HYBRID QUANTUM-CLASSICAL RECURRENT NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a new hybrid quantum-classical recurrent neural network (RNN) architecture in which the recurrent core is realized as a parametrized quantum circuit (PQC) controlled by a nonlinear classical feedforward network. The hidden state is the quantum state of the PQC, residing in an exponentially large Hilbert space $\mathbb{C}^{2^n}$ and manipulable using only n qubits. The PQC is unitary by construction, making the hidden-state evolution inherently norm preserving without external constraints. To evolve the recurrence, classical embeddings of the current input are combined with mid-circuit readouts from the previous timestep's quantum state and processed by a feedforward network. The resulting outputs parameterize the PQC, which then evolves unitarily to produce the updated hidden state. This enables per-timestep readouts while avoiding attempts to emulate nonlinearities with inherently linear quantum dynamics. We evaluate the model in simulation with up to 14 qubits on sentiment analysis, MNIST, permuted MNIST, copying memory, and language modeling, adopting projective measurements as a limiting case to obtain mid-circuit readouts while maintaining a coherent quantum memory across timesteps. We also devise a soft attention mechanism over readouts in a sequence-to-sequence model and show that the network is effective for machine translation. To our knowledge, this is the first model (RNN or otherwise) grounded in quantum operations to achieve superior or 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. To address this, various strategies have been proposed (Mikolov, 2012; Pascanu et al., 2013; Le et al., 2015). In particular, 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. These models perform well on synthetic memory tasks, but their results on broader benchmarks vary.

The introduction of the Transformer model (Vaswani et al., 2017) appeared to relegate recurrent architectures by bypassing the hidden-state bottleneck. Yet there is now renewed interest, with recent work demonstrating that recurrent inductive bias remains highly competitive and provides representational advantages that cannot be matched by the Transformer (Gu and Dao, 2023; Orvieto et al., 2023; Bhattamishra et al., 2024; Beck et al., 2024).

With the advent of quantum computing (Arute et al., 2019; Bondesan and Welling, 2020; Pan et al., 2023; Reichardt et al., 2024), including quantum backpropgation (Abbas et al., 2023), parametrized

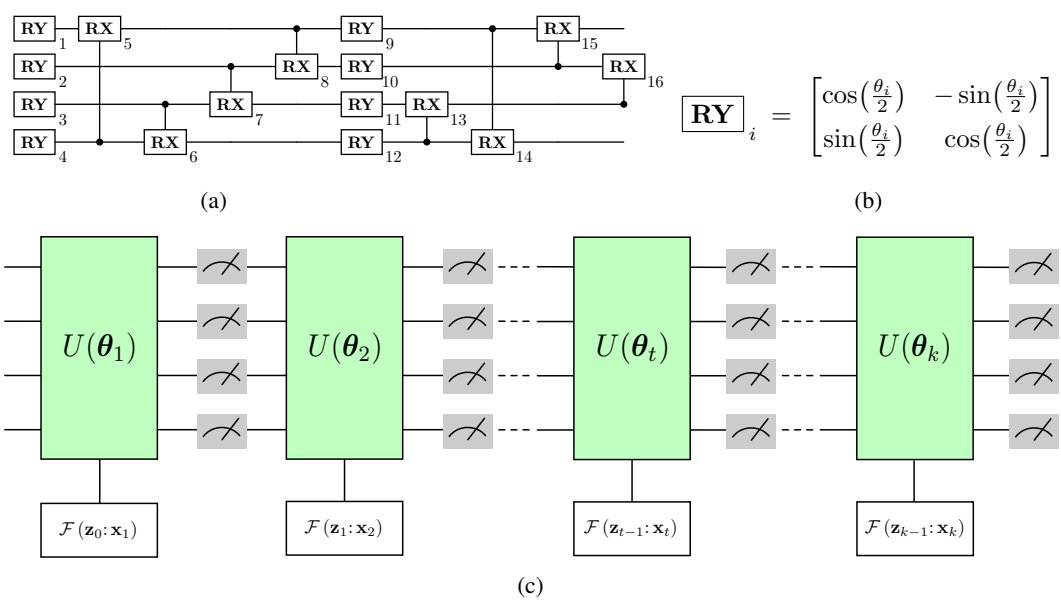


Figure 1: Hybrid QRNN architecture. (a) Recurrent core PQC with $\mathbf{n} = 4$ qubits (illustrative) with 16 parametrized gates and a Hilbert space $\mathbb{C}^{2^\mathbf{n}}$. (b) Example $\mathbf{RY}$ gate; each gate (either $\mathbf{RX}$ or $\mathbf{RY}$) is parametrized by a rotation angle θ_i, $1 \leq i \leq 16$. (c) QRNN unrolled for a sequence of length k, where $\nearrow$ denote qubit measurements. At each step t, $1 \leq t \leq k$, outputs from all measurements at step $t-1$ are aggregated into the readout vector $\mathbf{z}_{t-1}$, which is concatenated with the input embedding $\mathbf{x}_t$. The feedforward network $\mathcal{F}$ takes the combined vector $(\mathbf{z}_{t-1}{:}\,\mathbf{x}_t)$ and produces $|\boldsymbol{\theta}_t| = 16$ parameters per timestep that control the PQC structure from (a), denoted $U(\boldsymbol{\theta}_t)$.

quantum circuits (PQCs), which are a core component of variational hybrid quantum classical models, have emerged as an alternative mechanism for structured function approximation (Du et al., 2019; Schuld et al., 2021; Pérez-Salinas et al., 2021; Yu et al., 2024b). PQCs implement unitary transformations by construction, which naturally preserve norms (§3.1). Acting on n qubits, they evolve quantum states within a Hilbert space in $\mathbb{C}^{2^n}$, enabling expressive transformations over exponentially large states. Although such spaces are classically intractable beyond moderate n, they can be manipulated with only n physical qubits on quantum hardware.

In this work, we present a new hybrid quantum–classical RNN (QRNN) grounded in quantum operations, where the entire recurrent core is realized as a PQC. We introduce a mechanism for evolving the quantum hidden state while maintaining coherence across timesteps via mid-circuit measurements, as a limiting case simulated via projective measurements, enabling the state to remain coherent across time without collapsing and allowing per-timestep readouts. We use a classical feedforward network to dynamically parametrize the PQC at each step, which provides a principled way to couple classical nonlinearity with mid-circuit measurements of a unitary quantum recurrent core, avoiding attempts to emulate nonlinearities with intrinsically linear quantum operations, leaving the PQC strictly for coherent unitary evolution in an exponentially large Hilbert space.

Fig. 1 illustrates both the PQC (with four qubits shown for illustration) and the unrolled QRNN architecture:

- The input at step t is mapped to a classical embedding $\mathbf{x_t}$ via a learnable embedding layer.
- A classical feedforward network $\mathcal{F}$ takes as input the concatenation of the previous readout $\mathbf{z}_{t-1}$ and the current input $\mathbf{x}_t$. It outputs the circuit parameters $\boldsymbol{\theta}_t$. These parameters configure a PQC with a fixed gate layout (Fig. 1a), denoted $U(\boldsymbol{\theta}_t)$, which is applied at timestep t. For instance, the PQC in Fig. 1a comprises 16 parametrized gates (the square $\mathbf{RX}$ and $\mathbf{RY}$ boxes).[1] Accordingly, the feedforward output vector at step t is $\boldsymbol{\theta}_t \in \mathbb{R}^{16}$,

[1]More precisely, each $\mathbf{RX}$ gate is a controlled $\mathbf{RX}$ gate, activated by a connected qubit.

$1 \leq t \leq k$, (Fig. 1c), with each element of $\boldsymbol{\theta}_t$, denoted θ_i, $i \in [1, 16]$, assigned to one gate of the PQC at step t, specifying its rotation angle (Fig. 1b).

- The quantum state encoded by the PQC evolves unitarily through applications of the parametrized unitary gates, yielding the updated state. Residing in an exponentially large Hilbert space, this state provides the core recurrent memory of the model, which persists across timesteps.
- Mid-circuit readouts $\mathbf{z}_t$ (or final readouts at the end of the sequence) are applied to extract classical features from the hidden state. These readouts, obtained via measurements of the quantum state, serve two purposes: (i) as recurrent feedback $\mathbf{z}_{t-1}$ at step t; and (ii) as the input to task-specific classical layers, e.g., for classification.

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 rooted in quantum operations (RNN or otherwise) demonstrated to achieve superior or competitive performance with up to 14 qubits in classical simulation across a set of six realistic sequence modeling tasks. Our results show QRNN outperforms scoRNN, a classical orthogonal RNN specifically designed for norm preservation, on five of six tasks, and outperforms LSTM on four of six, while achieving competitive results on language modeling and machine translation. Experiments also show that classical non-linear control and feedback is effective, with the non-linear models outperforming their linear counterparts, and we find that the unitary quantum recurrent core maintains more stable gradients than LSTMs (§4.6).

Our architecture 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 elementary one- and two-qubit gates that should be supported on any hardware platform. It replaces conventional recurrence with expressive unitary transformations that are physically grounded. The model performs competitively in simulation, providing a hardware-aware base case and a plausible path toward future hardware implementations.

2 Related Work

Bausch (Bausch, 2020) developed a QRNN based on a quantum neuron construction. These quantum neurons are composed into layers, forming a QRNN with persistent quantum memory. However, the nonlinearities are emulated within PQCs and incur significant overhead due to probabilistic repeated-until-success circuit execution and postselection steps. This overhead comes from the fact that quantum computation is inherently linear, and attempts to emulate nonlinearities within PQCs often require substantial complexity, and the available forms of quantum nonlinearity remain 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) and Siemaszko et al. (2023) also model recurrences with PQCs, while supporting per-timestep outputs, but they rely entirely on linear quantum dynamics of the PQC without explicit nonlinearities or classical control.

Experiments of the existing models have primarily focused on domain-specific evaluations such as fraud detection (Ubale et al., 2025), low-resource text classification (Yu et al., 2024a), or scaled-down MNIST (Bausch, 2020; Siemaszko et al., 2023). We instead present the first QRNN to demonstrate competitive performance across six full-scale sequence modeling tasks.

Another way to interpret our hybrid model is through the lens of fast and slow weights in RNNs, which provides a mechanism for memory across different timescales (Schmidhuber, 1992; Ba et al., 2016). The PQC functions as fast weights, controlled and reconfigured at each step by a classical

feedforward network that plays the role of slow weights. The quantum state evolves under this control and persists across timesteps under unitary evolution.

3 MODEL

3.1 PQC

Unitary evolution. A PQC is shown in Fig. 1a, where each horizontal line represents a qubit. The square boxes denote quantum gates, which by definition are always unitary transformations acting on one or more qubits. Single-qubit gates apply local transformations, while multi-qubit gates can generate superposition and entanglement between qubits.[2] A typical PQC consists of entirely unitary operations U acting on quantum states $|\psi\rangle \in \mathbb{C}^{2^n}$, with

$$U^\dagger U = I \quad \Rightarrow \quad \||U|\psi\rangle\|_2 = \||\psi\rangle\|_2,$$

ensuring norm preservation by construction.[3]

Parametrized unitary gates. The unitary gates on a PQC can also serve as learnable transformations acting on a quantum state and the sequence of gates or unitary matrices is analogous to "layers" in a neural network, where parametrized gates are defined by unitary matrices with classically adjustable parameters. The parametric gates in Fig. 1a are the $\mathbf{RX}$ and $\mathbf{RY}$ *rotation* gates,[4] with the unitary matrices

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

The rotation angles above, when used in a PQC, can be provided and controlled by external inputs, allowing the quantum computation to incorporate classical data.

Measurements. A PQC typically starts from an initial quantum state, often initialized as all zero, and applies a series of gates arranged from left to right. To intercept the quantum state, we perform measurements to obtain real-valued readouts (in the case of our hybrid model, without collapsing the state). These readouts provide partial observations of the state, and any required number of measurements, on any of the qubits, can be combined for downstream tasks. For instance, a measurement through the Pauli-$\mathbf{Z}$ observable with the unitary

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

assigns scalar values (e.g., $+1$ for $|0\rangle$ and -1 for $|1\rangle$) in the computational basis in the single-qubit case. For a general quantum state however, the outcome is probabilistic: it yields $+1$ with probability $|\alpha|^2$ and -1 with probability $|\beta|^2$. The expectation value of this measurement is given by $|\alpha|^2 - |\beta|^2$, which can be used as a real-valued readout in hybrid quantum-classical models.

Although the outputs obtained via measurement are a nonlinear function of the gate parameters, particularly those used in parametrized rotation gates such as $\mathbf{RX}$, it is typically a weak form of nonlinearity (§4).

3.2 HYBRID MODEL

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

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

[2] See Appendix A for a basic description of qubits and superposition.

[3] $U^\dagger$ denotes the conjugate transpose (Hermitian adjoint) of U.

[4] All $\mathbf{RX}$ gates only activate conditionally if the connected control qubit in the PQC is in the $|1\rangle$ state (Fig. 1a).

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 $\mathbf{x}_t$:

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

where f is a transformation (e.g., a basic RNN or LSTM cell) parametrized by $\mathbf{\Theta}$. 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.

Let $\mathbf{x}_t$ be the input embedding at timestep t, and let $\mathbf{z}_{t-1}$ be the measurement-based readout from the previous timestep. In the most generic form of the hybrid model,[5] these values are combined 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 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 non-linear 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., $1 \le i \le d$ and $d = 16$ in Fig. 1a).

The PQC itself is defined by a unitary operator $U(\boldsymbol{\theta}_t)$, parametrized by $\boldsymbol{\theta}_t$.[6] After the gates in $U(\boldsymbol{\theta}_t)$ are applied to the quantum state $\mathbf{h}_{t-1} = |\psi_{t-1}\rangle$, the resulting state $\mathbf{h}_t = U(\boldsymbol{\theta}_t)\,|\psi_{t-1}\rangle$ is measured to obtain classical readouts

$$\mathbf{z}_t = \textbf{Measure}(\mathbf{h}_t), \tag{3}$$

which serve as a proxy for the quantum state and are combined with the next input $\mathbf{x}_{t+1}$ to evolve the recurrence. To preserve coherence across timesteps, we simulate mid-circuit measurements, allowing recurrent structure without collapsing the full quantum state, retaining the quantum memory throughout the sequence.

We train the entire hybrid model end-to-end using classical backpropagation, optimizing the parameters $\mathbf{\Theta} = \{\mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2\}$ via standard optimizers, such as Adam (Kingma and Ba, 2014). Because each $\mathbf{z}_t$ is real-valued, it can be used both as a per-timestep output and as a contextual embedding for soft attention in sequence-to-sequence 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.[7] 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 $\mathbf{Measure}$ in Eq. 3, measurements are performed in each of the Pauli-$\mathbf{X}$, Pauli-$\mathbf{Y}$, and Pauli-$\mathbf{Z}$ observables across all wires in the PQC, and the measurement outcomes are combined to form $\mathbf{z}_t$ (Eq. 3). For the feedforward network $\mathcal{F}$ (Eq. 1 and Eq. 2), we experimented with ReLU, leaky ReLU, GLU and GELU non-linearities. [8] For both language modeling and translation, we first transform

[5] We may add extra transformations to the measurement outcomes before classifications or feeding them to the next step; see §4.

[6] We slightly abuse notation by writing $U(\boldsymbol{\theta}_t)$ to denote all unitary operations composed of multiple parametrized gates, each acting on one or more qubits with parameters drawn from $\boldsymbol{\theta}_t$.

[7] See AppendixB for details on the PQC design and expressibility evaluation methodology.

[8] 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 measurements 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	5.2K
$\text{QRNN}_{\text{LeakyReLU}}$	87.41	**87.00**	8_{24}	100	5.2K
$\text{QRNN}_{\text{GELU}}$	87.53	86.38	8_{24}	100	5.2K
$\text{QRNN}_{\text{Linear}}$	85.37	84.21	8_{24}	100	5.2K
$\text{QRNN}_{\text{Linear}}$	84.21	83.22	4_{12}	100	2.6K
RNN	87.64	86.96	50	50	5K
LSTM	88.40	86.79	25	25	5.1K
LSTM†	–	86.5	1,024	512	6.2M
scoRNN	84.05	83.14	170	100	31K

the measurement outcomes 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 from 4 minutes for MNIST (with 10 qubits) to 60 minutes for language modeling (with 14 qubits). Hyperparameters for the hybrid model common to all the tasks, including the Adam optimizer (Kingma and Ba, 2014) ($lr = 1 \times 10^{-3}$, $\lambda = 1 \times 10^{-4}$ and $\epsilon = 1 \times 10^{-10}$) with no learning rate decay. Dropout, with task-dependent drop probabilities, is applied to the measurement outcomes from the previous timestep and the current input embedding, which are encoded by the feedforward network $\mathcal{F}$ into the rotation angles of the PQC. We apply full-sequence backpropagation without truncation, except for language modeling, where sequences are truncated to 35 tokens per standard practice. 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).in Helfrich et al. (2018) 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.

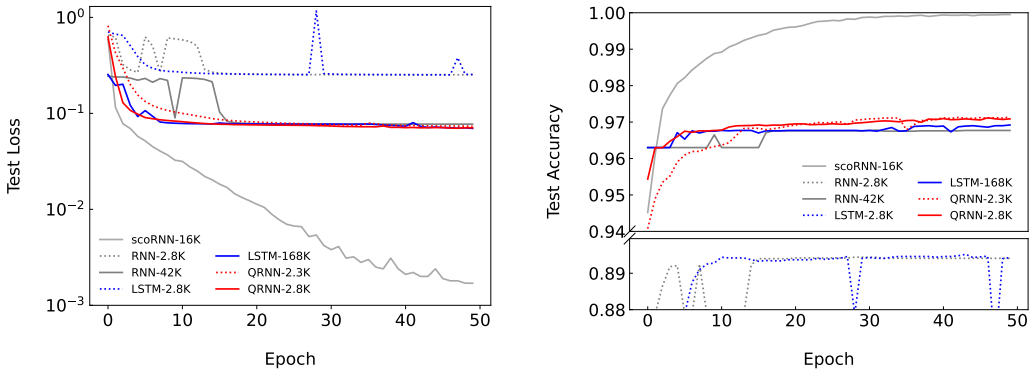
The hybrid model for this task follows the generic hybrid architecture described in §3.2. At the final input token, we apply an affine transformation to the measurement outcomes to produce two logits, which are used for classification via cross-entropy loss. Table 1 summarizes the results. The QRNN with LeakyReLU nonlinearity 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 down-sampling. We use the same model architecture 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 given the current limitations of the toolkit. Therefore we experimented with a simplified version in which the pixels of

Table 2: Classification accuracy on MNIST and pMNIST. Qubit count q, total measurements 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	3.9K
scoRNN	97.94	97.12	96.86	95.56	170	28	16K



(a) (b)

Figure 3: Test loss (a) and accuracy (b) for the copying memory problem with $T = 200$.

each digit are first permuted and 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 (where nonlinearity is ablated). 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 $\text{QRNN}_{\text{GELU}}$.

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 matches the large

Table 3: PTB word-level language modelling (PPL). Qubit count q, total measurements 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}	650	130K
$\text{QRNN}_{\text{LeakyReLU}}$	131.41	126.58	14_{42}	650	130K
$\text{QRNN}_{\text{GELU}}$	136.62	131.07	14_{42}	650	130K
$\text{QRNN}_{\text{LeakyReLU}}$	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	131K

Table 4: Multi-30 German to English translation (BLEU). Qubit count q, total measurements 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	390K
$\text{QRNN}_{\text{LeakyRELU}}$	29.22	28.99	13_{39}	512	340K
$\text{QRNN}_{\text{GELU}}$	29.95	29.14	13_{39}	512	340K
QRNN_{GLU}	30.16	31.51	10_{30}	512	360K
QRNN_{GLU}	27.63	29.66	5_{15}	512	270K
RNN	29.17	29.20	512	256	390K
LSTM	29.20	**32.20**	256	124	390K

168K-parameter LSTM (loss 0.07, accuracy 97%) and outperforms the smaller 2.8K LSTM (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 tested scoRNN on this task but found that it did not converge to a good solution. The LSTM achieved the best result, with 120.30 perplexity (PPL), followed closely by $\text{QRNN}_{\text{ReLU}}$ at 126.69.

4.5 Machine Translation

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 German-to-English translation using the Multi30k dataset (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 ten qubits closely matches the performance of the LSTM baseline, followed by QRNN_{GLU} with five 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,

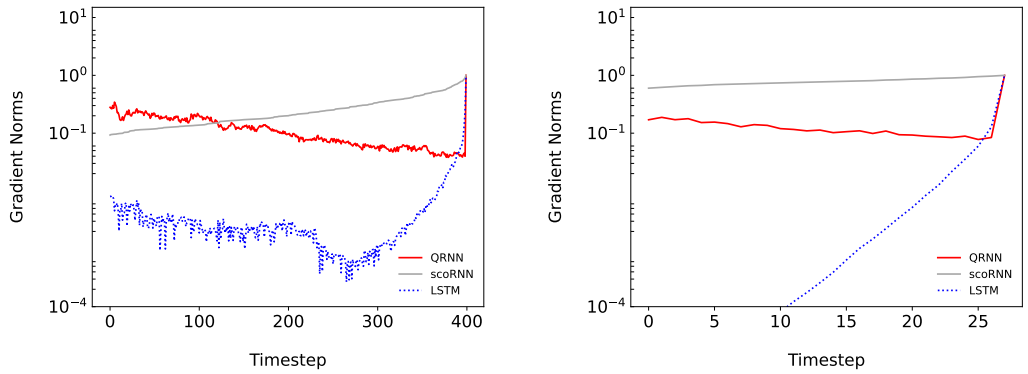


Figure 4: Normalized per-timestep gradient norms $\|\partial\mathcal{L}/\partial\mathbf{h}_t\|_2$ during training, averaged over one mini-batch containing samples of identical T (batch size $= 16$). Curves are normalized by the final timestep ($t = T$) gradient to compare decay shape; higher gradient values closer to $T = 0$ indicate less vanishing. (a) IMDB, $T = 400$. (b) pMNIST, $T = 28$.

despite intermediate measurements, sufficient information is retained and propagated across timesteps. We qualitatively interpret the learned soft alignments in Appendix D.

4.6 Hidden State Gradients

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) during training from saved checkpoints and computing $\|\partial\mathcal{L}/\partial\mathbf{h}_t\|_2$. For each mini-batch we normalize the curve by the last-step norm $\|\partial\mathcal{L}/\partial\mathbf{h}_T\|_2$ to compare decay shape, then average across samples in a mini-batch. As shown in Fig. 4, the QRNN curves remains consistently above the LSTM curves on both IMDB and pMNIST, indicating slower decay and less vanishing through time toward the start of the sequences. Both curves reach 1.0 at $t = T$ by construction (normalization), but the relative elevation of the QRNN curve at earlier timesteps demonstrates better gradient propagation.

5 Discussion and Conclusion

Different quantum hardware platforms currently require distinct control stacks and implementations and architectural choices do not translate one-for-one across devices. Our goal is not to prescribe a hardware roadmap but to analyze a hardware-realistic base case in an idealized simulator to isolate the computation performed by the architecture. The PQC gate set is native on most platforms and involves no nonstandard gates. We model mid-circuit observation using projective measurements as a limiting case. With better simulation toolchain and budgets allowing more qubits (e.g., future multi-GPU toolkits based on `cuQuantum` (Bayraktar et al., 2023)), direct projective reads can be replaced by an ancilla-mediated scheme, in which extra auxiliary qubits are entangled with the PQC, measured, and reset if needed, while the memory register remains coherent. Such mid-circuit measure-and-reset operations are already supported on several platforms (DeCross et al., 2022; Lis et al., 2023; Norcia et al., 2023) and align naturally with our design.

This paper contributes to sequence learning by demonstrating recurrent networks grounded in quantum operations, with a recurrent core implemented as a parametrized unitary quantum circuit and a lightweight classical controller that steers the recurrent evolution. The unitary dynamics preserve norms, promoting stable gradient propagation, while the classical controller injects nonlinearity and task adaptivity needed for expressiveness. The result is a compact, principled architecture that unifies unitary recurrence, partial observation, and classical control. As simulation techniques improve and quantum hardware matures, this points toward practical, hardware-realistic quantum models for sequential learning.

A 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 *amplitutes* 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 $|\beta|^2$.

B 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.

C Experimental Settings and Test Accuracy Statistics Across Runs

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

Task	b	d	e_{init}
IMDB	200	0.25	Xavier Uniform
MNIST	200	0.0	-
PTB	64	0.5	Xavier Uniform
Multi-30	64	0.25	Xavier Uniform

Table 6: Accuracy statistics on IMDB test set across 100 runs for each nonlinearity variant. Qubit count q, total measurements m; embedding dimension e; parameter count p. 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 nonlineariteis 40, 42 and 25 failed runs were observed each. The results also indicate that GELU nonlineartiy reduces the senstivity compared with the other two.

Model	min	max	μ	min^*	μ^*	q_m	e	p
$\text{QRNN}_{\text{ReLU}}$	49.55	85.96	71.18	71.74	83.11	8_{24}	100	5.2K
$\text{QRNN}_{\text{LeakyReLU}}$	49.63	87.00	70.23	75.77	83.44	8_{24}	100	5.2K
$\text{QRNN}_{\text{GELU}}$	49.98	86.38	77.18	70.39	83.75	8_{24}	100	5.2K

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.

Table 7: Accuracy statistics on MNIST and pMNIST test sets across 50 runs for each nonlinearity variant. Qubit count q and total measurements m; embedding dimension e; parameter count p.

	MNIST			**pMNIST**					
Model	min	max	μ	min	max	μ	q_m	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

Table 8: BLEU evaluations on the Multi-30 German to English test set across 20 runs for each nonlinearity variant. Qubit count q, total measurements m; embedding dimension e; parameter count p.

Model	min	max	μ	q_m	e	p
QRNN_{GLU}	19.83	31.92	27.88	13_{39}	512	390K
$\text{QRNN}_{\text{LeakyRELU}}$	24.52	29.87	28.55	13_{39}	512	340K
$\text{QRNN}_{\text{GELU}}$	25.71	30.29	29.09	13_{39}	512	340K

D Non-Monotonic Alignments

To qualitatively analyze the model's learned soft attention alignments we selected four sentences from test set which required non-monotonic alignments and interpreted the hybrid model translations and alignments (Fig. 5).

We observe that the hybrid model can manage spatial and syntactic shifts while capturing clause-level structure and semantics through its measurement-driven 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. 5a) and *"Zwei grün gekleidete Männer bereiten in einem Restaurant Essen zu"* (Fig. 5b), 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"* $\rightarrow$ *"a crowd of people"* (Fig. 5a) and *"gekleidete Männer"* $\rightarrow$ *"men in green outfits"* (Fig. 5b) 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. 5a) *"Menschen, die vor einem großen Gebäude im Kreis sitzen"* (Fig. 5c). In the former, German's verb-final structure is reorganized into a mid-sentence English verb phrase, while

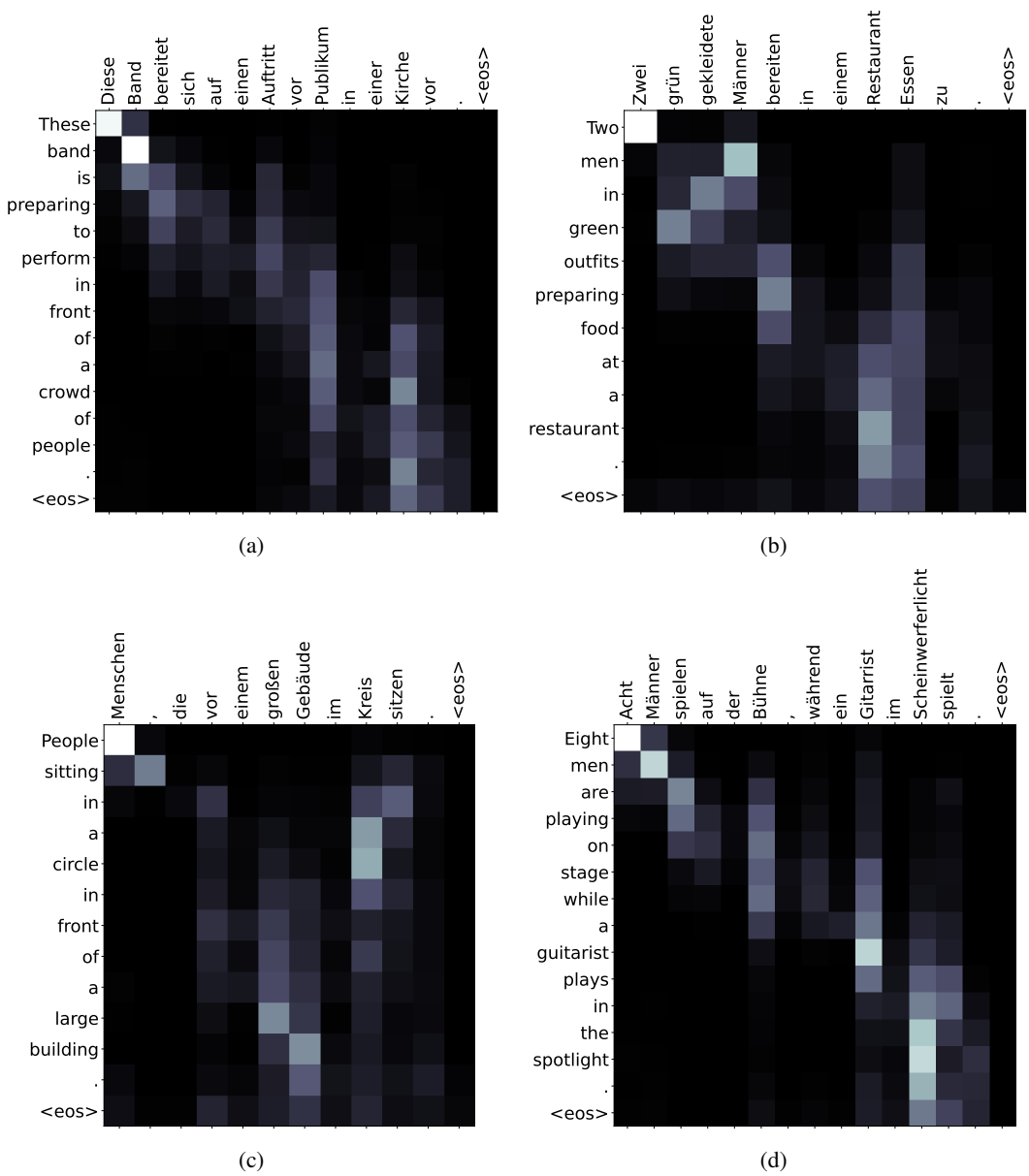


Figure 5: Non-monotonic alignments produced by the QRNN encoder-decoder model.

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. 5d). 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. 5c), where the model tracks relative clause dependencies and maps them onto compact English constructions.

E LLM USAGE

We used LLMs to assist with spelling, grammar, and wording improvements.

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.

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.

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.

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.

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.

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.

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.

Xiaoxuan Pan, Zhide Lu, Weiting Wang, Ziyue Hua, Yifang Xu, Weikang Li, Weizhou Cai, Xuegang Li, Haiyan Wang, Yi-Pu Song, Chang-Ling Zou, Dong-Ling Deng, and Luyan Sun. Deep quantum neural networks on a superconducting processor. *Nature Communications*, 14(1):4006, 2023. doi: 10.1038/s41467-023-39785-8. URL https://www.nature.com/articles/s41467-023-39785-8.

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.

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

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.

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.

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.