Biomedical Testing & Sensing
2026, VOL. 5, NO. 1, 38-41
DOI: 10.47297/wspbtsWSP2755-952108.20260501

Quantum-enhanced and Quantum-inspired Training for Deep Sequence Models: A Concise Review

Yifeng Hou

School of Aerospace Engineering and Applied Mechanics, Tongji University, Shanghai, 201804, China

ABSTRACT
Deep sequence models are increasingly combined with quantum computing and quantum-inspired optimization, yet the resulting literature uses inconsistent terminology and reports uneven evidence. This concise review distinguishes quantum-enhanced models, which execute parameterized quantum circuits, from quantum-inspired methods implemented on classical hardware. It organizes recent work by the point at which quantum principles enter the learning pipeline: recurrent gates, attention and feature maps, kernel representations, and training optimizers. Quantum long short-term memory networks and related hybrid architectures demonstrate expressive low-parameter models on small forecasting and classification tasks, while emerging superposition- and tunnelling-inspired optimizers target difficult neural-network loss landscapes. However, most reported gains are based on simulators, limited datasets, and comparisons that do not control computational budgets. The review therefore proposes a compact evaluation framework emphasizing strong classical baselines, wall-clock and energy costs, repeated trials, noise-aware experiments, and transparent definitions of quantum advantage. These practices are necessary to determine whether quantum mechanisms provide practical value for recurrent networks, Transformers, and future large-scale sequence models.

KEYWORDS
Quantum machine learning; Quantum-inspired optimization; Sequence models; QLSTM; Transformer

1 Introduction

Sequence modelling underpins forecasting, natural-language processing, signal analysis and decision making. Long short-term memory (LSTM) networks introduced a durable mechanism for learning temporal dependencies [1], while Transformers replaced recurrence with attention and became the dominant architecture for large-scale sequence learning [2]. Their success has also exposed costly training, high-dimensional non-convex optimization and growing memory demands. Quantum machine learning (QML) offers a different computational vocabulary based on superposition, entanglement and measurement [3]. In the noisy intermediate-scale quantum (NISQ) era [4], practical research mainly uses hybrid models in which a classical optimizer trains parameterized quantum circuits (PQCs) [5].

The literature now contains two related but technically distinct directions. Quantum-enhanced methods execute a quantum circuit, or simulate one, as part of a model. The quantum LSTM (QLSTM), for example, replaces classical transformations in recurrent gates with variational quantum circuits [6]. Quantum-inspired methods instead translate concepts such as probabilistic superposition, tunnelling or interference into classical algorithms. Recent examples include adaptive superposition optimization [7] and circuit-perturbed gradient descent for Transformer training [8]. Calling both directions simply 'quantum AI' obscures their resource requirements and the strength of their empirical claims.

This concise review focuses on how quantum and quantum-inspired mechanisms affect the representation and training of deep sequence models. It contributes an entry-point taxonomy, compares representative studies, and identifies the evidence required for credible progress. Because the field is young and heterogeneous, the aim is a critical narrative synthesis rather than a claim of exhaustive systematic coverage.

2 Scope and Taxonomy

A useful taxonomy asks where the non-classical mechanism enters the learning pipeline. First, architecture-level approaches replace a recurrent gate, attention block or output map with a PQC. Second, representation-level approaches use quantum feature maps, kernels or Fourier transforms while retaining a largely classical predictor. Third, optimizer-level approaches modify candidate selection or gradient updates using quantum-inspired dynamics. These categories are not mutually exclusive: a hybrid sequence model may use a PQC for representation but still rely on Adam for parameter updates.

The distinction between quantum-enhanced and quantum-inspired is essential. A simulated PQC preserves the mathematical operations of a quantum circuit but does not demonstrate a hardware speed-up. A classical algorithm

* **Corresponding Author:** Yifeng Hou, huat@tongji.edu.cn

named after superposition or tunnelling may be useful, but its advantage must be attributed to the implemented search rule rather than to quantum computation. Table 1 summarizes the main mechanisms and the claims that should be tested.

Table 1 Taxonomy of quantum-enhanced and quantum-inspired sequence learning

Entry point	Representative approach	Execution	Intended benefit	Primary evidence gap
Recurrent gates	QLSTM with PQCs [6]	Quantum device or simulator	Expressive gates with fewer trainable parameters	Small tasks and costly circuit evaluations
Attention	Quantum-inspired attention [9]	Classical hardware	Richer weighting or interactions	Unclear contribution of the quantum analogy
Feature maps	Quantum feature maps and QFT kernels [13,17]	Quantum device or simulator	High-dimensional temporal representation	State preparation and measurement overhead
Candidate search	Adaptive superposition optimization [7]	Classical hardware	Gradient-free exploration and escape from local minima	Scaling to modern sequence models
Gradient updates	Circuit-perturbed gradient descent [8]	Hybrid or simulated	Improved exploration during training	Fair comparison with tuned AdamW

3 Quantum-enhanced Sequence Architectures

3.1 Recurrent Models

The QLSTM proposed by Chen, Yoo and Fang is the clearest architecture-level example [6]. It retains the classical cell-state recurrence but uses six variational quantum circuits to produce gate values, hidden states and outputs. Classical inputs are encoded through rotations, entangling layers transform the quantum state, and expectation-value measurements return real-valued features. This design can increase functional expressivity per trainable parameter, but every time step requires repeated circuit executions and measurements. Consequently, an apparent parameter advantage can coexist with slower wall-clock training.

The architecture builds on the broader development of quantum neural networks and parameterized quantum-circuit learning [10-12]. These foundations explain how trainable circuits can act as non-linear function approximators, but they do not remove the recurrent execution cost introduced when a circuit is evaluated at every time step.

Subsequent sequence applications have expanded from synthetic signals to forecasting and decision systems. A QLSTM-based forecaster has been combined with a quantum-enhanced actor-critic agent for algorithmic trading [16]. Such studies show that hybrid recurrent modules can be embedded in larger pipelines, yet many are evaluated on noiseless simulators and a single train-test split. Their results therefore support feasibility more strongly than general quantum advantage. Recurrent quantum neural networks also provide a broader formulation in which a persistent quantum state carries temporal information, although trainability and hardware depth remain central constraints [15].

3.2 Attention, Kernels and Transforms

Not all sequence methods replace recurrent gates. Zhu and Lu introduced quantum-inspired attention parameters into an sLSTM groundwater-inversion model and reported improved accuracy and stability [9]. Because the model runs classically, the result should be interpreted as an attention innovation motivated by quantum theory, not a quantum-computing gain. This distinction is valuable: quantum inspiration can generate useful inductive biases without incurring quantum-hardware overhead, but it must be justified through ablation against equally expressive classical alternatives.

Representation-level approaches map temporal windows into quantum feature spaces. Quantum kernels estimate similarities through state overlaps [13-14], and a recent QFT-enhanced kernel demonstrated solar-irradiance forecasting on a five-qubit noiseless simulator [17]. Short-time quantum Fourier transform processing extends windowed signal analysis into the quantum domain and addresses overlap-add reconstruction [18]. These approaches are promising for periodic or local temporal structure, but their end-to-end value depends on data loading, repeated measurements and output extraction. Efficient circuit complexity alone does not guarantee efficient learning on classical datasets.

4 Quantum-inspired Optimization for Sequence Models

Optimizer-level work is smaller but strategically important because it can be applied without replacing an established sequence architecture. Quantum-inspired adaptive superposition optimization (QIASO) maintains a probability distribution over candidate neural-network weights and updates it through a KL-regularized amplification rule; stochastic tunnelling perturbations are introduced to escape poor local minima [7]. The method is gradient-free and conceptually

relevant to recurrent models whose loss surfaces may be difficult to navigate. However, maintaining and evaluating multiple candidate networks can become expensive as parameter counts grow.

Superpositional Gradient Descent instead augments a conventional update with perturbations derived from quantum-circuit or interference-inspired functions, and has been demonstrated on sequence classification and Transformer fine-tuning [8]. The broader opportunity is to use quantum-inspired updates as plug-in optimizers for LSTMs, Transformers and parameter-efficient fine-tuning. The appropriate baselines are therefore not only stochastic gradient descent and Adam [21], but also decoupled weight decay through AdamW [22] and low-rank adaptation methods such as LoRA and QLoRA [23-24]. Comparisons must equalize training tokens, parameter budgets, tuning effort and random seeds.

Hybrid PQC models introduce an additional optimization problem. Circuit gradients can be computed by parameter-shift rules [20], but repeated evaluations increase cost, and deep or poorly initialized circuits may exhibit barren plateaus with exponentially small gradients [19]. Quantum-inspired optimizers cannot by themselves remove these constraints. Architecture, encoding, optimizer and noise must therefore be studied jointly rather than reported as independent components.

5 Evidence Quality and Practical Limitations

Four recurring limitations prevent strong conclusions. First, most experiments use small datasets, few qubits and shallow circuits. A model that performs well under these conditions may not scale to long contexts or large feature spaces. Second, noiseless simulation removes device errors while retaining substantial classical simulation cost. Hardware experiments should report shot counts, queue time, error-mitigation procedures and variability. Third, many studies compare against an untuned LSTM or Transformer rather than strong contemporary baselines. Fourth, accuracy is often reported without wall-clock time, energy, memory or circuit-execution counts.

The meaning of advantage must also be specified. A representation advantage means improved predictive quality under a comparable resource budget; a parameter advantage means similar quality with fewer trainable parameters; a computational advantage requires lower end-to-end cost; and a hardware quantum advantage requires evidence unattainable by competitive classical methods. Expressivity results for quantum neural networks [27] do not automatically establish practical learning advantage. Data encoding can dominate performance [25], while carefully constructed problems are currently required for rigorous quantum learning speed-ups [26].

For compact but credible evaluation, future studies should report at least five elements: a tuned classical baseline; repeated trials with confidence intervals; parameter, time and energy budgets; simulator and hardware results when feasible; and ablations isolating the quantum or quantum-inspired component. Public code and fixed data splits are equally important. Benchmarking guidance should treat negative results as informative, since identifying where quantum models fail is necessary for mapping their useful regime [28].

6 Research Agenda

A first priority is a shared benchmark suite spanning synthetic memory tasks, multivariate forecasting, text classification and controlled Transformer fine-tuning. Each task should include matched classical, quantum-enhanced and quantum-inspired variants. A second priority is resource-aware design. Rather than replacing every gate or attention block, quantum modules should target bottlenecks where their expressive bias is measurable. Recent classical sequence advances such as xLSTM and improved long-horizon LSTM variants [29-30] should be included as baselines, because improvements over an older LSTM do not establish competitiveness.

A third direction is quantum-inspired parameter-efficient training. Candidate-distribution and interference-inspired updates may be more realistic for adapters or low-rank modules than for all model weights. Finally, studies should connect optimizer behaviour to measurable properties of the loss landscape, such as curvature, gradient noise and basin diversity. This would convert appealing quantum analogies into testable mechanisms and clarify when they outperform simpler stochastic perturbations.

7 Conclusion

Quantum-enhanced and quantum-inspired sequence learning is progressing from isolated proof-of-concept models toward a broader study of representation and training. QLSTM and quantum feature methods show that temporal models can incorporate PQCs, while emerging quantum-inspired optimizers offer a hardware-independent route to exploring difficult loss landscapes. Current evidence, however, is fragmented and rarely demonstrates end-to-end computational advantage. Progress will depend less on quantum terminology than on strong baselines, resource accounting,

reproducibility and mechanism-driven experiments. Under these standards, the field can determine which quantum ideas genuinely improve deep sequence modelling and which remain useful only as conceptual inspiration.

References

[1] Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. Neural Comput., 9: 1735-1780.
[2] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In: Advances in Neural Information Processing Systems, 30.
[3] Biamonte, J., Wittek, P., Pancotti, N., et al. (2017). Quantum machine learning. Nature, 549: 195-202.
[4] Preskill, J. (2018). Quantum computing in the NISQ era and beyond. Quantum, 2: 79.
[5] Cerezo, M., Arrasmith, A., Babbush, R., et al. (2021). Variational quantum algorithms. Nat. Rev. Phys., 3: 625-644.
[6] Chen, S.Y.C., Yoo, S., Fang, Y.L.L. (2022). Quantum long short-term memory. In: Proc. IEEE ICASSP. pp. 8622-8626.
[7] Sajjad, I., AL Sobhi, M. M. (2026). The quantum-inspired adaptive superposition optimization for neural network training. AIMS Math., 11: 243-271.
[8] Pamuk, A.E., Ozdemir, E.K., Kocabay, S.T. (2025). Superpositional gradient descent: harnessing quantum principles for model training. In: IEEE International Conference on Quantum Artificial Intelligence.
[9] Zhu, L., Lu, W. (2026). A quantum-inspired attention integrated scalar long short-term memory model for accurate and stable groundwater contaminant source inversion. Environ. Monit. Assess., 198: 124.
[10] Schuld, M., Sinayskiy, I., Petruccione, F. (2014). The quest for a quantum neural network. Quantum Inf. Process., 13: 2567-2586.
[11] Mitarai, K., Negoro, M., Kitagawa, M., Fujii, K. (2018). Quantum circuit learning. Phys. Rev. A, 98: 032309.
[12] Benedetti, M., Lloyd, E., Sack, S., Fiorentini, M. (2019). Parameterized quantum circuits as machine learning models. Quantum Sci. Technol., 4: 043001.
[13] Schuld, M., Killoran, N. (2019). Quantum machine learning in feature Hilbert spaces. Phys. Rev. Lett., 122: 040504.
[14] Havlicek, V., Corcoles, A.D., Temme, K., et al. (2019). Supervised learning with quantum-enhanced feature spaces. Nature, 567: 209-212.
[15] Bausch, J. (2020). Recurrent quantum neural networks. In: Advances in Neural Information Processing Systems, 33.
[16] Chen, J.H., Huang, Y.C., Tsai, Y.C., Chen, S.Y.C. (2025). Quantum-enhanced forecasting for deep reinforcement learning in algorithmic trading. arXiv:2509.09176.
[17] Mechiche-Alami, N., Rodriguez, E., Cardemil, J.M., Lopez Droguett, E. (2025). Quantum Fourier transform based kernel for solar irradiance forecasting. arXiv:2511.17698.
[18] Nair, S. R., Southwell, B., Ferrie, C. (2025). Short-time quantum Fourier transform processing. In: Proc. IEEE ICASSP. DOI: 10.1109/ICASSP49660.2025.10890646.
[19] McClean, J.R., Boixo, S., Smelyanskiy, V.N., Babbush, R., Neven, H. (2018). Barren plateaus in quantum neural network training landscapes. Nat. Commun., 9: 4812.
[20] Schuld, M., Bergholm, V., Gogolin, C., Izaac, J., Killoran, N. (2019). Evaluating analytic gradients on quantum hardware. Phys. Rev. A, 99: 032331.
[21] Kingma, D.P., Ba, J. (2015). Adam: a method for stochastic optimization. In: International Conference on Learning Representations.
[22] Loshchilov, I., Hutter, F. (2019). Decoupled weight decay regularization. In: International Conference on Learning Representations.
[23] Hu, E. J., Shen, Y., Wallis, P., et al. (2022). LoRA: low-rank adaptation of large language models. In: International Conference on Learning Representations.
[24] Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L. (2023). QLoRA: efficient finetuning of quantized LLMs. In: Advances in Neural Information Processing Systems, 36.
[25] Huang, H.Y., Broughton, M., Mohseni, M., et al. (2021). Power of data in quantum machine learning. Nat. Commun., 12: 2631.
[26] Liu, Y., Arunachalam, S., Temme, K. (2021). A rigorous and robust quantum speed-up in supervised machine learning. Nat. Phys., 17: 1013-1017.
[27] Abbas, A., Sutter, D., Zoufal, C., et al. (2021). The power of quantum neural networks. Nat. Comput. Sci., 1: 403-409.
[28] Bowles, J., Ahmed, S., Schuld, M. (2024). Better than classical? The subtle art of benchmarking quantum machine learning models. arXiv: 2403.07059.
[29] Beck, M., Poppel, K., Spanring, M., et al. (2024). xLSTM: extended long short-term memory. arXiv:2405.04517.
[30] Kong, Y., Wang, Z., Nie, Y., et al. (2025). Unlocking the power of LSTM for long-term time series forecasting. In: Proc. AAAI Conference on Artificial Intelligence.