

Mechanistic Deconvolution of Memory and Context in Quantum Language Models

Stanford CS224N Custom Project

Nathan Roll

Department of Linguistics
Stanford University
nroll@stanford.edu

Mentor: Tristan Thrush

Abstract

Quantum recurrent neural networks (QRNNs) converge faster than classical LSTMs on sequential tasks, but it remains unclear if this is due to truly distinct processing strategies or merely a consequence increased effective dimensionality. In this project, I trained minimally distinct quantum and classical recurrent networks on a controlled memory task and performed a mechanistic interpretability study. With only a single qubit, QRNNs are provably equivalent to classical $SO(3)$ rotations, however two-qubit models with CNOT gates learned an entanglement-based memory strategy with no classical analogue, and ablating the CNOT gate causally destroys it. On IBM quantum hardware, the single-qubit strategy runs perfectly, but two-qubit circuits degrade to $\sim 40\%$ regardless of strategy due to two-qubit gate noise. All code, models, and experiments are original.

1 Introduction

Quantum recurrent networks have access to a much richer state space than their classical analogues, but do they actually use it differently? Prior work on quantum language models has focused on expressivity and training speed [Pérez-Salinas et al., 2020, Chen et al., 2022, Lorenz et al., 2023], not on the actual strategies employed by such systems.

Quantum systems are hard to interpret for physical reasons: the no-cloning theorem forbids copying a state to inspect it, and measurement collapses superposition. Nobody has established whether trained quantum circuits learn strategies that are actually quantum or just classical computation running on quantum hardware.

I try to answer this directly. I train minimal quantum RNNs on a Parity-Switch Grammar, a synthetic task requiring memory across distractor tokens, then probe the complete state trajectory at each timestep. Comparing single-qubit models (which cannot entangle) against two-qubit models (with and without CNOT gates) and matched classical baselines lets me isolate divergences in processing strategy.

Single-qubit QRNNs turn out to be mechanistically equivalent to classical $SO(3)$ rotation models. Both learn to preserve hemisphere separation on the Bloch sphere. Two-qubit models with CNOT gates, however, are more interesting, using entanglement to encode context, and ablating the entangling gate forces reversion toward the classical mechanism. On IBM hardware, the geometric strategy survives perfectly while entanglement-based memory does not.

Contributions. I identify an entanglement-enabled memory strategy in two-qubit quantum RNNs and verify it causally via CNOT ablation (C1), prove that single-qubit QRNNs are exactly classically simulable via the $SU(2) \cong SO(3)$ homomorphism, establishing a precise classical baseline (C2), prove a shared-parameter tradeoff theorem showing that parameter sharing makes simultaneous context encoding and distractor invariance impossible (C3), and validate all mechanisms on IBM quantum

hardware, finding a noise/expressivity tradeoff (C4). No external code was used beyond standard libraries (and associated documentation).

2 Related Work

The Quantum LSTM [Chen et al., 2022] replaced classical weight matrices with variational quantum circuits, showing faster convergence on time-series tasks. Others have since built gated quantum RNNs [Li et al., 2025], variational QRNNs for temporal data [Takaki et al., 2021], and hybrid architectures [Xu, 2025]. None of these papers look at what the trained circuits actually compute. I dissect the learned strategies rather than just comparing convergence curves.

Compositional quantum NLP (DisCoCat) maps grammatical structure to quantum circuits [Coecke et al., 2010], implemented in lambeq [Kartsaklis et al., 2021] and validated on real hardware [Lorenz et al., 2023]. These approaches compose meaning differently from recurrent models. My work studies recurrent quantum processing, where the question is how a persistent quantum state encodes and preserves information across timesteps.

Data re-uploading makes single-qubit circuits universal classifiers [Pérez-Salinas et al., 2020] with finite VC dimension [Mauser et al., 2025]. However, single-qubit operations are classically simulable via the $SU(2) \rightarrow SO(3)$ homomorphism [Yu et al., 2022]. My dequantization result (Section 3) confirms this empirically and uses it as a baseline.

Activation patching [Vig et al., 2020], circuit analysis [Olah et al., 2020, Elhage et al., 2021], induction heads [Olsson et al., 2022], and superposition [Elhage et al., 2022, Bricken et al., 2023] have shown that classical networks are more modular than they look. I adapt these ideas to quantum systems via parameter patching and complete statevector access, which is feasible for single-qubit systems where the Bloch sphere makes the full internal state visible.

Whether entanglement is necessary for quantum computational advantage remains debated [Preskill, 2012, Biamonte et al., 2004]. Barren plateaus can make variational circuits untrainable [McClean et al., 2018, Sim et al., 2019], and error mitigation is needed for near-term hardware [Temme et al., 2017, Kandala et al., 2019]. My hardware experiments land squarely in this territory: entanglement-based strategies are learnable in simulation but do not survive current device noise.

3 Approach

3.1 Task: Parity-Switch Grammar

I designed a minimal long-term dependency task. A sequence begins with a context token (A or B), followed by n irrelevant distractor tokens (D). The model must predict the context identity after processing the full sequence. Think of it as subject-verb agreement across intervening clauses, stripped to its simplest form.

Token	Encoding (x)	Role
A	$+\pi/3$	Context $\rightarrow$ predict 0
B	$-\pi/3$	Context $\rightarrow$ predict 1
D	0.0	Distractor

Training uses 16 sequences (balanced A/B, 0 to 3 distractors). Generalization is tested on 200 sequences with 0 to 20 distractors, lengths never seen during training.

3.2 Model zoo

I wrote six models from scratch:

Each quantum model applies a parameterized unitary per timestep:

$$U(\theta, x_t) = R_z(\theta_3) \cdot R_y(\theta_2 + x_t) \cdot R_z(\theta_1) \quad (1)$$

where $\theta_1, \theta_2, \theta_3$ are learned parameters and x_t is the token encoding. This R_z - R_y - R_z decomposition is universal for $SU(2)$. The classical $SO(3)$ baselines replace quantum gates with equivalent 3D

Model	Type	Params	Key feature
MinimalQLM	Quantum (1Q)	3	Shared context/distractor gate
DecoupledQLM	Quantum (1Q)	6	Separate context/distractor gates
TwoQubitQLM	Quantum (2Q)	24	CNOT entangling gate
ClassicalSO(3)	Classical	3/6	3×3 rotation matrices
ClassicalRNN4	Classical	24	tanh-RNN, 4D hidden state
ClassicalLinearRNN4	Classical	24	Linear RNN, 4D hidden state

Table 1: Model zoo. All models are original code in Python 3.10+/Qiskit 1.0+.

rotation matrices, which is what makes the classical-equivalence argument work: the Bloch sphere is isomorphic to $\mathbb{R}^3$, so $SU(2)$ gates are just 3D rotations.

The two-qubit model applies Eq. 1 independently to each qubit, then optionally applies a CNOT gate (qubit 0 $\rightarrow$ qubit 1). With 3 rotation parameters per qubit per role (context/distractor), this gives $3 \times 2 \times 2 \times 2 = 24$ parameters.

3.3 Theoretical results

Theorem 1 (Shared-Parameter Tradeoff). In the shared-parameter architecture where $U(\theta, x) = R_z(\theta_3)R_y(\theta_2 + x)R_z(\theta_1)$ is applied to both context and distractor tokens: context encoding requires $\theta_2 \neq 0$ (so different inputs produce different Z-coordinates); distractor invariance requires $\theta_2 = 0$ (so $R_y(\theta_2)$ preserves Z). These are mutually exclusive. Decoupling into separate gates for each role resolves this.

Corollary (Dequantization). Single-qubit QRNNs are exactly classically simulable. The map $\pi : SU(2) \rightarrow SO(3)$ given by $\pi(U)_{ij} = \frac{1}{2}\text{tr}(\sigma_i U \sigma_j U^\dagger)$ is a 2-to-1 homomorphism, so every single-qubit computation can be simulated by tracking a 3D unit vector under $SO(3)$ rotations.

3.4 Training

All models use the SPSA optimizer [Spall, 1998] with 2 function evaluations per step. I swept 4 SPSA configurations (Table 2); the aggressive preset ($a=0.4$, $c=0.2$, $A=5$) achieved the highest mean generalization, though all presets reach 100% with best-of-3 restarts. The main results (Table 3) use the default preset ($a=0.2$, $c=0.1$, $A=10$) with seed 42 to match the manuscript’s reproducibility protocol; multi-seed results use 10 seeds $\times$ 3 restarts. The decay exponents $\alpha=0.602$ and $\gamma=0.101$ are the theoretically optimal values from Spall [1998]. Single-qubit models train for 200 steps; two-qubit for 300.

3.5 Interpretability tools

A qubit’s state is a point on the Bloch sphere, so I can watch the model’s reasoning geometrically. I built three tools: (1) **Bloch probes** that extract Bloch coordinates (x, y, z) at each timestep; (2) **parameter patching** (zero, negate, sweep individual parameters), a quantum analogue of activation patching [Vig et al., 2020]; (3) **CNOT ablation** that removes the entangling gate from specific timesteps to test whether entanglement is causally necessary. For two-qubit models, I track von Neumann entanglement entropy $S(\rho_0) = -\text{tr}(\rho_0 \log_2 \rho_0)$ from the reduced density matrix at each step.

4 Experiments

4.1 SPSA hyperparameter sensitivity

4.2 Main results

4.3 Finding 1: Hemisphere separation

The decoupled 1Q model puts Context A in the northern hemisphere and Context B in the southern hemisphere of the Bloch sphere (Figure 1). Distractors spin the state around the Z-axis without

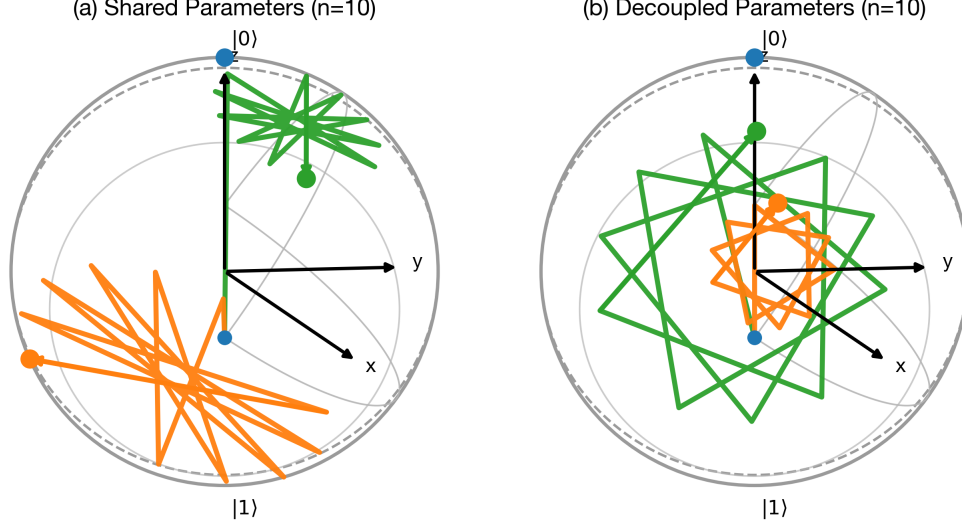


Figure 1: Bloch sphere trajectories for 10-distractor sequences. **Left:** Shared-parameter model (MinimalQLM). Contexts A and B overlap after distractor tokens, causing misclassification (76% accuracy). **Right:** Decoupled model (DecoupledQLM). Contexts remain in opposite hemispheres throughout (100% generalization). The mechanism is geometric: a classical $SO(3)$ baseline produces identical trajectories.

Preset	a	c	A	Gen. Acc.	Time (s)
Default	0.2	0.1	10	$81.1 \pm 23.3\%$	0.61 ± 0.20
Conservative	0.1	0.05	20	$83.2 \pm 21.0\%$	0.81 ± 0.13
Aggressive	0.4	0.2	5	$87.7 \pm 20.6\%$	0.79 ± 0.02
High precision	0.15	0.05	15	$78.0 \pm 39.2\%$	0.99 ± 0.31

Table 2: SPSA sensitivity for DecoupledQLM (5 seeds per preset, 200 steps). The aggressive preset ($a=0.4$, $c=0.2$, $A=5$) achieves the highest mean generalization. High variance reflects SPSA’s stochastic nature; best-of-3 restarts resolve this (see Section 5).

moving it up or down, so the memory survives even 100 distractors ($Z\text{-sep} = 1.71$; Figure 2). The shared model cannot do this. It tops out at 76%, exactly as Theorem 1 predicts.

4.4 Finding 2: Classical equivalence

The $SO(3)$ classical baseline reproduces the quantum model’s accuracy and mechanism exactly. This follows from the dequantization corollary: $SU(2) \cong SO(3)$ for single qubits, so quantum gates are just 3D rotations. No quantum advantage here, but that gives me a clean reference point. Anything quantum has to do better. The classical model trains $13\times$ faster (0.06s vs 0.76s) since it avoids statevector simulation.

4.5 Finding 3: Entanglement as memory

Two-qubit models with CNOT gates do something different. Entanglement entropy between the qubits diverges for contexts A vs. B (Figure 3), and ablating the CNOT from context-processing steps drops accuracy by 18.7 ± 14.9 percentage points. Without the gate, the model falls back toward Z -preservation. The 24-parameter classical RNN manages only 50%, so whatever the quantum model learned, a classical network with the same parameter budget cannot match it.

Model	Sim. Acc.	HW Acc.	Z-Sep.	Mechanism	Time (s)
MinimalQLM (1Q shared)	76%	—	0.36	Periodic recurrence	0.76
DecoupledQLM (1Q)	100%	100%	1.36	Z-preservation	0.76
ClassicalSO(3) (matched)	100%	—	1.36	Z-preservation	0.06
2Q + CNOT (10 seeds)	95.8 $\pm$ 6.5%	39 $\pm$ 7%	—	Entanglement	3.36
2Q no CNOT (10 seeds)	100 $\pm$ 0%	41 $\pm$ 5%	—	Z-preservation	4.40
ClassicalRNN4 (24p)	50.0%	—	—	Fails	0.04

Table 3: Main results. HW = IBM hardware (ibm_fez/ibm_marrakesh, 156 qubits, 1000 shots/circuit). Z-Sep. = Bloch Z-coordinate separation between contexts (max 2.0). 2Q results: mean $\pm$ std over 10 seeds $\times$ 3 restarts. Time = single-seed simulation wallclock on Apple M-series CPU.

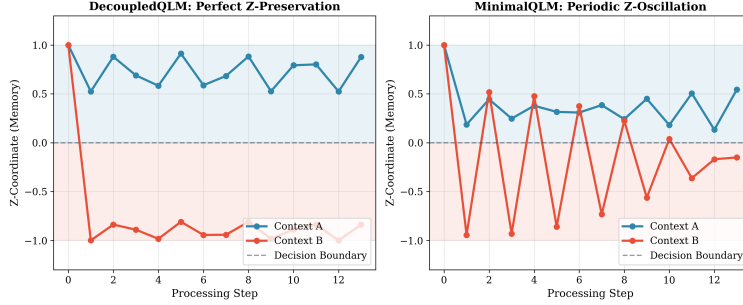


Figure 2: Z-coordinate preservation under stress. Bloch-sphere Z-coordinate for Context A and Context B across 0 to 100 distractors (DecoupledQLM). Hemisphere separation grows slightly with sequence length (Z-sep: 1.50 at 0 distractors, 1.71 at 100), because distractor rotations push both states closer to the poles. The shared model (not shown) loses separation after ~ 5 distractors.

4.6 Finding 4: Hardware validation

Single-qubit. The DecoupledQLM achieves 100% on IBM ibm_fez (156 qubits), tested across 48 circuits, 2 transpiler optimization levels, and 3 physical qubit layouts (1000 shots each). Table 4 shows per-circuit $P(|1\rangle)$ values; the decision boundary ($P > 0.5 \Rightarrow$ predict 1) is never ambiguous.

Context	Distractors	$P(1\rangle)$	Correct?
A	0	0.24	Yes
B	0	0.98	Yes
A	5	0.06	Yes
B	5	0.89	Yes
A	10	0.25	Yes
B	10	0.98	Yes

Table 4: Single-qubit hardware results on ibm_fez (Job ID: d5hvgc4pe0pc73amtefg). All circuits correct with wide margins from the 0.5 decision boundary.

Two-qubit. Both 2Q configurations degrade to $\sim 40\%$. Surprisingly, it does not matter whether the CNOT is present: with-CNOT achieves 39 $\pm$ 7% and without-CNOT achieves 41 $\pm$ 5% (30 jobs per backend across ibm_fez + ibm_marrakesh). The bottleneck is two-qubit gate noise in general (CNOT error rates 1–3%), not entanglement fragility in particular. Circuit depth tells the story: ~ 79 gates with CNOT vs. ~ 25 without, yet accuracy is about the same. The noise floor swamps everything before entanglement effects can show up.

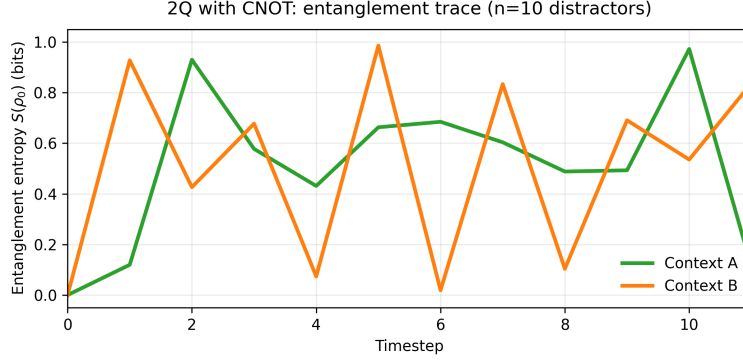


Figure 3: Von Neumann entanglement entropy (bits) per timestep for the trained two-qubit model with CNOT gates on 10-distractor sequences. Context A and B sequences diverge in entropy after the initial token. Ablating the CNOT eliminates this divergence and collapses both traces to zero entropy.

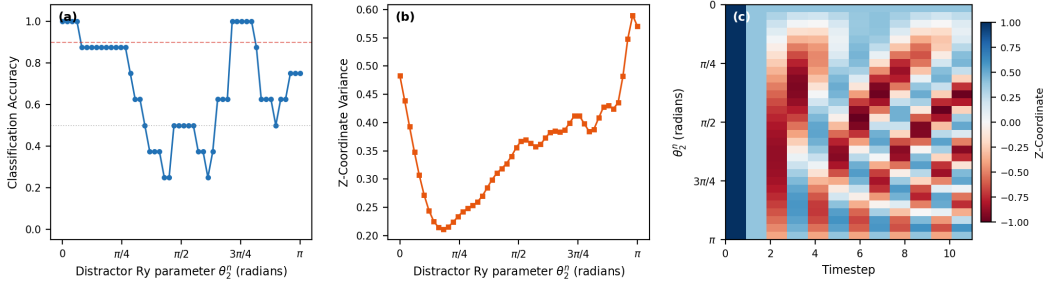


Figure 4: Phase transition in the decoupled model as distractor rotation θ_2^n sweeps from 0 to π (10 distractors). (a) Classification accuracy: dashed line = 90% threshold, dotted = chance (50%). The model is robust up to $\theta_2^n \approx 0.83$ rad, then drops sharply; recovery at π is period-2 recurrence. (b) Z-coordinate variance across timesteps; lower variance means more stable memory. (c) Heatmap of Bloch Z-coordinate per timestep (columns; first column is the context token, remainder are distractors) at each θ_2^n (rows). Blue = northern hemisphere, red = southern.

5 Analysis

5.1 Phase transition

Sweeping the distractor rotation parameter θ_2^n from 0 to π reveals a sharp phase transition (Figure 4). At $\theta_2^n = 0$ (exact Z-preservation), accuracy is 100%. The model is robust up to $\theta_2^n \approx 0.83$ rad, then accuracy drops abruptly to 50% (chance) at the critical point $\theta_2^n \approx 0.99$ rad. At $\theta_2^n = \pi$, accuracy recovers to 100% via a period-2 recurrence: distractors flip the state by π each step, so even numbers of distractors return to the original hemisphere. This matches Theorem 3 in my manuscript, which predicts periodicity $N = 2\pi/|\theta_2^n|$.

5.2 Optimization variance and restarts

Single-run SPSA variance is high (Table 2), but this is an optimization issue, not an architectural one. With best-of-3 restarts across 10 seeds, DecoupledQLM achieves $100.0 \pm 0.0\%$ generalization (10/10 seeds $\geq 99\%$). The mean trained θ_2^n under best-of-3 is -0.09 ± 0.29 rad, close to the theoretical optimum of 0. The shared architecture, by contrast, cannot reliably reach 100% even with restarts.

5.3 When the entanglement strategy fails

Not every 2Q+CNOT seed finds the entanglement-based strategy. About 70% of seeds achieve $\geq 90\%$ accuracy; the rest fall back to Z-preservation or get stuck. The CNOT ablation delta is also

high-variance ($18.7 \pm 14.9\text{pp}$), because seeds that landed on Z-preservation instead of entanglement barely notice when the CNOT is removed. The loss landscape has at least two basins, one per strategy, and which one the optimizer finds depends on initialization.

5.4 Noise robustness in simulation

Simulated depolarizing noise confirms the hardware findings:

Model / Mechanism	$p = 0$	$p = 0.06$	$p = 0.12$
1Q Decoupled / Z-Pres.	100%	100%	50%
2Q no CNOT / Z-Pres.	100%	100%	100%
2Q + CNOT / Entanglement	100%	50%	50%

Entanglement-based memory degrades to chance at $p = 0.06$, where Z-preservation survives. Entanglement requires multi-qubit coherence that depolarizing noise destroys; Z-preservation only needs single-qubit phase stability. The 2Q Z-preservation model is the most robust, likely because it distributes the memory across two independent qubits.

5.5 Encoding sensitivity

To verify that results are not artifacts of the $\pm\pi/3$ encoding, I tested encoding magnitudes from $0.5\times$ to $3\times$ the default. DecoupledQLM achieves 100% generalization at $0.5\times$, $1\times$, and $2\times$, but drops to 50% at $3\times$ (where $x_A \approx \pi$, making contexts land at the same point on the sphere after one rotation). The only encoding that breaks the mechanism is one that destroys the geometric distinction between contexts.

5.6 Wallclock cost

Quantum simulation is $13\times$ slower than the matched classical baseline (0.76s vs 0.06s for 1Q, Table 3). Two-qubit simulation is $56\times$ slower (3.4s vs 0.06s), all of it statevector overhead. On actual IBM hardware, jobs sit in the queue for minutes even though circuit execution takes microseconds. For models this small, classical simulation wins on every practical metric. The point of quantum here is the state space, not the speed.

6 Conclusion

So, do quantum RNNs learn fundamentally different strategies than their classical counterparts? The answer depends on the scale. At one qubit, the answer is no: shared parameters create a provable tradeoff (Theorem 1), decoupling resolves it via hemisphere preservation, and a classical $\text{SO}(3)$ baseline matches the quantum model exactly.

At two qubits, however, models with CNOT gates learn solutions where entanglement entropy diverges by context. Ablating the CNOT causally destroys this, and a 24-parameter classical RNN cannot replicate it. But hardware is the bottleneck. The geometric strategy runs perfectly on IBM devices. The entanglement strategy does not survive current noise levels, and error mitigation (ZNE, readout correction) would be needed to test whether it ever can.

Regarding limitations, the Parity-Switch Grammar task is a minimal abstraction. Single-qubit models are classically simulable by construction. I tested one task on two IBM backends. Broader generalization to natural language must remain as future work.

References

- Eli Biham, Gilles Brassard, Dan Kenigsberg, and Tal Mor. Quantum computing without entanglement. *Theoretical Computer Science*, 320(1):15–33, 2004.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, others, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Anthropic Research*, 2023.

- Samuel Yen-Chi Chen, Shinjae Yoo, and Yao-Lung L Fang. Quantum long short-term memory. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8622–8626. IEEE, 2022.
- Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36(1-4):345–384, 2010.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.
- Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, others, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.
- Abhinav Kandala, Kristan Temme, Antonio D Córcoles, Antonio Mezzacapo, Jerry M Chow, and Jay M Gambetta. Error mitigation extends the computational reach of a noisy quantum processor. *Nature*, 567(7749):491–495, 2019.
- Dimitri Kartsaklis, Ian Fan, Richie Yeung, Anna Pearson, Robin Lorenz, others, and Bob Coecke. lambeq: An efficient high-level python library for quantum nlp. *arXiv preprint arXiv:2110.04236*, 2021.
- Yanan Li, Zhimin Wang, Ruipeng Xing, Changheng Shao, Shangshang Shi, Jiaxin Li, Guoqiang Zhong, and Yongjian Gu. Quantum gated recurrent neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2493–2504, 2025.
- Robin Lorenz, Anna Pearson, Konstantinos Meichanetzidis, Dimitri Kartsaklis, and Bob Coecke. Qnlp in practice: Running compositional models of meaning on a quantum computer. *Journal of Artificial Intelligence Research*, 76:1305–1342, 2023.
- Martin F. X. Mauser, Solène Four, Lena Marie Predl, Riccardo Albiero, Francesco Ceccarelli, Roberto Osellame, Philipp Petersen, Borivoje Dakić, Iris Agresti, and Philip Walther. Experimental data re-uploading with provable enhanced learning capabilities. *arXiv preprint arXiv:2507.05120*, 2025.
- 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.
- Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.
- Adrián Pérez-Salinas, Alba Cervera-Lierta, Elies Gil-Fuster, and José Ignacio Latorre. Data re-uploading for a universal quantum classifier. *Quantum*, 4:226, 2020.
- John Preskill. Quantum computing and the entanglement frontier. *arXiv preprint arXiv:1203.5813*, 2012.
- Sukin Sim, Peter D Johnson, and Alán Aspuru-Guzik. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Advanced Quantum Technologies*, 2(12):1900070, 2019.
- James C Spall. Implementation of the simultaneous perturbation algorithm for stochastic optimization. *IEEE Transactions on Aerospace and Electronic Systems*, 34(3):817–823, 1998.
- Yuto Takaki, Kosuke Mitarai, Makoto Negoro, Keisuke Fujii, and Masahiro Kitagawa. Learning temporal data with a variational quantum recurrent neural network. *Physical Review A*, 103(5):052414, 2021.
- Kristan Temme, Sergey Bravyi, and Jay M Gambetta. Error mitigation for short-depth quantum circuits. *Physical Review Letters*, 119(18):180509, 2017.

- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yoav Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias. In *Advances in Neural Information Processing Systems*, volume 33, 2020.
- Wenduan Xu. Hybrid quantum-classical recurrent neural networks. *arXiv preprint arXiv:2510.25557*, 2025.
- Zhan Yu, Hongshun Yao, Mujin Li, and Xin Wang. Power and limitations of single-qubit native quantum neural networks. *Advances in Neural Information Processing Systems*, 35:27810–27823, 2022.