-qubit PQC, residing in an exponentially large Hilbert space $\mathbb{C}^{2^n}$, which provides a high-capacity coherent quantum memory. Nonlinearity is introduced through classical computation rather than approximated within the quantum circuit, leaving the PQC strictly for unitary evolution of the hidden state. Fig. 1 illustrates both the PQC (with four qubits shown for illustration) and the unrolled QRNN:

- At timestep t, the input is mapped to a learnable representation $\mathbf{x}_t$ via an embedding layer.
- A classical feedforward network $\mathcal{F}$ takes as input the concatenation of $\mathbf{z}_{t-1}$ (readout outputs at timestep $t-1$) and the current input $\mathbf{x}_t$. It outputs the PQC parameters $\boldsymbol{\theta}_t$ which configure the PQC with a fixed gate layout (Fig. 1a), denoted $U(\boldsymbol{\theta}_t)$, applied at timestep t (Fig. 1c).

[1] All $\mathbf{RX}$ gates are controlled rotations that apply only when the connected control qubit is in the $|1\rangle$ state: $\mathbf{CRX}(\theta_i) = |0\rangle\langle 0| \otimes I \; + \; |1\rangle\langle 1| \otimes \mathbf{RX}(\theta_i)$.

[2] $U(\boldsymbol{\theta}_t)$ is a single unitary for encoding the outputs of $\mathcal{F}$ and transforming the state.

- The PQC applies the parametrized unitary gates to evolve the quantum state, yielding the updated state. Residing in an exponentially large Hilbert space, this state persists across timesteps and provides the model's core recurrent memory.
- The mid-circuit readout $\mathbf{z}_t$ (or the final readout at the end of the sequence) is a real-valued vector obtained from the quantum state via Pauli expectation-value readouts and is used: (i) as recurrent feedback $\mathbf{z}_{t-1}$ at timestep t, and (ii) as the input to task-specific classical layers.

We develop the models on GPUs, allowing us to simulate and train quantum recurrence via classical backpropagation, with the expectation that such models will become classically unsimulatable as the number of qubits increases. To our knowledge, this is the first model (RNN or otherwise) grounded in quantum operations demonstrated in classical simulation with up to 14 qubits across six realistic sequence-modeling tasks, achieving competitive performance with LSTM and the scaled Cayley orthogonal scoRNN designed for norm preservation (Helfrich et al., 2018). Experiments also show that classical nonlinear control and feedback are effective, with nonlinear variants outperforming their linear counterparts, and that the unitary quantum recurrent core maintains more stable gradients than LSTMs on sequences of up to 400 tokens (§4.6).

The QRNN is motivated in part by the memory and gradient problems of RNNs, but its main aim is to explore a hybrid quantum–classical recurrent model in an idealized proof of principle that allows us to study its computational behavior under best-case conditions across a broad class of sequence learning tasks. The PQC (Sim et al., 2019) uses only one- and two-qubit gates without nonstandard operations and is feasible on quantum hardware (Xu et al., 2024). The overall architecture provides a hardware-aware base case and a plausible path toward future hardware implementations, which would require fault-tolerant devices capable of sustaining long coherent recurrences and real-time classical control for realistic-scale sequence modeling.

Another way to view the QRNN is via fast and slow weights in RNNs, which function as different types of memory across multiple timescales (Schmidhuber, 1992; Ba et al., 2016). The PQC parameters serve as the short-term memory, analogous to the hidden activities of classical RNNs, and are controlled and reconfigured at each timestep by a classical feedforward network whose slow weights encode the long-term memory. The quantum state, updated via unitary transformations, evolves on a faster timescale than the slow weights, persists across timesteps, and acts as a third, higher-capacity memory in the Hilbert space, retaining information that influences subsequent computation (Hinton and Plaut, 1987; Schmidhuber, 1993).

2 Related Work

Bausch (2020) proposes a QRNN whose recurrence is implemented by iterating a quantum cell built from "quantum neurons". Nonlinearity is induced *inside* the circuit via measurement-and-postselection primitives (repeat-until-success), together with amplitude amplification to make the procedure near-deterministic. Consequently, the per-timestep update is not strictly unitary and introduces back-action on the memory. Because the inherent linear dynamics of PQCs, the effective nonlinear maps achievable by such schemes are structurally constrained, and the repertoire of admissible nonlinearities remains limited (Yan et al., 2020; Moreira et al., 2023; Zi et al., 2024).

The so-called QLSTMs embed PQCs into the gating mechanisms of classical LSTMs (Chen et al., 2020; Yu et al., 2024a; Ubale et al., 2025), replacing dense layers in the LSTM gates with PQCs. However, all memory and recurrence remain entirely classical, governed by standard hidden and cell state updates. These architectures are best viewed as classical LSTMs augmented with auxiliary PQCs, rather than quantum recurrent models.

Li et al. (2023); Siemaszko et al. (2023) also model recurrences with PQCs and support per-timestep readouts, but they rely entirely on linear quantum dynamics without other nonlinearity or classical control. Nikoloska et al. (2023) stacked a PQC on top of a classical RNN, which modulates whether a quantum unitary is applied or skipped.[3]

Experiments with the existing models have focused on domain-specific tasks such as fraud detection (Ubale et al., 2025), low-resource text classification (Yu et al., 2024a), or scaled-down

[3]We experimented with a similar stacking architecture concurrently [link to code].

MNIST (Bausch, 2020; Siemaszko et al., 2023). We instead present the first QRNN to demonstrate competitive performance across six full-scale sequence modeling tasks.

3 MODEL

3.1 PQC

Unitary evolution. A PQC typically starts from the all-zero state $|\psi\rangle = |0\rangle^{\otimes n} \in \mathbb{C}^{2^n}$ and applies a series of gates arranged from left to right.[4] An example PQC with $n = 4$ qubits is shown in Fig. 1a, where each horizontal line represents a qubit. The square boxes denote quantum gates, which by definition are unitary transformations acting on one or more qubits. Single-qubit gates apply local transformations, while multi-qubit gates can generate superposition and entanglement.[5]

Let U denote the composition (product) of a collection of unitary gates, hence $U^\dagger U = I$. For any state $|\psi\rangle$,

$$\|U|\psi\rangle\|^2 = \langle\psi|U^\dagger U|\psi\rangle = \langle\psi|I|\psi\rangle = \langle\psi|\psi\rangle = \|\psi\|^2,$$

which ensures norm preservation by construction.[6]

Parametrized unitary gates. In a PQC, gates can be either fixed or parametrized. Fixed gates implement structural operations and remain constant throughout training,[7] while the latter contain learnable parameters, which function like trainable weight matrices analogous to neural-network "layers". The PQC in Fig. 1a consists of entirely parametrized gates.

Readouts via Pauli expectations. A succinct way to summarize a quantum state is via *expectation values* of Pauli observables. For each qubit k, the Pauli operators X_k, Y_k, Z_k are the three fundamental single-qubit observables that act on that qubit (with identity on all other qubits). For example, the Pauli-Z operator measures the *computational basis* with

$$Z \;=\; \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

while X and Y measure superposition states in orthogonal bases. For a single qubit $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, the Pauli-$Z$ expectation is $\langle Z\rangle = \langle\psi|Z|\psi\rangle = |\alpha|^2 - |\beta|^2 \in [-1, 1]$. For an n-qubit state, per-qubit expectations $\{\langle X_k\rangle, \langle Y_k\rangle, \langle Z_k\rangle\}_{k=1}^n$ and correlation terms (*Pauli strings*) $\langle P_1 \otimes \cdots \otimes P_n\rangle$ with $P_\ell \in \{I, X, Y, Z\}$ yield bounded, hardware-agnostic real-valued summaries useful for analysis and downstream modeling.

Although expectation-value readouts are nonlinear functions of parametrized rotation angles (e.g., in $\mathbf{RX}$ gates), the resulting nonlinearity is generally insufficient on its own (§4).

3.2 HYBRID MODEL

RNNs parametrize a conditional distribution with a function that depends on a hidden state $\mathbf{h}_{t-1}$, which compacts past inputs $(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1})$ into a fixed-dimensional representation:

$$p(\mathbf{x}_t \mid \mathbf{x}_1, \ldots, \mathbf{x}_{t-1}) \;\approx\; p(\mathbf{x}_t \mid \mathbf{h}_{t-1}).$$

At each timestep t, the hidden state $\mathbf{h}_t$ is updated based on the previous hidden state $\mathbf{h}_{t-1}$ and the current input $\mathbf{x}_t$:

$$\mathbf{h}_t \;=\; f\big(\mathbf{h}_{t-1}, \mathbf{x}_t; \boldsymbol{\Theta}\big),$$

where f is a transformation (e.g., a basic RNN or LSTM cell) parametrized by $\boldsymbol{\Theta}$. In the hybrid model (Fig. 1c), we replace the hidden state with a quantum state represented by the PQC in Fig. 1a, which is controlled by a classical feedforward network and evolved by applying the unitary gates.

[4] $\otimes$ denotes the tensor product.

[5] See Appendix D for a basic description of qubits and superposition.

[6] $U^\dagger$ denotes the conjugate transpose (Hermitian adjoint) of U. Formally, if the PQC consists of L gates, $U = u_L u_{L-1} \cdots u_1$, where each u_i is a unitary operator acting on some subset of qubits, then $U^\dagger = u_1^\dagger u_2^\dagger \cdots u_L^\dagger$, and hence $U^\dagger U = I$.

[7] For example, the $\mathbf{CNOT}$ gate flips the target qubit if the control is in the $|1\rangle$ state.

Let $\mathbf{x}_t$ be the input embedding at timestep t, and let $\mathbf{z}_{t-1}$ be the readout vector from the previous timestep. In the most generic form of the hybrid model,[8] the two are concatenated into a single vector $\mathbf{u}_t = (\mathbf{z}_{t-1}; \mathbf{x}_t)$ and passed through a classical feedforward network $\mathcal{F}$ with one hidden layer and a nonlinearity.

The first transformation in $\mathcal{F}$ maps the input $\mathbf{u}_t$ to a hidden representation $\mathbf{v}_t$:

$$\mathbf{v}_t = \phi(\mathbf{W}_1\mathbf{u}_t + \mathbf{b}_1), \tag{1}$$

where ϕ is a nonlinear activation function. The second transformation maps $\mathbf{v}_t$ to

$$\boldsymbol{\theta}_t = \mathbf{W}_2\mathbf{v}_t + \mathbf{b}_2, \tag{2}$$

where $\boldsymbol{\theta}_t \in \mathbb{R}^d$ represents the parameters that control the PQC's unitary operations at timestep t. Each element of $\boldsymbol{\theta}_t$ denoted θ_i is mapped to a rotation angle in a parametrized quantum gate within the PQC (e.g., $1 \leq i \leq d$ and $d = 16$ in Fig. 1a).

The PQC itself is defined by a unitary operator $U(\boldsymbol{\theta}_t)$ parametrized by $\boldsymbol{\theta}_t$.[9] Applying $U(\boldsymbol{\theta}_t)$ to the prior state $\mathbf{h}_{t-1} = |\psi_{t-1}\rangle$ yields the updated state

$$\mathbf{h}_t = U(\boldsymbol{\theta}_t)\,|\psi_{t-1}\rangle.$$

A classical readout vector is then computed from Pauli expectation values:

$$\mathbf{z}_t = \mathrm{Readout}(\mathbf{h}_t) = \big(\,\langle X_1\rangle_t,\ \langle Y_1\rangle_t,\ \langle Z_1\rangle_t,\ \ldots,\ \langle X_n\rangle_t,\ \langle Y_n\rangle_t,\ \langle Z_n\rangle_t\,\big), \tag{3}$$

where $\langle P_k\rangle_t = \langle\psi_t|P_k|\psi_t\rangle$ for $P_k \in \{X_k, Y_k, Z_k\}$ and $|\psi_t\rangle = \mathbf{h}_t$.[10]

Serving as a proxy for the quantum state, $\mathbf{z}_t$ is concatenated with the next input $\mathbf{x}_{t+1}$ to produce $\boldsymbol{\theta}_{t+1}$ via the classical controller. Because the Pauli expectation-value readouts do not alter the state, coherence of the quantum state is preserved across timesteps, allowing it to function as a coherent recurrent memory.

We train the entire hybrid model end-to-end using classical backpropagation, optimizing the parameters $\boldsymbol{\Theta} = \{\mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2\}$ via standard optimizers, such as Adam (Kingma and Ba, 2014). For sequence-to-sequence learning, $\mathbf{z}_t$ provides per-timestep outputs and serves as contextual embeddings for soft attention decoding.

4 Experiments

We use the ansatz shown in Fig.1a (scaled to more qubits when required) as the core circuit for the QRNN. Sim et al. (2019) demonstrate experimentally that this ansatz is expressive, capable of generating strong entanglement, and able to represent a significant portion of the Hilbert space, even compared to deeper circuits built from less expressive ansätze.[11] We implement and simulate the model using `TorchQuantum` (Wang et al., 2022), which remains less optimized than classical toolkits due to the lack of efficient kernels for hybrid operations involving tight classical–quantum feedback, particularly in recurrent settings. Our ansatz balances expressivity, implementation simplicity, and simulation efficiency.

For the feedforward network $\mathcal{F}$ (Eq. 1 and Eq. 2), we experimented with ReLU, LeakyReLU, GLU and GELU nonlinearities.[12] For both language modeling and translation, we first transform the readouts with a separate feedforward layer and use the result both for vocabulary classification and as input to the next timestep.

All experiments are run on a single A100/A30 GPU and we select the best models on the validation split across different random seeds and report the test results. The per-epoch training runtime ranges

[8] Extra transformations may be applied to the readouts before classifications or feeding them to the next step (for some tasks); see §4.

[9] Here $U(\boldsymbol{\theta}_t)$ denotes the product of the circuit's parametrized gates, each acting on one or more qubits with its parameter drawn from $\boldsymbol{\theta}_t$.

[10] For computational efficiency, we use only single-qubit expectations rather than multi-qubit Pauli strings.

[11] See Appendix E for details on the PQC design and expressibility evaluation methodology.

[12] GLU requires projecting to twice the output dimensionality, effectively increasing the parameter count compared to standard nonlinearities like ReLU, when all other dimensions are held constant.

Table 1: Classification accuracy on IMDB. Qubit count q, total readouts m; or hidden state size h (for RNN, LSTM and scoRNN only); embedding dimension e; parameter count p. † indicates the LSTM in Dai and Le (2015).

Model	**Val**	**Test**	$q_m \vee h$	e	p
$\text{QRNN}_{\text{ReLU}}$	87.25	85.37	8_{24}	100	5K
$\text{QRNN}_{\text{LeakyReLU}}$	87.41	**87.00**	8_{24}	100	5K
$\text{QRNN}_{\text{GELU}}$	87.53	86.38	8_{24}	100	5K
$\text{QRNN}_{\text{Linear}}$	85.37	84.21	8_{24}	100	5K
$\text{QRNN}_{\text{Linear}}$	84.21	83.22	4_{12}	100	2K
RNN	87.64	86.96	50	50	5K
LSTM	88.40	86.79	25	25	5.2K
LSTM†	–	86.5	1,024	512	6.2M
scoRNN	87.25	86.48	50	50	5K

from ~4 minutes for MNIST (with 10 qubits) to ~36 minutes for language modeling (with 14 qubits). Hyperparameters are tuned on the validation splits, the shared hyperparameters across all the tasks include the Adam optimizer without learning rate decay ($lr = 1 \times 10^{-3},\ \lambda = 1 \times 10^{-4}$, and $\epsilon = 1 \times 10^{-10}$) and dropout applied to the input at each step, with task-dependent drop rates. We apply full-sequence backpropagation without truncation, except for language modeling, where sequences are truncated to 35 tokens. No pretrained word embeddings are used. Additional hyperparameters and test set statistics (mean, min, max across runs) are provided in Appendix C. For scoRNN, we use a hidden size of 170 and the hyperparameters from Helfrich et al. (2018) are used throughout.

4.1 Sentiment Analysis

The IMDB sentiment dataset (Maas et al., 2011) is a balanced binary classification benchmark with 25K labeled reviews each for training and testing. The average review length is 241 tokens, with a maximum length of 2,500 tokens. We use 7.5K reviews from the training set for validation and truncate all reviews to a maximum length of 400 tokens across all models.

The hybrid model for this task follows the generic hybrid architecture described in §3.2. At the final input token, we apply an affine transformation to the readouts to produce two logits, which are used for classification via cross-entropy. Table 1 summarizes the results. $\text{QRNN}_{\text{LeakyReLU}}$ achieves the highest test accuracy. Ablating the classical nonlinearity (Eq. 1) degrades performance, though increasing the number of qubits in the linear model still yields some accuracy gains. Adding the nonlinearity results in a substantial improvement, outperforming all baselines. On this task, the orthogonal scoRNN underperforms other models, despite having a larger hidden state and over five times more parameters.

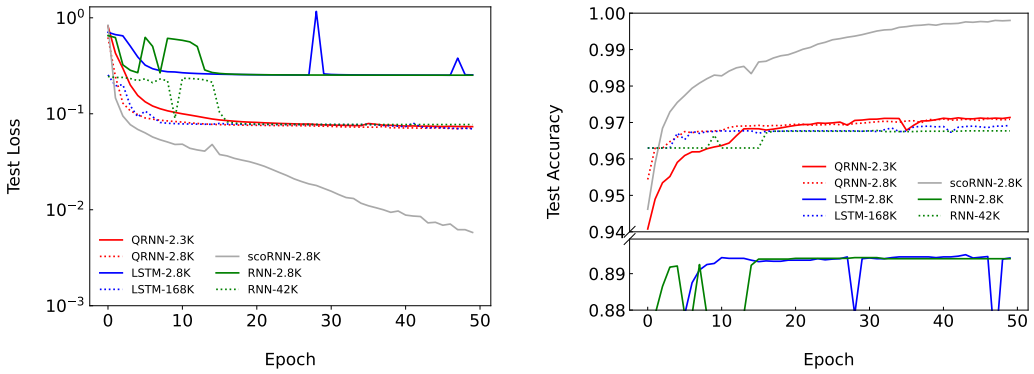
4.2 MNIST and Permuted MNIST

We report results on the full MNIST dataset without downsampling using the same model as for IMDB, except with 10 output classes instead of binary classification. The standard pixel-by-pixel permuted MNIST (pMNIST) setup (Le et al., 2015; Arjovsky et al., 2016) requires 784 steps to process each 28×28 digit, which makes simulation prohibitively slow. Here we permute the pixels of each digit first, which are then reshaped back to 28×28. In both the standard and permuted cases, we use the same hyperparameters.

Table 2 shows that QRNNs with three different types of nonlinearity outperform the classical baselines on both tasks, clearly demonstrating the benefit of adding classical nonlinearities compared to the $\text{QRNN}_{\text{Linear}}$ models. We observe that permutation leads to an accuracy drop across all models: 2.45% for $\text{QRNN}_{\text{GELU}}$, 3.00% for the RNN, 3.51% for the LSTM, and 1.51% for scoRNN, which achieves comparable performance to $\text{QRNN}_{\text{GELU}}$.

Table 2: Classification accuracy on MNIST and pMNIST. Qubit count q, total readouts m; or hidden state size h (for RNN, LSTM and scoRNN only); embedding dimension e; parameter count p. † indicates the QRNN model of (Bausch, 2020) with 13 qubits and each digit downscaled to 4×4 and binarized.

	MNIST		pMNIST				
Model	**Val**	**Test**	**Val**	**Test**	$q_m \vee h$	e	p
$\text{QRNN}_{\text{ReLU}}$	98.10	97.83	94.86	95.05	10_{30}	28	3.9K
$\text{QRNN}_{\text{LeakyReLU}}$	98.01	97.96	95.13	94.86	10_{30}	28	3.9K
$\text{QRNN}_{\text{GELU}}$	98.17	**98.03**	95.38	**95.58**	10_{30}	28	3.9K
$\text{QRNN}_{\text{Linear}}$	97.06	96.80	94.94	94.13	10_{30}	28	3.9K
$\text{QRNN}_{\text{Linear}}$	94.31	93.87	91.10	90.55	5_{15}	28	1.3K
QRNN†	–	96.70	–	–	$q = 13$	1	3.1K
RNN	97.42	97.28	95.16	94.28	50	28	3.9K
LSTM	97.61	97.44	94.92	93.93	20	28	4K
scoRNN	96.68	95.62	94.50	92.37	50	28	3.9K



(a) (b)

Figure 2: Test loss (a) and accuracy (b) for copying memory with $T = 200$.

4.3 Copying Memory

The copying memory problem tests a model's ability to retain and recall information over long sequences (Hochreiter and Schmidhuber, 1997; Arjovsky et al., 2016). Each input sequence has $T + 20$ tokens, where the first $k = 10$ are random digits from 1 to 8 (n_{classes}), followed by zeros, and the last 11 ($k + 1$) positions are filled with the digit '9' with the first '9' acting as a delimiter. The model must learn to detect the delimiter and recall the original digits right after it in the output sequence. We randomly generated 5K training and 1K test samples with $T = 200$ (for training efficiency of QRNNs). A random guess baseline yields a loss of $\frac{k \cdot \log(n_{\text{classes}}-1)}{T+2k} \approx 0.095$, reflecting the expected cross-entropy when choosing uniformly from incorrect digits. On this task, QRNN-2.3K matches LSTM-168K (loss 0.07, accuracy 97%) and outperforms LSTM-2.8K (loss 0.25, accuracy 89.4%). scoRNN, specialized for this task, achieves near-perfect results, highlighting a performance gap between general-purpose and tailored models.

4.4 Word-Level Language Modeling

The PTB dataset (Mikolov et al., 2011) consists of 929K training tokens, 73K validation tokens, and 82K test tokens. As is standard, we use a vocabulary size of 10K, converting `OOV` tokens to `UNK`. We

Table 3: PTB word-level language modeling (PPL). Qubit count q, total readouts m; or hidden state size h (for RNN and LSTM only); embedding dimension e; parameter count p.

Model	**Val**	**Test**	$q_m \vee h$	e	p
$\text{QRNN}_{\text{ReLU}}$	131.81	126.69	14_{42}	512	131K
$\text{QRNN}_{\text{LeakyReLU}}$	131.41	126.58	14_{42}	512	131K
$\text{QRNN}_{\text{GELU}}$	136.62	131.07	14_{42}	512	131K
$\text{QRNN}_{\text{LeakyReLU}}$	135.00	130.35	10_{30}	512	78K
$\text{QRNN}_{\text{LeakyReLU}}$	169.17	161.09	5_{15}	512	39K
RNN	151.96	139.13	256	256	131K
LSTM	124.22	**120.30**	128	128	132K

Table 4: Multi30K German-to-English translation (BLEU). Qubit count q, total readouts m; or hidden state size h (for RNN and LSTM only); embedding dimension e; parameter count p.

Model	**Val**	**Test**	$q_m \vee h$	e	p
QRNN_{GLU}	31.08	31.92	13_{39}	512	412K
$\text{QRNN}_{\text{LeakyReLU}}$	29.22	28.99	13_{39}	512	359K
$\text{QRNN}_{\text{GELU}}$	29.95	29.14	13_{39}	512	359K
QRNN_{GLU}	30.16	31.51	10_{30}	512	378K
QRNN_{GLU}	27.63	29.66	5_{15}	512	320K
RNN	29.50	29.50	512	293	412K
LSTM	32.10	32.00	256	145	412K
scoRNN	32.70	**33.60**	512	293	412K

tested scoRNN on this task, but it did not converge to a good solution. The LSTM achieved the best result, with 120.30 perplexity (PPL), followed closely by $\text{QRNN}_{\text{LeakyReLU}}$ at 126.58.

4.5 Machine Translation

Soft attentions can be implemented using various formulations, such as additive attention or dot-product attention (Luong et al., 2015), but they share the same core principle: at each decoder timestep, compute a similarity score between the current decoder state and each encoder state, normalize these scores via a softmax, and form a context vector by summation, which is then combined with the decoder's hidden state to generate the next output token.

The attention mechanism implemented here follows the additive attention of Bahdanau et al. (2015). At each decoding step, the decoder hidden state is concatenated with encoder outputs, passed through a tanh activation followed by a linear projection to compute alignment scores. A softmax then normalizes these scores into attention weights, with masking applied to exclude padded positions.

We applied the model to Multi30k German-to-English translation (Elliott et al., 2016), with vocabulary sizes of 19.2K for German and 10.8K for English, and an average of 11 tokens per sentence in both languages. The training set contains 29K sentence pairs, with 1K each for validation and testing.

Results in Table 4 show that QRNN_{GLU} with 13 qubits closely matches the LSTM, followed by QRNN_{GLU} with 10 qubits. For the QRNN, it is somewhat surprising that intermediate readouts can still support mechanisms like soft attention, since these readouts capture only partial projections of the quantum state rather than the full hidden state. This suggests that, despite mid-circuit readouts, sufficient information is retained and propagated across timesteps. We qualitatively interpret the learned soft alignments on a few examples where the translations required non-trivial linguistic interpretations in Appendix A.

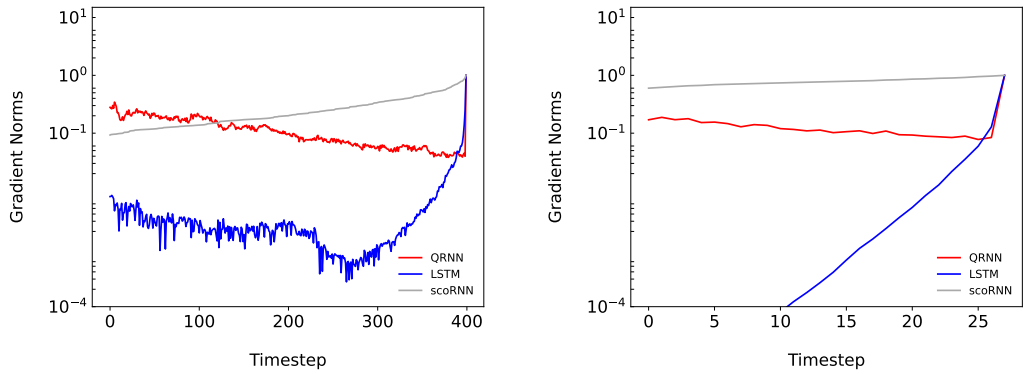


Figure 3: Normalized per-timestep gradient norms $\|\partial\mathcal{L}/\partial\mathbf{h}_t\|_2$, averaged over one mini-batch containing samples of identical T (batch size $= 16$). Curves are normalized by the final timestep ($t = T$) gradient to compare decay shape; higher gradient values closer to $T = 0$ indicate less vanishing. (a) IMDB, $T = 400$. (b) pMNIST, $T = 28$.

4.6 Hidden State Gradients and Visualization

We measure per-timestep gradient norms on IMDB ($T = 400$) and pMNIST ($T = 28$) by retaining gradients on the per-timestep readouts (QRNN) and hidden states (LSTM) from saved checkpoints and computing $\|\partial\mathcal{L}/\partial\mathbf{h}_t\|_2$. Gradients are averaged across samples in a mini-batch and normalized by the last-step norm $\|\partial\mathcal{L}/\partial\mathbf{h}_T\|_2$ to compare decay shape.

As shown in Fig. 3, the QRNN curves remain consistently above the LSTM on both IMDB and pMNIST, indicating less vanishing through time toward the start of the sequences. All curves start with 1.0 at $t = T$ (normalization), but the relative elevation of the QRNN curve at earlier timesteps demonstrates more stable gradient propagation. The LSTM gradient norm decays rapidly, collapsing below 10^{-4} on the relatively short pMNIST sequences.

We visualize the per-timestep state trajectories for a single pMNIST test example in Appendix B.

5 Discussion and Conclusion

Different quantum hardware platforms currently require distinct control stacks, and architectural choices do not translate one-to-one across devices, with factors such as native gate sets, qubit connectivity, and the implementation of mid-circuit readout all affecting the realization of a given circuit. The aim here is not to prescribe a hardware roadmap but to analyze a hardware-realistic base case under idealized classical simulation to study the empirical properties of the architecture, where we model mid-circuit observations via expectation-value readouts.

As more efficient and scalable toolchains become available (e.g., future multi-GPU toolkits based on `cuQuantum` (Bayraktar et al., 2023)), we anticipate more faithful simulations via ancilla-mediated schemes in which auxiliary qubits are entangled with the main circuit, measured, and reset as needed while the recurrent memory remains coherent. This aligns with mid-circuit measure-and-reset operations already supported on several platforms (DeCross et al., 2022; Lis et al., 2023; Norcia et al., 2023).

This paper bridges quantum operations and recurrent learning by introducing a new hybrid QRNN whose recurrent core is implemented as a PQC steered by a classical controller. The unitary dynamics preserve norms, promoting stable gradient propagation; mid-circuit, per-timestep readouts inject task adaptability; and the classical controller supplies the nonlinearity and feedback for expressiveness. As techniques improve (Abbas et al., 2023) and quantum hardware matures, the architecture provides a path toward hardware-realistic quantum models for sequential learning.

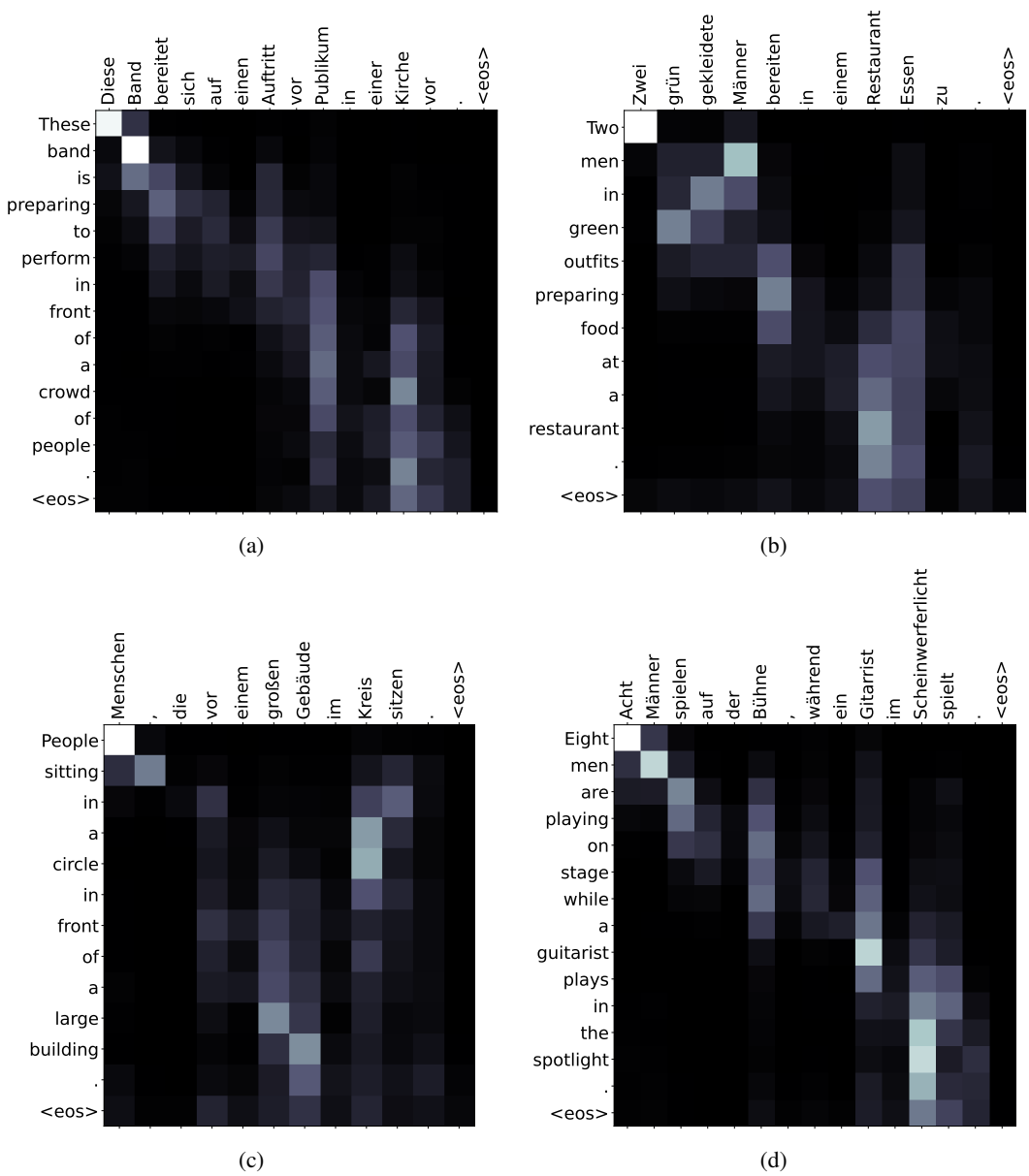


Figure 4: Soft attention alignments produced by the QRNN encoder-decoder model.

A Attention Alignments

To qualitatively analyze the model's learned soft attention alignments we selected four sentences from test set and interpreted the hybrid model translations and alignments (Fig. 4).

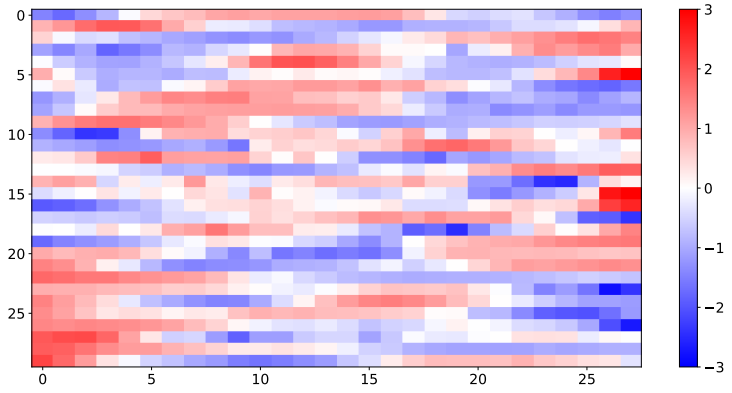
We observe that the hybrid model can manage spatial and syntactic shifts while capturing clause-level structure and semantics through its readout-based hidden states and soft attention as well as the LSTM baseline. It is evident that the model handles **compound verb constructions** and **semantic expansion**, in sentences like *"Diese Band bereitet sich auf einen Auftritt vor Publikum in einer Kirche vor"* (Fig. 4a) and *"Zwei grün gekleidete Männer bereiten in einem Restaurant Essen zu"* (Fig. 4b), where German separable verbs—*"bereitet ... vor"* and *"bereiten ... zu"*—are correctly reconstructed into the English verb phrases *"is preparing to perform"* and *"preparing"*, respectively. The soft attention allowed the model to attend across non-contiguous source tokens, enabling reassembly of

verb phrases. Additionally, lexical expansions such as *"Publikum"* → *"a crowd of people"* (Fig. 4a) and *"gekleidete Männer"* → *"men in green outfits"* (Fig. 4b) demonstrate contextually appropriate semantic elaboration beyond literal translation.

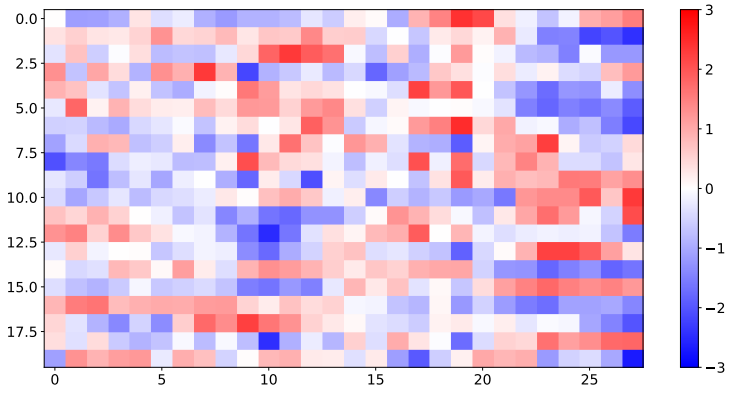
The model also displays **syntactic reordering** and **clause realignment**, necessitated by divergences between German and English word order. This is shown in both *"Diese Band ... vor Publikum ... vor"* and (Fig. 4a) *"Menschen, die vor einem großen Gebäude im Kreis sitzen"* (Fig. 4c). In the former, German's verb-final structure is reorganized into a mid-sentence English verb phrase, while handling nested prepositional phrases. In the latter, the relative clause *"die ... sitzen"* is compressed into the participial phrase *"sitting"*, dropping auxiliaries and pronouns to better fit English syntactic norms. Similarly, the location and positional phrases *"im Kreis"* and *"vor einem großen Gebäude"* are reordered into *"in a circle in front of a large building"*

Lastly, for **multi-clause coordination**, **tense adaptation**, and **long-range dependency tracking**, as seen in *"Acht Männer spielen auf der Bühne, während ein Gitarrist im Scheinwerferlicht spielt"* (Fig. 4d). The model successfully disentangles two coordinated clauses and renders them with the correct English conjunction *"while"*, while adjusting verb forms from German's uniform *"spielen"* to *"are playing"* and *"plays"*, based on subject plurality. Finally, this ability to flexibly adapt clause boundaries and maintain coherence is also reflected in the *"Menschen ... im Kreis sitzen"* example (Fig. 4c), where the model tracks relative clause dependencies and maps them onto compact English constructions.

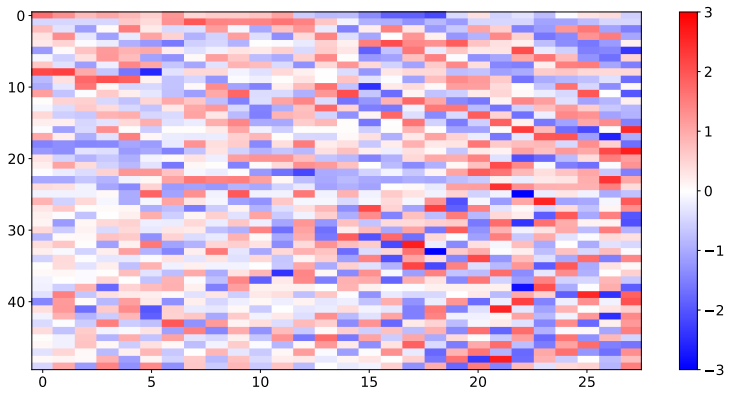
B Hidden State Visualization



(a) QRNN.



(b) LSTM.



(c) scoRNN.

The above plots are obtained on a pMNIST sample. All models are those obtained the highest accuracies on the test set. For comparability across models with different dynamic ranges, each unit's trajectory is z-scored over time (per unit) and displayed using the same color scale (clipped to $[-3, 3]$). Qualitatively, the QRNN exhibits more coherent, banded temporal structure (units with sustained positive/negative excursions across multiple timesteps), whereas the LSTM and scoRNN activations appear more spatially fragmented and locally varying.

C Experimental Settings and Test Accuracy Statistics Across Runs

For RNN language modeling, we use the code from `https://github.com/yandex/faster-rnnlm`, and hyperparameters following the results and recommended settings in the documentation as below:

```
faster-rnnlm/rnnlm -rnnlm rnn.sig.256.1 -train ~/data/lm/ptb.train.txt
--valid ~/data/lm/ptb.valid.txt --hidden 256 --hidden-type sigmoid
--nce 50 --alpha 0.1 --bptt-skip 8 --bptt 35
```

For LSTM language modeling, we use the code from Zaremba et al. (2015). The hyperparameters are listed below:

```
batch_size=64, dropout=0.5, factor=1.0, factor_epoch=6, hidden_size=128,
layer_num=1, learning_rate=0.001, lstm_type='pytorch', max_grad_norm=5,
opt='adam', seq_length=35, total_epochs=49, winit=0.05
```

Table 5: QRNN Hyperparameters: batch size b, dropout rate d; embedding initialization e_{init}.

Task	b	d	e_{init}	max_grad_norm
IMDB	200	0.25	Xavier Uniform	5
MNIST	200	–	–	5
Copying (LeakyReLU)	50	–	–	5
PTB	64	0.5	Xavier Uniform	5
Multi30K	64	0.25	Xavier Uniform	7

Table 6: RNN and LSTM Hyperparameters: batch size b, embedding dropout rate d_e; final output hidden state dropout rate d_o; embedding initialization e_{init}. The Adam hyperparameters are the same as the QRNN.

Task	b	d_e	d_o	e_{init}	max_grad_norm
IMDB	100	0.2	0.2	$\mathcal{N}(0, 1)$	5
MNIST	200	–	0.2	–	1
Copying	50	–	–	–	–
Multi30K	64	0.25	–	Xavier Uniform	5

Table 7: scoRNN Hyperparameters: batch size b, embedding dropout rate d_e; final output hidden state dropout rate d_o; embedding initialization e_{init}. On MT, we use an internal PyTorch implementation of scoRNN which uses the same optimizer settings as our other models as described in the experimental settings. On other tasks, the hyperparameters are from the scoRNN paper's code including the optimizer hyperparamters.

Task	b	d_e	n_{neg_ones}	e_{init}
IMDB	64	0.2	$h/10$	Uniform $(-1, 1)$
MNIST	50	–	$h/10$, $h/2$ (permuted)	–
Copying	50	–	$h/2$	–
Multi30K	64	0.5	$h/10$	Xavier Uniform

Table 8: Accuracy statistics on MNIST and pMNIST test sets across 50 runs for each nonlinearity variant. Qubit count q and total readouts m; embedding dimension e; parameter count p. scoRNN is reported over 10 runs.

	MNIST			**pMNIST**					
Model	min	max	μ	min	max	μ	$q_m \vee h$	e	p
$\text{QRNN}_{\text{ReLU}}$	97.51	98.25	97.84	94.33	95.31	94.83	10_{30}	28	3.9K
$\text{QRNN}_{\text{LeakyReLU}}$	97.42	98.15	97.88	94.33	95.38	94.80	10_{30}	28	3.9K
$\text{QRNN}_{\text{GELU}}$	97.62	98.22	97.96	94.72	95.58	95.12	10_{30}	28	3.9K
RNN	96.36	97.32	96.92	93.66	94.53	94.11	50	28	3.9K
LSTM	96.62	97.63	97.23	93.35	94.09	93.74	20	28	4K
scoRNN	94.82	95.92	95.26	92.15	92.83	92.42	50	28	3.9K

Table 9: PPL on PTB test set across 5 runs for each nonlinearity variant. Qubit count q, total readouts m; embedding dimension e; parameter count p. With the `faster-rnnlm` toolkit for RNN, we obtained the same results on multiple runs as Table 3.

Model	min	max	μ	$q_m \vee h$	e	p
$\text{QRNN}_{\text{LeakyReLU}}$	126.53	128.82	127.77	14_{42}	512	131K
LSTM	119.83	121.99	120.57	128	128	132K

Table 10: BLEU evaluations on the Multi30K German to English test set across 20 runs for each nonlinearity variant. Qubit count q, total readouts m; embedding dimension e; parameter count p.

Model	min	max	μ	$q_m \vee h$	e	p
QRNN_{GLU}	19.83	31.92	27.88	13_{39}	512	412K
$\text{QRNN}_{\text{LeakyReLU}}$	24.52	29.87	28.55	13_{39}	512	359K
$\text{QRNN}_{\text{GELU}}$	25.71	30.29	29.09	13_{39}	512	359K
RNN	27.60	29.90	29.16	512	293	412K
LSTM	32.00	33.10	32.42	256	145	412K
scoRNN	29.50	33.60	32.34	256	293	412K

Table 11: Accuracy statistics on IMDB test set across 100 runs for each nonlinearity variant. Qubit count q, total readouts m; embedding dimension e; parameter count p. RNN and LSTM are reported over 50 runs. scoRNN is reported over 5 runs.

Model	min	max	μ	min^*	μ^*	$q_m \vee h$	e	p
$\text{QRNN}_{\text{ReLU}}$	49.55	85.96	71.18	71.74	83.11	8_{24}	100	5K
$\text{QRNN}_{\text{LeakyReLU}}$	49.63	87.00	70.23	75.77	83.44	8_{24}	100	5K
$\text{QRNN}_{\text{GELU}}$	49.98	86.38	77.18	70.39	83.75	8_{24}	100	5K
RNN	85.24	87.19	–	–	86.39	50	50	5K
LSTM	86.28	88.00	–	–	86.99	25	25	5.2K
scoRNN	85.57	86.48	–	–	86.19	50	50	5K

Among all tasks, IMDB showed the greatest variability in QRNN performance across random seeds in development. This behavior may align with known sensitivities in training variational PQCs (Grant et al., 2019). We therefore also report stats where we remove failed runs ($< 70\%$ accuracy, well below simple baselines such as BoW), indicated by $*$. For the three nonlinearities 40, 42 and 25 failed runs were observed each. The results also indicate that GELU nonlinearity reduces the sensitivity compared with the other two.

While parametrized quantum circuits (PQCs) can suffer from vanishing gradients in deep or wide settings due to the barren plateau phenomenon (McClean et al., 2018), there is no general impossibility theorem that barren plateaus must occur in all parametrized quantum circuits; their presence and severity are known to depend on the ansatz, cost function, initialization, training strategy, and noise, and remain an empirical matter at practical scales. Several studies provide insights into how it arises or design principles that prevent or mitigate plateaus (Cerezo et al., 2019; Grant et al., 2019; Patti et al., 2021; Sack et al., 2022). These results indicate that barren plateaus are not inevitable, and that careful design yields a tractable and stable training landscape in practice. In particular, some architectures such as quantum convolutional neural networks avoid barren plateaus by construction (Pesah et al., 2021), which supports the view that appropriate architectural choices can produce stable and trainable quantum models.

D Quantum States and Superposition

Unlike a classical bit, a qubit exists in a *superposition* of the states 0 and 1 in a two-dimensional complex Hilbert space: $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle = [\alpha \quad \beta]^T \in \mathbb{C}^2$ and $|0\rangle = [1 \quad 0]^T$ and $|1\rangle = [0 \quad 1]^T$ are elements of the *computational basis* for the Hilbert space. The coefficients α and β are complex numbers referred to as the *amplitudes* that satisfy $|\alpha|^2 + |\beta|^2 = 1$. For a state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, the probability of obtaining $|0\rangle$ is $|\alpha|^2$, and the probability of obtaining $|1\rangle$ is $|\beta|^2$.

E PQC TEMPLATE

We have chosen the PQC template based on the benchmarking study in Sim et al. (2019), which evaluates 19 different parametrized quantum circuits (PQCs) up to depth 5 (i.e., the base circuit repeated up to five times and used a single PQC). Each PQC is assessed using two key metrics: expressibility and entangling capability. The architecture referred to as ansatz-14 in Sim et al. (2019) which we use here in a single layer configuration was shown to score highly on both. This gives a good balance of simulation cost and "goodness" of the PQC.

Expressibility is quantified by comparing the distribution of pairwise fidelities between states generated by the PQC to the theoretical fidelity distribution of Haar-random states, which represent uniform randomness over the composite Hilbert space (the tensor product of individual qubit spaces). Instead of generating Haar-random states directly, the method in (Sim et al., 2019) uses the analytical form of the Haar fidelity distribution as a reference. PQC output states are obtained by sampling random parameters, and their pairwise fidelities are used to construct an empirical distribution. The KL divergence between this empirical distribution and the Haar reference provides a scalar expressibility score, with lower values indicating greater expressiveness.

There has also been recent work on Quantum Deep Equilibrium Models (QDEQs), which use deep equilibrium ideas to train PQC-based quantum models with effectively lower circuit depth and fewer parameters (Schleich et al., 2024). Orthogonal to the above, this suggests a promising path to shallower circuits while maintaining or enhancing expressivity.

F COPYING TASK RESULTS SUMMARY

Table 12: Copying task summary results of Fig. 2 for models with the same number of parameters.

Model	**Acc.**	**Loss**	$q_m \vee h$	e	p
$\text{QRNN}_{\text{LeakyReLU}}$	97.09	0.0706	8_{24}	10	2.8K
RNN	89.41	0.2524	48	10	2.8K
LSTM	89.43	0.2543	22	10	2.8K
LSTM	96.92	0.0697	200	10	169.6K
scoRNN	**99.80**	**0.0058**	48	10	2.8K

G WALL-CLOCK TRAINING TIME COMPARISON

Table 13: Per-epoch wall-clock training time for QRNN and LSTM, using the best-performing model configuration for each task.

Task	**QRNN ($\sim$ mins)**	**LSTM ($\sim$ secs)**
IMDB	8	10
MNIST	4	7
Copying	9	1
PTB	36	5
Multi30K	31	38

REFERENCES

Amira Abbas, Robbie King, Hsin-Yuan Huang, William J. Huggins, Ramis Movassagh, Dar Gilboa, and Jarrod R. McClean. On quantum backpropagation, information reuse, and cheating

measurement collapse. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8c3caae2f725c8e2a55ecd600563d172-Abstract-Conference.html.

Rajeev Acharya, Dmitry A. Abanin, Laleh Aghababaie-Beni, Igor Aleiner, Trond I. Andersen, Markus Ansmann, Frank Arute, Kunal Arya, Abraham Asfaw, Nikita Astrakhantsev, Juan Atalaya, Ryan Babbush, Dave Bacon, Brian Ballard, Joseph C. Bardin, Johannes Bausch, Andreas Bengtsson, Alexander Bilmes, Sam Blackwell, Sergio Boixo, Gina Bortoli, Alexandre Bourassa, Jenna Bovaird, Leon Brill, Michael Broughton, David A. Browne, Brett Buchea, Bob B. Buckley, David A. Buell, Tim Burger, Brian Burkett, Nicholas Bushnell, Anthony Cabrera, Juan Campero, Hung-Shen Chang, Yu Chen, Zijun Chen, Ben Chiaro, Desmond Chik, Charina Chou, Jahan Claes, Agnetta Y. Cleland, Josh Cogan, Roberto Collins, Paul Conner, William Courtney, Alexander L. Crook, Ben Curtin, Sayan Das, Alex Davies, Laura De Lorenzo, Dripto M. Debroy, Sean Demura, Michel Devoret, Agustin Di Paolo, Paul Donohoe, Ilya Drozdov, Andrew Dunsworth, Clint Earle, Thomas Edlich, Alec Eickbusch, Aviv Moshe Elbag, Mahmoud Elzouka, Catherine Erickson, Lara Faoro, Edward Farhi, Vinicius S. Ferreira, Leslie Flores Burgos, Ebrahim Forati, Austin G. Fowler, Brooks Foxen, Suhas Ganjam, Gonzalo Garcia, Robert Gasca, Élie Genois, William Giang, Craig Gidney, Dar Gilboa, Raja Gosula, Alejandro Grajales Dau, Dietrich Graumann, Alex Greene, Jonathan A. Gross, Steve Habegger, John Hall, Michael C. Hamilton, Monica Hansen, Matthew P. Harrigan, Sean D. Harrington, Francisco J. H. Heras, Stephen Heslin, Paula Heu, Oscar Higgott, Gordon Hill, Jeremy Hilton, George Holland, Sabrina Hong, Hsin-Yuan Huang, Ashley Huff, William J. Huggins, Lev B. Ioffe, Sergei V. Isakov, Justin Iveland, Evan Jeffrey, Zhang Jiang, Cody Jones, Stephen Jordan, Chaitali Joshi, Pavol Juhas, Dvir Kafri, Hui Kang, Amir H. Karamlou, Kostyantyn Kechedzhi, Julian Kelly, Trupti Khaire, Tanuj Khattar, Mostafa Khezri, Seon Kim, Paul V. Klimov, Andrey R. Klots, Bryce Kobrin, Pushmeet Kohli, Alexander N. Korotkov, Fedor Kostritsa, Robin Kothari, Borislav Kozlovskii, John Mark Kreikebaum, Vladislav D. Kurilovich, Nathan Lacroix, David Landhuis, Tiano Lange-Dei, Brandon W. Langley, Pavel Laptev, Kim-Ming Lau, Loïck Le Guevel, Justin Ledford, Joonho Lee, Kenny Lee, Yuri D. Lensky, Shannon Leon, Brian J. Lester, Wing Yan Li, Yin Li, Alexander T. Lill, Wayne Liu, William P. Livingston, Aditya Locharla, Erik Lucero, Daniel Lundahl, Aaron Lunt, Sid Madhuk, Fionn D. Malone, Ashley Maloney, Salvatore Mandrà, James Manyika, Leigh S. Martin, Orion Martin, Steven Martin, Cameron Maxfield, Jarrod R. McClean, Matt McEwen, Seneca Meeks, Anthony Megrant, Xiao Mi, Kevin C. Miao, Amanda Mieszala, Reza Molavi, Sebastian Molina, Shirin Montazeri, Alexis Morvan, Ramis Movassagh, Wojciech Mruczkiewicz, Ofer Naaman, Matthew Neeley, Charles Neill, Ani Nersisyan, Hartmut Neven, Michael Newman, Jiun How Ng, Anthony Nguyen, Murray Nguyen, Chia-Hung Ni, Murphy Yuezhen Niu, Thomas E. O'Brien, William D. Oliver, Alex Opremcak, Kristoffer Ottosson, Andre Petukhov, Alex Pizzuto, John Platt, Rebecca Potter, Orion Pritchard, Leonid P. Pryadko, Chris Quintana, Ganesh Ramachandran, Matthew J. Reagor, John Redding, David M. Rhodes, Gabrielle Roberts, Eliott Rosenberg, Emma Rosenfeld, Pedram Roushan, Nicholas C. Rubin, Negar Saei, Daniel Sank, Kannan Sankaragomathi, Kevin J. Satzinger, Henry F. Schurkus, Christopher Schuster, Andrew W. Senior, Michael J. Shearn, Aaron Shorter, Noah Shutty, Vladimir Shvarts, Shraddha Singh, Volodymyr Sivak, Jindra Skruzny, Spencer Small, Vadim Smelyanskiy, W. Clarke Smith, Rolando D. Somma, Sofia Springer, George Sterling, Doug Strain, Jordan Suchard, Aaron Szasz, Alex Sztein, Douglas Thor, Alfredo Torres, M. Mert Torunbalci, Abeer Vaishnav, Justin Vargas, Sergey Vdovichev, Guifre Vidal, Benjamin Villalonga, Catherine Vollgraff Heidweiller, Steven Waltman, Shannon X. Wang, Brayden Ware, Kate Weber, Travis Weidel, Theodore White, Kristi Wong, Bryan W. K. Woo, Cheng Xing, Z. Jamie Yao, Ping Yeh, Bicheng Ying, Juhwan Yoo, Noureldin Yosri, Grayson Young, Adam Zalcman, Yaxing Zhang, Ningfeng Zhu, and Nicholas Zobrist. Quantum error correction below the surface code threshold. *Nature*, 638(8052):920–926, December 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-08449-y. URL http://dx.doi.org/10.1038/s41586-024-08449-y.

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *ICML'16: Proceedings of the 33rd International Conference on International Conference on Machine Learning*, volume 48, page 1120–1128, 2016.

Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando Brandao, David Buell, Brian Burkett, Yu Chen, Jimmy Chen, Ben Chiaro, Roberto Collins, William Courtney, Andrew Dunsworth, Edward Farhi, Brooks Foxen, Austin Fowler, Craig Michael Gidney, Marissa Giustina, Rob Graff, Keith Guerin, Steve Habegger,

Matthew Harrigan, Michael Hartmann, Alan Ho, Markus Rudolf Hoffmann, Trent Huang, Travis Humble, Sergei Isakov, Evan Jeffrey, Zhang Jiang, Dvir Kafri, Kostyantyn Kechedzhi, Julian Kelly, Paul Klimov, Sergey Knysh, Alexander Korotkov, Fedor Kostritsa, Dave Landhuis, Mike Lindmark, Erik Lucero, Dmitry Lyakh, Salvatore Mandrà, Jarrod Ryan McClean, Matthew McEwen, Anthony Megrant, Xiao Mi, Kristel Michielsen, Masoud Mohseni, Josh Mutus, Ofer Naaman, Matthew Neeley, Charles Neill, Murphy Yuezhen Niu, Eric Ostby, Andre Petukhov, John Platt, Chris Quintana, Eleanor G. Rieffel, Pedram Roushan, Nicholas Rubin, Daniel Sank, Kevin J. Satzinger, Vadim Smelyanskiy, Kevin Jeffery Sung, Matt Trevithick, Amit Vainsencher, Benjamin Villalonga, Ted White, Z. Jamie Yao, Ping Yeh, Adam Zalcman, Hartmut Neven, and John Martinis. Quantum supremacy using a programmable superconducting processor. *Nature*, 574:505–510, 2019. URL https://www.nature.com/articles/s41586-019-1666-5.

Jimmy Ba, Geoffrey Hinton, Volodymyr Mnih, Joel Z Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1409.0473.

Johannes Bausch. Recurrent quantum neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 1368–1379. Curran Associates, Inc., 2020.

Harun Bayraktar, Ali Charara, David Clark, Saul Cohen, Timothy Costa, Yao-Lung L. Fang, Yang Gao, Jack Guan, John Gunnels, Azzam Haidar, Andreas Hehn, Markus Hohnerbach, Matthew Jones, Tom Lubowe, Dmitry Lyakh, Shinya Morino, Paul Springer, Sam Stanwyck, Igor Terentyev, Satya Varadhan, Jonathan Wong, and Takuma Yamaguchi. cuquantum sdk: A high-performance library for accelerating quantum science, 2023. URL https://arxiv.org/abs/2308.01999.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. *Advances in Neural Information Processing Systems*, 37:107547–107603, 2024.

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

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

Satwik Bhattamishra, Michael Hahn, Phil Blunsom, and Varun Kanade. Separations in the representational capabilities of transformers and recurrent architectures. *Advances in Neural Information Processing Systems*, 37:36002–36045, 2024.

Roberto Bondesan and Max Welling. Quantum deformed neural networks, 2020.

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, 2019. doi: 10.1038/s41467-021-21728-w. URL https://doi.org/10.1038/s41467-021-21728-w.

Samuel Yen-Chi Chen, Shinjae Yoo, and Yao-Lung L. Fang. Quantum long short-term memory. *arXiv preprint arXiv:2009.01783*, 2020.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1179.

Jasmine Collins, Jascha Sohl-Dickstein, and David Sussillo. Capacity and trainability in recurrent neural networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=BydARw9ex.

Andrew M. Dai and Quoc V. Le. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems, NIPS*, 2015.

M. DeCross, R. Haghshenas, M. Liu, E. Rinaldi, J. Gray, Y. Alexeev, C.H. Baldwin, J.P. Bartolotta, M. Bohn, E. Chertkov, J. Cline, J. Colina, D. DelVento, J.M. Dreiling, C. Foltz, J.P. Gaebler, T.M. Gatterman, C.N. Gilbreth, J. Giles, D. Gresh, A. Hall, A. Hankin, A. Hansen, N. Hewitt, I. Hoffman, C. Holliman, R.B. Hutson, T. Jacobs, J. Johansen, P.J. Lee, E. Lehman, D. Lucchetti, D. Lykov, I.S. Madjarov, B. Mathewson, K. Mayer, M. Mills, P. Niroula, J.M. Pino, C. Roman, M. Schecter, P.E. Siegfried, B.G. Tiemann, C. Volin, J. Walker, R. Shaydulin, M. Pistoia, S.A. Moses, D. Hayes, B. Neyenhuis, R.P. Stutz, and M. Foss-Feig. Computational power of random quantum circuits in arbitrary geometries. *Physical Review X*, 15(2), May 2025. ISSN 2160-3308. doi: 10.1103/physrevx.15.021052. URL http://dx.doi.org/10.1103/PhysRevX.15.021052.

Matthew DeCross, Eli Chertkov, Megan Kohagen, and Michael Foss-Feig. Qubit-reuse compilation with mid-circuit measurement and reset, 2022. URL https://arxiv.org/abs/2210.08039.

Y Du, MH Hsieh, T Liu, and D Tao. The expressive power of parameterized quantum circuits. arxiv 2018. *arXiv preprint arXiv:1810.11922*, 2019.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3210. URL https://aclanthology.org/W16-3210/.

Edward Grant, Leonard Wossnig, Mateusz Ostaszewski, and Marcello Benedetti. An initialization strategy for addressing barren plateaus in parametrized quantum circuits. *Quantum*, 3: 214, December 2019. ISSN 2521-327X. doi: 10.22331/q-2019-12-09-214. URL http://dx.doi.org/10.22331/q-2019-12-09-214.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Kyle Helfrich, Devin Willmott, and Qiang Ye. Orthogonal recurrent neural networks with scaled Cayley transform. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1969–1978. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/helfrich18a.html.

Geoffrey E Hinton and David C Plaut. Using fast weights to deblur old memories. In *Proceedings of the ninth annual conference of the Cognitive Science Society*, pages 177–186, 1987.

Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Li Jing, Çağlar Gülçehre, John Peurifoy, Yichen Shen, Max Tegmark, Marin Soljačić, and Yoshua Bengio. Gated orthogonal recurrent units: On learning to forget. *Neural Computation*, 31(4): 765–783, 2019. doi: 10.1162/neco_a_01174. URL https://doi.org/10.1162/neco_a_01174.

Bobak Kiani, Randall Balestriero, Yann LeCun, and Seth Lloyd. projunn: Efficient method for training deep networks with unitary matrices. In *Advances in Neural Information Processing Systems*, volume 35, pages 14448–14463, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Quoc V Le, Navdeep Jaitly, and Geoffrey E Hinton. A simple way to initialize recurrent networks of rectified linear units. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1133–1141, 2015.

Yanan Li, Zhimin Wang, Rongbing Han, Shangshang Shi, Jiaxin Li, Ruimin Shang, Haiyong Zheng, Guoqiang Zhong, and Yongjian Gu. Quantum recurrent neural networks for sequential learning. *Neural Networks*, 166:148–161, 2023.

Joanna W. Lis, Aruku Senoo, William F. McGrew, Felix Rönchen, Alec Jenkins, and Adam M. Kaufman. Mid-circuit operations using the omg-architecture in neutral atom arrays, 2023. URL https://arxiv.org/abs/2305.19266.

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1166. URL https://aclanthology.org/D15-1166/.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015/.

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, 2018. doi: 10.1038/s41467-018-07090-4. URL https://www.nature.com/articles/s41467-018-07090-4.

Tomas Mikolov. *Statistical Language Models Based on Neural Networks*. PhD thesis, Brno University of Technology, 2012. URL https://www.fit.vutbr.cz/research/view_pub.php?id=9848.

Tomas Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. Empirical evaluation and combination of advanced language modeling techniques. In *Interspeech*, 2011.

MS Moreira, Gian Giacomo Guerreschi, Wouter Vlothuizen, Jorge F Marques, Jeroen van Straten, Shavindra P Premaratne, Xiang Zou, Hany Ali, Nandini Muthusubramanian, Christos Zachariadis, et al. Realization of a quantum neural network using repeat-until-success circuits in a superconducting quantum processor. *npj Quantum Information*, 9(1):118, 2023.

Ivana Nikoloska, Osvaldo Simeone, Leonardo Banchi, and Petar Veličković. Time-warping invariant quantum recurrent neural networks via quantum-classical adaptive gating. *Mach.Learn.Sci.Tech.*, 4(4), 2023. doi: 10.1088/2632-2153/acff39. URL https://www.osti.gov/biblio/1923309.

M.A. Norcia, W.B. Cairncross, K. Barnes, P. Battaglino, A. Brown, M.O. Brown, K. Cassella, C.-A. Chen, R. Coxe, D. Crow, J. Epstein, C. Griger, A.M.W. Jones, H. Kim, J.M. Kindem, J. King, S.S. Kondov, K. Kotru, J. Lauigan, M. Li, M. Lu, E. Megidish, J. Marjanovic, M. McDonald, T. Mittiga, J.A. Muniz, S. Narayanaswami, C. Nishiguchi, R. Notermans, T. Paule, K.A. Pawlak, L.S. Peng, A. Ryou, A. Smull, D. Stack, M. Stone, A. Sucich, M. Urbanek, R.J.M. van de Veerdonk, Z. Vendeiro, T. Wilkason, T.-Y. Wu, X. Xie, X. Zhang, and B.J. Bloom. Midcircuit qubit measurement and rearrangement in a yb 171 atomic array. *Physical Review X*, 13(4), November 2023. ISSN 2160-3308. doi: 10.1103/physrevx.13.041034. URL http://dx.doi.org/10.1103/PhysRevX.13.041034.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 26670–26698. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/orvieto23a.html.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318, 2013.

Taylor L. Patti, Khadijeh Najafi, Xun Gao, and Susanne F. Yelin. Entanglement devised barren plateau mitigation. *Physical Review Research*, 3(3), July 2021. ISSN 2643-1564. doi: 10.1103/physrevresearch.3.033090. URL http://dx.doi.org/10.1103/PhysRevResearch.3.033090.

Adrián Pérez-Salinas, David López-Núñez, Artur García-Sáez, Pol Forn-Díaz, and José I Latorre. One qubit as a universal approximant. *Physical Review A*, 104(1):012405, 2021.

Arthur Pesah, M Cerezo, Samson Wang, Andrew T Sornborger, Lukasz Cincio, and Patrick J Coles. Absence of barren plateaus in quantum convolutional neural networks. *Physical Review X*, 11 (4):041011, 2021. doi: 10.1103/PhysRevX.11.041011. URL https://doi.org/10.1103/PhysRevX.11.041011.

Ben W. Reichardt, David Aasen, Rui Chao, Alex Chernoguzov, Wim van Dam, John P. Gaebler, Dan Gresh, Dominic Lucchetti, Michael Mills, Steven A. Moses, Brian Neyenhuis, Adam Paetznick, Andres Paz, Peter E. Siegfried, Marcus P. da Silva, Krysta M. Svore, Zhenghan Wang, and Matt Zanner. Demonstration of quantum computation and error correction with a tesseract code. *arXiv preprint arXiv:2409.04628*, 2024. URL https://arxiv.org/abs/2409.04628.

Stefan H. Sack, Raimel A. Medina, Alexios A. Michailidis, Richard Kueng, and Maksym Serbyn. Avoiding barren plateaus using classical shadows. *PRX Quantum*, 3(2), June 2022. ISSN 2691-3399. doi: 10.1103/prxquantum.3.020365. URL http://dx.doi.org/10.1103/PRXQuantum.3.020365.

Philipp Schleich, Marta Skreta, Lasse B. Kristensen, Rodrigo A. Vargas-Hernández, and Alán Aspuru-Guzik. Quantum deep equilibrium models, 2024. URL https://arxiv.org/abs/2410.23940.

Jürgen Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992.

Jürgen Schmidhuber. Reducing the ratio between learning complexity and number of time varying variables in fully recurrent nets. In *International Conference on Artificial Neural Networks*, pages 460–463. Springer, 1993.

Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3):032430, 2021.

Michał Siemaszko, Adam Buraczewski, Bertrand Le Saux, and Magdalena Stobińska. Rapid training of quantum recurrent neural networks. *Quantum Information Processing*, 22(1):1–15, 2023.

Sukin Sim, Peter D. Johnson, and Alan Aspuru-Guzik. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Advanced Quantum Technologies*, 2(12), 2019.

Rushikesh Ubale, Sujan K. K., Sangram Deshpande, and Gregory T. Byrd. Toward practical quantum machine learning: A novel hybrid quantum lstm for fraud detection, 2025. URL https://arxiv.org/abs/2505.00137.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. URL https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Hanrui Wang, Yongshan Ding, Jiaqi Gu, Zirui Li, Yujun Lin, David Z Pan, Frederic T Chong, and Song Han. QuantumNAS: Noise-adaptive search for robust quantum circuits. In *The 28th IEEE International Symposium on High-Performance Computer Architecture (HPCA-28)*, 2022.

Wenduan Xu, Stephen Clark, Douglas Brown, Gabriel Matos, and Konstantinos Meichanetzidis. Quantum recurrent architectures for text classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, November 2024. URL https://aclanthology.org/2024.emnlp-main.1000/.

Shilu Yan, Hongsheng Qi, and Wei Cui. Nonlinear quantum neuron: A fundamental building block for quantum neural networks. *Physical Review A*, 102(5):052421, 2020.

Wenbin Yu, Lei Yin, Chengjun Zhang, Yadang Chen, and Alex X. Liu. Application of quantum recurrent neural network in low-resource language text classification. *IEEE Transactions on Quantum Engineering*, 5:1–13, 2024a. doi: 10.1109/TQE.2024.3373903.

Zhan Yu, Qiuhao Chen, Yuling Jiao, Yinan Li, Xiliang Lu, Xin Wang, and Jerry Yang. Non-asymptotic approximation error bounds of parameterized quantum circuits. *Advances in Neural Information Processing Systems*, 37:99089–99127, 2024b.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization, 2015. URL https://arxiv.org/abs/1409.2329.

Wei Zi, Siyi Wang, Hyunji Kim, Xiaoming Sun, Anupam Chattopadhyay, and Patrick Rebentrost. Efficient quantum circuits for machine learning activation functions including constant t-depth relu. *Phys. Rev. Res.*, 6:043048, Oct 2024. doi: 10.1103/PhysRevResearch.6.043048. URL https://link.aps.org/doi/10.1103/PhysRevResearch.6.043048.