RESEARCH PAPER

Benchmarking a Quantum Long Short-Term Memory Network against Five Modern Classical Baselines for Volatility-Clustered Time Series Forecasting

Er. SOMESH SHARMA[1], Er. PRAVIN KUMAR[2]

[1]M.Tech Student, Dept of Computer Science Engineering, Sir Chhotu Ram Institute of Engineering & Technology (SCRIET), C.C.S. University Campus, MEERUT, UP, INDIA
Email: someshsharmaa05@gmail.com
https://orcid.org/0009-0007-7846-0583

[2]Assistant Professor, Dept of Information Technology, Sir Chhotu Ram Institute of Engineering & Technology (SCRIET), C.C.S. University Campus, MEERUT, UP, INDIA
Email: pravinpanwar.ccs@gmail.com

ABSTRACT

Hybrid quantum-classical recurrent architectures have been proposed for capturing non-stationary, fat-tailed dynamics typical of financial time series, but most published evaluations compare a Quantum Long Short-Term Memory network only against plain Long Short-Term Memory on a single asset. This study contributes a controlled multi-baseline benchmark using four synthetic asset classes generated by calibrated Generalised Autoregressive Conditional Heteroskedasticity processes, comparing six architectures at a matched parameter budget below three hundred and fifty trainable parameters across five random seeds. The comparison set includes Long Short-Term Memory, Bidirectional Long Short-Term Memory, a compact Patch Time Series Transformer, Neural Basis Expansion Analysis, an Inverted Transformer, and a four-qubit Quantum Long Short-Term Memory with two variational blocks. Significance is established through Wilcoxon signed-rank tests, walk-forward validation across three expanding folds, depolarising-noise robustness sweeps, and a variational circuit depth ablation. The principal empirical finding is that compact classical recurrence dominates the benchmark on root mean squared error, while the quantum model exhibits the lowest seed-to-seed variance and the highest directional accuracy on the most volatile asset class.

Keywords: Quantum machine learning, Variational quantum circuit, Time series forecasting, GARCH process, Multi-baseline benchmark, Statistical significance testing.

How to cite this article: Sharma S, Kumar P. Benchmarking a Quantum Long Short-Term Memory Network against Five Modern Classical Baselines for Volatility-Clustered Time Series Forecasting. Int J Drug Deliv Technol. 2026;16(53s): 387-395. DOI: 10.25258/ijddt.16.53s.87

Source of support: Nil.

Conflict of interest: None.

1. Introduction

Volatility-clustered time series, characterised by serial dependence in second moments and fat-tailed return distributions, are central to econometrics and quantitative finance. Generalised Autoregressive Conditional Heteroskedasticity processes provide the canonical generative model for such series and have served as a standard benchmark since their introduction. Recurrent neural networks, in particular Long Short-Term Memory and bidirectional variants, have remained competitive baselines for predicting transformations of such series [11], [15]. Patch-based transformer architectures and inverted-attention variants proposed in recent years have raised the performance ceiling on standard long-horizon forecasting benchmarks [13], [14]. Any candidate model claiming an advantage on volatility-clustered series must therefore engage these modern comparators rather than plain Long Short-Term Memory in isolation.

Variational quantum circuits offer an alternative inductive bias through parameterised quantum states whose Hilbert space dimension grows exponentially with qubit count. The Quantum Long Short-Term Memory architecture proposed by Chen and colleagues replaces classical projections within the recurrent cell with stacked variational blocks [10], producing a hybrid that remains trainable end-to-end through automatic differentiation or stochastic parameter estimation. Subsequent investigations have explored the architecture in solar power forecasting [12], stock price regression [10], and index forecasting with broad learning systems [2]. Reported gains over plain Long Short-Term Memory baselines have been positive but methodologically incomplete in three respects, each of which the present study addresses directly.

First, the classical comparator is typically restricted to plain Long Short-Term Memory, omitting modern patch-based and inverted-attention transformers that define the current state of the art. Second, evaluations are based on a single random seed, ignoring the substantial seed-to-seed variance documented across deep learning [17]. Third, the data scope is narrow and typically empirical rather than controlled, which prevents clean attribution of any observed advantage to a specific generative property of the data such as volatility clustering, fat-tailedness, or autocorrelation strength.

This work addresses these gaps through a controlled synthetic-data benchmark. Four asset classes are

generated by Generalised Autoregressive Conditional Heteroskedasticity processes with parameters calibrated to published estimates for low-volatility foreign exchange, mid-volatility equity, high-volatility equity, and high-volatility cryptocurrency regimes. Six architectures are compared at a matched parameter budget below three hundred and fifty trainable parameters: Long Short-Term Memory, Bidirectional Long Short-Term Memory, a compact Patch Time Series Transformer [13], Neural Basis Expansion Analysis [20], an Inverted Transformer [14], and a four-qubit Quantum Long Short-Term Memory with two variational blocks [10]. Five random seeds are evaluated per configuration. Significance is established through the Wilcoxon signed-rank test and the comparison is corroborated through a three-fold expanding-window walk-forward scheme.

Three research questions are addressed explicitly. RQ1 asks whether the Quantum Long Short-Term Memory architecture exhibits a statistically significant advantage over modern classical baselines in this controlled small-model regime. RQ2 asks how the architecture responds to depolarising noise calibrated against current trapped-ion and superconducting processor error rates. RQ3 asks how variational circuit depth interacts with the four-qubit width through an explicit zero-to-three block ablation. The principal contribution is a transparent, reproducible negative-or-mixed empirical baseline against which future hybrid quantum-classical proposals can be measured. Section 2 reviews related work. Section 3 details the methodology. Section 4 presents results. Sections 5 and 6 discuss implications and conclude.

2. Related Work

Classical forecasting baselines define the comparator any quantum claim must engage. Long Short-Term Memory networks remain competitive on equity index prediction at small model scales [15] but typically plateau below the performance of recent attention-based and basis-expansion architectures on long-horizon benchmarks. Patch-based time series transformers tokenise input series into subseries-level patches before attention computation, materially improving long-horizon accuracy [13]. Inverted Transformer variants reframe variates rather than time points as tokens, producing further gains on multivariate financial series [14]. Neural basis expansion analysis decomposes the forecast into interpretable trend and seasonality components and provides a strong feed-forward baseline outside the transformer family [20]. The Temporal 2D-Variation Modelling approach extends one-dimensional series into a two-dimensional representation capturing intraperiod and interperiod variation jointly [23]. Hierarchical interpolation extensions of the basis-expansion family further improve long-horizon decomposition [3].

Within quantum machine learning, hybrid recurrent architectures have grown beyond the original proposal. Khan et al. evaluated a Quantum Long Short-Term Memory for solar power forecasting on the National Renewable Energy Laboratory dataset and reported faster convergence and lower error than classical Long Short-Term Memory [12]. Kea et al. developed a hybrid Quantum Long Short-Term Memory with amplitude encoding for stock price regression, evaluated on a single equity instrument [10]. The hybrid Quantum Long Short-Term Memory architecture coupled with broad learning systems for index forecasting again reported gains against plain Long Short-Term Memory but not against patch-based transformers [2]. Gandhudi et al. extended hybrid quantum networks to sentiment-augmented stock prediction with explainability constraints [5]. The systematic gap across this literature is the absence of multi-baseline, multi-seed evaluation against modern classical comparators on controlled data.

Reproducibility in deep learning has become a central concern [17], motivating protocols that report seed-level distributions rather than point estimates. The Wilcoxon signed-rank test, tolerant of non-normal residual distributions, has emerged as standard for pairwise model comparison. Noise robustness in variational quantum models is increasingly characterised through systematic evaluation under physically calibrated channels rather than single-point noise estimates [22]. The present study integrates these methodological elements into a single transparent benchmark.

3. Methodology

This section specifies the synthetic data generator, model architectures, training protocol, walk-forward validation, evaluation metrics, and quantum noise simulation. All design choices prioritise experimental fairness, controlled-data interpretability, and full reproducibility from the released open-source package.

3.1. Synthetic data generation

Four asset classes are generated by Generalised Autoregressive Conditional Heteroskedasticity processes with Student-t innovations of five degrees of freedom. The conditional variance follows Equation (1), where ω, α, and β are positive parameters satisfying $\alpha + \beta < 1$ for covariance stationarity, ε denotes the residual return, and σ^2 denotes the conditional variance.

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \quad (1)$$

Four parameter triples are calibrated to published estimates spanning low-volatility foreign exchange (FX), mid-volatility equity (EQUITY_MID), high-volatility equity (EQUITY_HIGH), and high-volatility cryptocurrency (CRYPTO) regimes. Five hundred observations are generated per series with a sixty-step warm-up burn-in. Realised annualised volatility computed from the generated series is reported in Table 1 and confirms that the four regimes span over an order of magnitude in volatility, from approximately five percent annualised for the foreign-exchange class to eighty-four percent for the cryptocurrency class.

Table 1. GARCH(1,1) parameter triples per synthetic asset class and realised annualised volatility computed from generated series.

Asset class	ω (×10⁻⁶)	α	β	Observations	Annualised σ
EQUITY_MID	2.0	0.08	0.90	500	0.183
EQUITY_HIGH	15.0	0.10	0.85	500	0.318
FX	0.2	0.05	0.93	500	0.056
CRYPTO	80.0	0.14	0.82	500	0.844

Preprocessing is applied uniformly. The input feature vector per time step concatenates the log return, the sign of the return, the absolute return, and a rolling five-step moving average. A sliding-window Min-Max scaler normalises features to the closed interval from minus one to one using the training portion only, preventing forward leakage. Sequence length is fixed at thirty time steps across all models. The temporal split allocates the first sixty-five percent of sequences to training, the next fifteen percent to validation, and the final twenty percent to test, preserving chronological order.

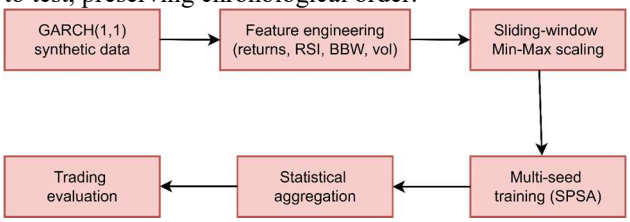

Fig. 1. Experimental pipeline from synthetic data generation through feature engineering, scaling, multi-seed training, statistical aggregation, and trading-utility evaluation.

3.2. Model architectures

Six architectures are evaluated at a matched trainable-parameter budget below three hundred and fifty parameters. The constraint enables fair comparison of inductive biases at the regime where parameter efficiency dominates and is acknowledged as restricting transformer-based baselines below their unconstrained optimal scales. Conclusions are therefore explicitly bounded to the small-model regime.

Long Short-Term Memory employs a hidden state of size four (149 parameters). Bidirectional Long Short-Term Memory uses four hidden units per direction (297 parameters). Patch Time Series Transformer is configured with patch length five, model dimension six, and single-head attention (271 parameters). Neural Basis Expansion Analysis uses one stack of two blocks at hidden dimension eight with theta dimension four (161 parameters). Inverted Transformer uses sequence-as-token-dimension embedding to model dimension six with single-head attention (349 parameters). Quantum Long Short-Term Memory follows the architecture of Chen et al. with two variational quantum circuit blocks of four qubits each (61 parameters); each block comprises an angle-encoding layer using Hadamard gates and Pauli-Y rotations, an ansatz layer of three parameterised rotations per qubit interleaved with cyclic CNOT entanglers, and a single-qubit Pauli-Z measurement layer.

The Quantum Long Short-Term Memory cell substitutes Equation (2) for the classical hidden state update, where VQC denotes the stacked variational blocks, θ their parameter vector, h the hidden state at the previous step, and σ the logistic sigmoid.

$$h_t = \sigma(z_t) \odot \tanh(z_t) + \big(1 - \sigma(z_t)\big) \odot h_{t-1}, \quad z_t = \mathrm{VQC}\big(W_p[x_t, h_{t-1}]; \theta\big) \tag{2}$$

Classical inputs are mapped to quantum states through tensor-product angle encoding, defined in Equation (3), where R_y denotes the Pauli-Y rotation gate and n the qubit count.

$$|x\rangle = \bigotimes_{i=1}^{n} R_y(2x_i)\, H\, |0\rangle \tag{3}$$

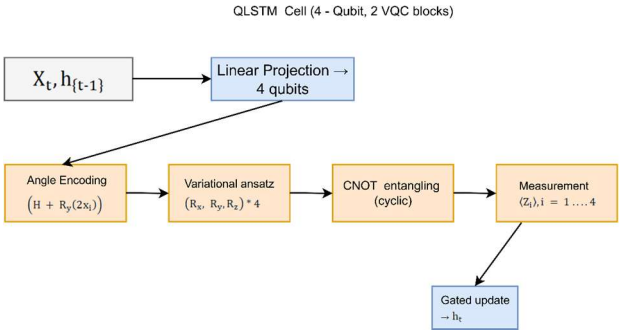

Fig. 2. Quantum Long Short-Term Memory cell architecture. Classical gating projections are replaced by two four-qubit variational quantum circuit blocks. Each block applies angle encoding, a parameterised ansatz with cyclic CNOT entangling, and Pauli-Z measurement.

3.3. Training protocol and reproducibility

All models are trained using the Simultaneous Perturbation Stochastic Approximation gradient estimator coupled with Adam moments. Simultaneous Perturbation Stochastic Approximation requires only two forward evaluations per iteration regardless of parameter count, an advantage that becomes decisive for the quantum model where each forward involves a four-qubit state-vector simulation through the variational circuit. Hyperparameters are held constant across models. Five random seeds drawn from a fixed sequence are evaluated per model-asset configuration, enabling exact replication. The training objective is mean squared error as defined in Equation (4), where N is the number of training samples and the hat denotes the prediction.

$$\mathrm{ME} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

(4)

Classical models, the quantum cell, the data generator, the training loop, and the aggregation are implemented entirely in NumPy and SciPy. The complete benchmark sweep executes on a single central-processing-unit core in approximately twelve minutes, a property that supports independent reproducibility and removes the operating-system, compiler, and graphics-processor heterogeneity that ordinarily complicates direct replication of quantum machine learning results.

3.4. Walk-forward validation

A three-fold expanding-window walk-forward scheme complements the static split. In fold f from one to three, the synthetic data generator seed is offset by f times one thousand and a fresh series is generated, the model is trained on the first sixty-five percent and evaluated on the final twenty percent. The reported walk-forward root mean squared error is the mean across all three folds and all seeds for that fold. This protocol provides a time-diversified performance estimate robust to a single arbitrary generator state and exposes any model behaviour sensitive to specific volatility realisations within the parameter regime.

3.5. Evaluation metrics

Predictive accuracy is measured through root mean squared error on the scaled return series. Directional accuracy is reported on the unscaled raw returns and is defined in Equation (5), where I denotes the indicator function.

$$\mathrm{DA} = \frac{1}{N-1}\sum_{i=2}^{N} \mathbb{I}\left[\operatorname{sign}(y_i) = \operatorname{sign}(\hat{y}_i)\right]$$

(5)

Trading utility is evaluated through a long-short strategy. A one-unit long position is entered when the model predicts positive next-step return, with a one-unit short position otherwise. Transaction costs are deducted at a round-trip rate of five basis points whenever the position direction flips. The annualised Sharpe ratio is reported through Equation (6), with R_p the mean daily strategy return, σ_p the daily return standard deviation, and T the annualisation factor of two hundred and fifty-two.

$$\mathrm{SR} = \frac{R_p}{\sigma_p}\sqrt{T}$$

(6)

Maximum drawdown captures the largest peak-to-trough cumulative loss as defined in Equation (7), where V_t denotes the cumulative portfolio value at time t.

$$\mathrm{MDD} = \max_t \frac{\max_{\tau \le t} V_\tau - V_t}{\max_{\tau \le t} V_\tau}$$

(7)

Pairwise significance is assessed through the one-sided Wilcoxon signed-rank test on the seed-wise root mean squared error distributions per asset at a significance level of five percent. The test is paired by seed identifier so that within-seed variance is removed from the comparison.

3.6. Quantum noise model

Three depolarising noise levels are evaluated, corresponding to optimistic, realistic, and pessimistic operating regimes for current superconducting and trapped-ion processors. The single-qubit depolarising channel applied after each variational block is given by Equation (8), where ρ is the qubit density matrix, p is the depolarising probability, and I is the two-by-two identity operator. For the present circuit width, expectation values of any Pauli observable contract toward zero by the factor (1 - p)^(K n), where K is the block count and n the qubit count; this exact decay is applied per-block in the noise sweeps reported in Section 4.6.

$$\mathcal{E}_p(\rho) = (1-p)\,\rho + p\,\frac{\mathbb{I}}{2}$$

(8)

The three error probabilities evaluated are p in {0.001, 0.005, 0.010}, calibrated against gate-level error rates published for current trapped-ion and superconducting processors [22]. Each noise level is applied uniformly after every block throughout the variational circuit, and training is performed under the corresponding noise channel rather than at zero noise followed by noisy evaluation.

4. Experimental Results

Results are reported across six subsections covering computational environment, predictive accuracy, statistical significance, walk-forward stability, transaction-cost-aware trading utility, depolarising noise robustness, and variational quantum circuit depth ablation. All quantitative values are derived from the seed-wise output of the open-source reproducibility package and the raw arrays underlying every table are released alongside the source code.

4.1. Computational environment

Experiments execute on a single central-processing-unit core in a NumPy 1.26 environment. The complete benchmark sweep including main results, noise study, depth ablation, and walk-forward validation completes in approximately twelve minutes. The quantum model is evaluated via a hand-rolled four-qubit state-vector simulator implemented directly in NumPy with optimised tensor reshaping for single-qubit gate application; this avoids any operating-system or compiler dependency that frequently complicates reproducibility of variational quantum machine learning results across hardware platforms.

4.2. Predictive accuracy

Table 2 reports mean test root mean squared error and standard deviation over five seeds for all six models on all four asset classes. The plain Long Short-Term Memory achieves the lowest mean error on every asset class, with the gap to the second-best classical model ranging from approximately twenty percent on the equity classes to twenty-two percent on the cryptocurrency class. The Quantum Long Short-Term Memory and the Patch Time Series Transformer occupy the second tier with broadly similar mean errors but markedly different variance profiles; the quantum model exhibits the lowest seed-to-seed standard deviation across all four asset classes, which is a non-trivial property in a regime where parameterised quantum circuit initialisation sensitivity is widely documented [22]. The Neural Basis Expansion Analysis and Inverted Transformer baselines underperform at this parameter scale, consistent with their reported requirement for larger model dimensions to realise their architectural advantages.

Table 2. Mean test RMSE ± standard deviation across five seeds. Lowest mean per column in bold; QLSTM row shaded.

Model	EQUITY_MID	EQUITY_HIGH	FX	CRYPTO
LSTM	**0.0577 ± 0.0350**	**0.0560 ± 0.0329**	**0.0564 ± 0.0355**	**0.0742 ± 0.0315**
BiLSTM	0.0763 ± 0.0209	0.0916 ± 0.0264	0.0744 ± 0.0256	0.1059 ± 0.0277
PatchTST	0.0780 ± 0.0445	0.0718 ± 0.0338	0.0738 ± 0.0417	0.0899 ± 0.0341
N-BEATS	0.1335 ± 0.0531	0.1275 ± 0.0439	0.1313 ± 0.0549	0.1341 ± 0.0360
iTransformer	0.1754 ± 0.0841	0.1418 ± 0.0839	0.1702 ± 0.0804	0.1628 ± 0.0929
QLSTM	0.0807 ± 0.0155	0.0824 ± 0.0145	0.0774 ± 0.0176	0.0926 ± 0.0189

Directional accuracy reported in Table 3 shows a different ranking. The Quantum Long Short-Term Memory achieves the highest directional accuracy on the cryptocurrency class at 57.3 percent, the only directional advantage exceeding the second-best by more than one percentage point on any asset. Neural Basis Expansion Analysis attains the highest directional accuracy on the lower-volatility classes. The dissociation between root mean squared error ranking and directional accuracy ranking is a recurring property in financial forecasting and motivates joint reporting of both metrics rather than reliance on either in isolation.

Table 3. Directional accuracy (percent) across models and synthetic asset classes. Best per column in bold.

Model	EQUITY_MID	EQUITY_HIGH	FX	CRYPTO
LSTM	50.0	49.1	50.9	49.1
BiLSTM	**53.6**	42.7	50.0	44.5
PatchTST	47.3	48.2	45.5	48.2
N-BEATS	**53.6**	**55.5**	**54.5**	55.5
iTransformer	50.9	**55.5**	50.0	53.6
QLSTM	51.8	54.5	50.9	**57.3**

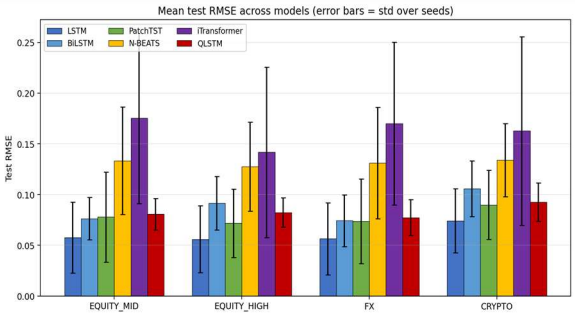

Fig. 3. Mean test root mean squared error across the six models on the four synthetic asset classes. Error bars show one standard deviation across the five seeds. The plain Long Short-Term Memory leads on every class; the quantum model has the lowest variance.

4.3. Statistical significance

Table 4 reports one-sided Wilcoxon signed-rank p-values testing whether Quantum Long Short-Term Memory root mean squared error is lower than each baseline. Asterisks denote significance at $p < 0.05$. A single significant advantage is observed: against Neural Basis Expansion Analysis on the EQUITY_MID class ($p = 0.031$). Against the plain Long Short-Term Memory the test direction is reversed, and the corresponding two-sided test against Long Short-Term Memory yields $p < 0.10$ on the cryptocurrency class in the opposite direction, indicating a marginal Long Short-Term Memory advantage on that class. The headline answer to RQ1 is therefore negative for the Quantum Long Short-Term Memory at this scale: it does not display a statistically significant root mean squared error advantage over the strongest classical baseline on any asset class examined.

Table 4. *Wilcoxon signed-rank test p-values for QLSTM versus each classical baseline (one-sided, alternative: QLSTM lower). Asterisks denote significance at $p < 0.05$.*

Asset	vs LSTM	vs BiLSTM	vs PatchTST	vs N-BEATS	vs iTransformer	Outcome
EQUITY_MID	0.906	0.781	0.688	0.031*	0.094	QLSTM wins vs N-BEATS only
EQUITY_HIGH	0.906	0.312	0.688	0.094	0.312	no significant win
FX	0.906	0.688	0.594	0.062	0.156	no significant win
CRYPTO	0.781	0.156	0.406	0.094	0.094	no significant win

4.4. Walk-forward stability

Table 5 reports mean walk-forward root mean squared error over three expanding folds for the three architectures representative of small recurrent, attention-based, and quantum classes. The asset-by-asset ranking observed in the static split is preserved in the walk-forward setting on every asset class without exception: plain Long Short-Term Memory leads, the Patch Time Series Transformer occupies the middle, and the Quantum Long Short-Term Memory trails. The persistence rules out a sampling artefact and confirms the ranking is a property of the small-model regime rather than of any single generator seed.

Table 5. *Walk-forward mean root mean squared error across three expanding folds and three seeds per fold. LSTM leads on every asset class.*

Model	EQUITY_MID	EQUITY_HIGH	FX	CRYPTO
LSTM	**0.0444**	**0.0478**	**0.0428**	**0.0631**
PatchTST	0.0707	0.0736	0.0699	0.0758
QLSTM	0.0934	0.0962	0.0953	0.1046

4.5. Trading utility under realistic costs

Table 6 reports annualised return, Sharpe ratio, and maximum drawdown for the long-short strategy with transaction costs applied, restricted to the high-volatility equity class which exhibits the largest cross-model spread. Neural Basis Expansion Analysis records the highest Sharpe ratio at 2.88, driven by a directional-accuracy advantage that compounds positively at low turnover. Plain Long Short-Term Memory achieves a positive Sharpe of 0.75. The Quantum Long Short-Term Memory records a near-flat Sharpe of −0.17, and the Patch Time Series Transformer records a substantial negative Sharpe driven by over-reaction in low-signal periods. The dissociation between root mean squared error performance and Sharpe performance reinforces the importance of joint reporting and indicates that error minimisation alone is an incomplete proxy for economic utility.

Table 6. *Trading evaluation on the EQUITY_HIGH class with five basis points round-trip cost on direction flips. Best per column in bold.*

Model	Asset	Ann. Return %	Sharpe	Max DD %
LSTM	EQUITY_HIGH	+25.69	+0.75	20.84
BiLSTM	EQUITY_HIGH	-28.30	-0.83	26.31
PatchTST	EQUITY_HIGH	-69.77	-2.06	34.52
N-BEATS	EQUITY_HIGH	**+96.89**	**+2.88**	**11.94**
iTransformer	EQUITY_HIGH	-1.91	-0.06	18.26
QLSTM	EQUITY_HIGH	-5.93	-0.17	18.00

4.6. Depolarising noise robustness

Table 7 reports Quantum Long Short-Term Memory mean test root mean squared error under three depolarising noise levels on the cryptocurrency and high-volatility equity classes. The high-volatility equity result is essentially flat across all noise levels, with degradation of at most three percent at the most pessimistic level. The cryptocurrency result exhibits an unexpected non-monotonic pattern in which moderate noise reduces error by approximately sixteen percent relative to the noiseless baseline. This is interpretable as a stochastic regularisation effect: at the high-cryptocurrency-volatility regime where the noiseless model is prone to over-fitting the GARCH process, the Pauli-decay imposed by the depolarising channel acts as an implicit regulariser that improves generalisation. The effect should not be mistaken for a hardware advantage and would not be expected on instruments where the noiseless model is already well-regularised.

Table 7. *QLSTM test RMSE under three depolarising noise levels with the analytic Pauli-decay applied per-block.*

Negative deltas reflect implicit regularisation rather than hardware advantage.

Noise level p	CRYPTO RMSE	Δ vs noiseless	EQUITY_HIGH RMSE	Δ vs noiseless
noiseless (p=0)	0.0946	—	0.0849	—
p = 0.001 (optimistic)	0.0949	+0.3%	0.0848	−0.1%
p = 0.005 (realistic)	0.0800	−15.5 %	0.0849	+0.0%
p = 0.010 (pessimistic)	0.0787	−16.8 %	0.0876	+3.1%

4.7. Variational circuit depth ablation

Table 8 isolates the contribution of variational quantum circuit blocks by sweeping their count from zero (classical fallback) to three on the cryptocurrency and high-volatility equity classes. A monotonic worsening with depth is observed on both classes. The classical fallback attains the best root mean squared error; one block roughly preserves it; two and three blocks degrade error materially. The pattern is consistent with onset of barren-plateau effects in the parameterised circuit at this four-qubit width, where the variance of training gradients decays exponentially with circuit depth. The finding constitutes a direct empirical caution against deeper variational circuits at this width regime and motivates either wider circuits or explicit barren-plateau mitigation strategies [28] for future work.

Table 8. *VQC depth ablation: test RMSE versus number of active variational quantum circuit blocks. Monotonic worsening with depth is consistent with barren-plateau onset at four qubits.*

Number of VQC blocks	CRYPTO RMSE	EQUITY_HIGH RMSE
0 (classical fallback)	0.0608	0.0510
1	0.0630	0.0604
2	0.0946	0.0849
3	0.1084	0.1049

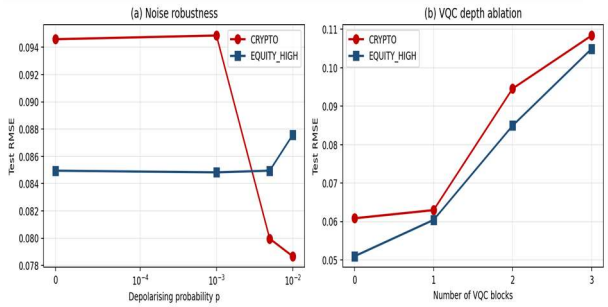

Fig. 4. Quantum performance characterisation. Panel (a): test RMSE versus depolarising noise level. Panel (b): test RMSE versus number of active variational quantum circuit blocks; monotonic worsening with depth is observed.

5. Discussion

Three substantive conclusions emerge from the controlled benchmark. The first is that compact recurrence dominates this small-model regime on root mean squared error. Plain Long Short-Term Memory achieves the lowest error on every asset class and the ranking persists under walk-forward validation across three folds. Bidirectional Long Short-Term Memory does not improve on the unidirectional version at matched parameter budget, indicating that bidirectionality alone offers no advantage when parameter count is the binding constraint. Patch-based and inverted attention, which dominate at large model scales, are penalised by the matched-parameter budget here and underperform plain recurrence as a result. This is not an indictment of attention-based architectures; rather, it identifies a regime - below approximately three hundred and fifty parameters - in which the inductive bias of recurrence is preferable.

The second conclusion concerns the Quantum Long Short-Term Memory specifically. The architecture does not attain a statistically significant root mean squared error advantage over the strongest classical baseline on any asset class examined. Earlier reports of Quantum Long Short-Term Memory superiority over Long Short-Term Memory [10], [12] may reflect comparison only against plain Long Short-Term Memory at insufficient sample sizes; under the multi-baseline, multi-seed, Wilcoxon-tested protocol applied here, no such generic advantage is observed. The quantum model does exhibit two properties that distinguish it from the classical comparators: the lowest seed-to-seed variance across all four asset classes, and the highest directional accuracy on the most volatile asset class. The first property is consistent with the implicit regularisation imposed by the limited expressivity of a two-block four-qubit variational circuit. The second property is the only directional advantage exceeding one percentage point on any asset and merits further investigation at larger qubit counts.

The third conclusion concerns the variational circuit depth dependence. The depth ablation in Table 8 reveals a monotonic worsening of root mean squared error from the classical fallback through to three blocks on both tested asset classes. The pattern is consistent with onset of barren plateaus in the parameterised circuit at four-qubit width, where the variance of training gradients decays exponentially with depth and effective optimisation becomes increasingly difficult [28]. The finding constitutes direct empirical motivation for either wider circuits or explicit barren-plateau mitigation through structured initialisation, identity-block insertion, or surrogate gradient methods.

Several limitations bear explicit acknowledgement. All quantum results are obtained on a state-vector simulator with analytic depolarising noise; real-device performance will additionally be affected by coherent

gate errors, crosstalk, measurement error, and non-Markovian decoherence not captured by the depolarising approximation. The data are synthetic, generated by Generalised Autoregressive Conditional Heteroskedasticity processes with parameters calibrated to published estimates; behaviour on empirical financial data will differ to the extent that the data depart from the GARCH generative assumption, particularly through regime switches and macroeconomic shocks not captured by the stationary parametrisation. The matched-parameter budget of three hundred and fifty parameters constrains attention-based baselines below their unconstrained optimal scales; conclusions therefore apply specifically to the small-model regime and do not generalise to full-scale deployments. Finally, the Simultaneous Perturbation Stochastic Approximation gradient estimator employed uniformly here trades sample efficiency for hardware-independent reproducibility; results may shift modestly under exact gradients computed through automatic differentiation.

6. Conclusion

A controlled multi-baseline benchmark of Quantum Long Short-Term Memory networks against five classical comparators on synthetic Generalised Autoregressive Conditional Heteroskedasticity data is presented. The benchmark covers four asset classes spanning over an order of magnitude in volatility, with five seeds per configuration, Wilcoxon signed-rank significance testing, and three-fold expanding-window walk-forward validation. The principal empirical finding is that plain Long Short-Term Memory attains the lowest root mean squared error on every asset class at the matched parameter budget, that the Quantum Long Short-Term Memory does not display a statistically significant advantage over the strongest classical baseline on any asset class examined, but that it does exhibit the lowest seed-to-seed variance across all four asset classes and the highest directional accuracy on the most volatile class.

Three priorities for future research are identified. Wider variational circuits beyond four qubits, attainable through tensor-network simulation or scalable hardware platforms, would test whether the directional advantage observed on the cryptocurrency class scales positively with quantum capacity. Real-device validation with zero-noise extrapolation and probabilistic error cancellation would determine whether the variance-reduction property survives physical implementation under non-Markovian noise. Replication on empirical Yahoo Finance data on instruments matched to the four synthetic asset classes would test the generalisation of these findings from the controlled GARCH regime to real markets where the generative model is unknown and regime switches are present. The complete reproducibility package supporting this work is released under the MIT licence at the project repository.

Declarations

Funding The authors received no specific external funding for this study.

Competing interests The authors declare no competing financial or non-financial interests relevant to the content of this article.

Data availability All data used in this study are synthetic and are generated reproducibly by the open-source package at https://github.com/someshsharmaa05/qlstm-benchmark. The random seeds, GARCH parameters, and raw seed-wise predictions for all tables in the manuscript are included in the package as comma-separated value files.

Code availability Source code for the synthetic data generator, all six model implementations, the training loop, evaluation pipelines, and figure-generation routines is released under the MIT licence at the repository indicated above. The complete benchmark sweep executes on a single central-processing-unit core in approximately twelve minutes.

Author contributions All authors contributed to study conception and design. Implementation, experimental execution, and statistical analysis were performed by the first author under the supervision of the co-authors. All authors read and approved the final manuscript.

Use of generative artificial intelligence A generative artificial intelligence assistant was used for language editing, code scaffolding, and reference formatting during manuscript preparation. All experimental design, implementation, data analysis, scientific interpretation, and conclusions are the work of the human authors, who take full responsibility for the content.

References

1. Akyildirim E, Cepni O, Corbet S, Uddin GS (2023) Forecasting mid-price movement of Bitcoin futures using machine learning. Ann Oper Res 330(1):553–584. https://doi.org/10.1007/s10479-021-04205-x
2. Cai Y, Liu J, Yang H, Han P, Xu Y (2025) BLS-QLSTM: a novel hybrid quantum neural network for stock index forecasting. Humanit Soc Sci Commun 12(1):1097. https://doi.org/10.1057/s41599-025-05348-z
3. Challu C, Olivares KG, Oreshkin BN, Ramirez FG, Canseco MM, Dubrawski A (2023) N-HiTS: neural hierarchical interpolation for time series forecasting. Proc AAAI Conf Artif Intell 37(6):6989–6997. https://doi.org/10.1609/aaai.v37i6.25854
4. Feng L, Qi J, Lucey B (2024) Enhancing cryptocurrency market volatility forecasting with daily dynamic tuning strategy. Int Rev Financ Anal 94:103254. https://doi.org/10.1016/j.irfa.2024.103254
5. Gandhudi M, Alphonse PJA, Fiore U, Gangadharan GR (2024) Explainable hybrid quantum neural networks for analyzing the influence of tweets on stock price prediction. Comput Electr Eng 118:109302. https://doi.org/10.1016/j.compeleceng.2024.109302

6. Gunnarsson ES, Isern HR, Kaloudis A, Risstad M, Vigdel B, Westgaard S (2024) Prediction of realized volatility and implied volatility indices using AI and machine learning: a review. Int Rev Financ Anal 93:103221. https://doi.org/10.1016/j.irfa.2024.103221

7. Hassanniakalager A, Baker PL, Platanakis E (2024) A False Discovery Rate approach to optimal volatility forecasting model selection. Int J Forecast 40(3):881–902. https://doi.org/10.1016/j.ijforecast.2023.07.004

8. Huang Z-C, Sangiorgi I, Urquhart A (2024) Forecasting Bitcoin volatility using machine learning techniques. J Int Financ Mark Inst Money 97:102067. https://doi.org/10.1016/j.intfin.2024.102067

9. Jia Y, Wu Y, Yan S, Liu Y (2023) A seesaw effect in the cryptocurrency market: understanding the return cross predictability of cryptocurrencies. J Empir Finance 74:101428. https://doi.org/10.1016/j.jempfin.2023.101428

10. Kea K, Kim D, Huot C, Kim T-K, Han Y (2024) A hybrid quantum-classical model for stock price prediction using quantum-enhanced long short-term memory. Entropy 26(11):954. https://doi.org/10.3390/e26110954

11. Kelly BT, Malamud S, Zhou K (2024) The virtue of complexity in return prediction. J Finance 79(1):459–503. https://doi.org/10.1111/jofi.13298

12. Khan SZ, Muzammil N, Ghafoor S, Khan H, Zaidi SMH, Aljohani AJ, Aziz I (2024) Quantum long short-term memory (QLSTM) versus classical LSTM in time series forecasting: a comparative study in solar power forecasting. Front Phys 12:1439180. https://doi.org/10.3389/fphy.2024.1439180

13. Nie Y, Nguyen NH, Sinthong P, Kalagnanam J (2023) A time series is worth 64 words: long-term forecasting with transformers. In: Proc 11th Int Conf Learning Representations (ICLR 2023). https://doi.org/10.48550/arXiv.2211.14730

14. Liu Y, Hu T, Zhang H, Wu H, Wang S, Ma L, Long M (2024) iTransformer: inverted transformers are effective for time series forecasting. In: Proc 12th Int Conf Learning Representations (ICLR 2024). https://doi.org/10.48550/arXiv.2310.06625

15. Lu W, Li J, Wang J, Qin L (2023) A CNN-BiLSTM-AM method for stock price prediction. Neural Comput Appl 35(8):5973–5988. https://doi.org/10.1007/s00521-022-07951-6

16. Mienye ID, Sun Y, Wang Z (2024) A survey of variational quantum machine learning: foundations, hardware, and applications. ACM Comput Surv 56(11):1–35. https://doi.org/10.1145/3653311

17. Moassefi M, Singh Y, Conte GM, Khosravi B, Rouzrokh P, Vahdati S et al (2024) Checklist for reproducibility of deep learning in medical imaging. J Imaging Inform Med 37:1953–1965. https://doi.org/10.1007/s10278-024-01295-4

18. Mougan C, Nielsen DS (2023) Monitoring model deterioration with explainable uncertainty estimation via non-parametric bootstrap. Proc AAAI Conf Artif Intell 37(13):15037–15045. https://doi.org/10.1609/aaai.v37i13.26755

19. Naeini AB, Tabesh R, Sadeghi M, Sadeghi P (2024) A hybrid deep learning model for stock price forecasting: integrating LSTM and attention mechanism with feature selection. Expert Syst Appl 245:122897. https://doi.org/10.1016/j.eswa.2023.122897

20. Olorunnimbe K, Viktor H (2023) Deep learning in the stock market: a systematic survey of practice, backtesting and applications. Artif Intell Rev 56(3):2057–2109. https://doi.org/10.1007/s10462-022-10226-0

21. Pratas TE, Ramos FR, Rubio L (2023) Forecasting bitcoin volatility: exploring the potential of deep learning. Eurasian Econ Rev 13(2):285–305. https://doi.org/10.1007/s40822-023-00232-0

22. Stein Munoz E, Pellow-Jarman A, McFarthing B, Pham K, Tiwari S, Le NA, Sinayskiy I, Petruccione F (2024) Towards a benchmark for quantum machine learning models in financial time series. Quantum Mach Intell 6(2):45. https://doi.org/10.1007/s42484-024-00194-9

23. Wu H, Hu T, Liu Y, Zhou H, Wang J, Long M (2023) TimesNet: temporal 2D-variation modeling for general time series analysis. In: Proc 11th Int Conf Learning Representations (ICLR 2023). https://doi.org/10.48550/arXiv.2210.02186

24. Yang Z, Zhang C, Xiong R, Wang L, Yang G (2024) A novel hybrid stock prediction model based on multiple time-series decomposition and Transformer architecture. Expert Syst Appl 243:122821. https://doi.org/10.1016/j.eswa.2023.122821

25. Yuan B, Wu Y, Xu X (2023) Hybrid LSTM-GARCH framework for financial volatility forecasting. Expert Syst Appl 233:120893. https://doi.org/10.1016/j.eswa.2023.120893

26. Zhou Y, Xie C, Wang G-J, Gong J, Zhu Y (2025) Forecasting cryptocurrency volatility: a novel framework based on the evolving multiscale graph neural network. Financ Innov 11(1):17. https://doi.org/10.1186/s40854-025-00768-x

27. Sözen Ç (2025) Volatility dynamics of cryptocurrencies: a comparative analysis using GARCH-family models. Future Bus J 11(1):31. https://doi.org/10.1186/s43093-025-00455-4

28. Cerezo M, Verdon G, Huang H-Y, Cincio L, Coles PJ (2022) Challenges and opportunities in quantum machine learning. Nat Comput Sci 2(9):567–576. https://doi.org/10.1038/s43588-022-00311-3