Can Machines Remember Markets? Long Memory, Market Efficiency, and Neural Forecasting in International Equity Markets

Dominik Owczarek[a], Małgorzata Snarska[a,*]

[a]*Cracow University of Economics, Kraków, Poland*

Abstract

This paper asks whether machine-learning models used in international equity-market forecasting learn the dependence structure that matters for market efficiency, volatility persistence and institutional risk control, or merely minimise pointwise forecast errors. The distinction is central because weak-form efficiency concerns directional predictability in returns, whereas market-risk systems depend on the persistence of return magnitudes. We develop a two-hypothesis framework in which the fractional integration parameter d is the object to be tested, preserved and estimated. The first hypothesis separates directional efficiency from volatility memory: raw returns should be short-memory, while absolute returns should display positive fractional integration. The second hypothesis states that conventional machine-learning accuracy is not memory fidelity: models with low mean-squared error or acceptable directional accuracy may still fail to reproduce d. Using daily data for the S&P 500, DAX30 and six large U.S. stocks over 2010–2025, synthetic fractionally integrated experiments, and the feasible exact local Whittle estimator as the econometric benchmark, we show that raw returns are short-memory or mildly anti-persistent, but absolute returns exhibit economically large long memory. For the S&P 500, $d(|r|)$ ranges from 0.305 to 0.498 across bandwidths; for the DAX30, it ranges from 0.412 to 0.447. The model comparison is equally clear. Linear models achieve low one-step errors but fail the memory-preservation test; recurrent architectures remain stable in recursive evaluation but often impose excessive persistence; the best real-market result is obtained by the embedding-attention LSTM, which matches held-out S&P 500 absolute-return memory with an average absolute d error of 0.048. The paper contributes a memory-fidelity validation layer for financial AI and shows why market-efficiency, volatility-forecasting and model-risk research should report not only how accurately a model predicts, but also what kind of market memory it reproduces.

Keywords: market efficiency, long memory, machine learning, international equity markets, FELW, LSTM, attention mechanisms, volatility persistence

JEL: C22, C45, C58, G14, G15

*Corresponding author.

1. Introduction

The empirical finance literature has long treated predictability as a test of market efficiency. If past prices help forecast future returns after costs and risk adjustment, weak-form efficiency is called into question (Fama, 1970; Lo, 1991). Yet this framing is incomplete for modern equity markets. International stock-index returns are difficult to predict in sign, but the magnitude of returns is persistent, clustered and often fractionally integrated (Mandelbrot, 1963; Ding et al., 1993). This creates a problem for machine-learning finance. A model can achieve low pointwise forecast error while failing to reproduce the long-range dependence that governs volatility, risk and market-state persistence. Conversely, a model may learn a stable memory representation while overshooting the true persistence of the data-generating process.

This paper reframes the use of machine learning in market-efficiency research around a single diagnostic question: *can machines remember markets in the same way that markets remember risk?* The question is not whether ML can improve one-step-ahead forecasts in a narrow predictive contest. It is whether the forecasts and learned representations preserve the fractional integration parameter d, which summarises the slow hyperbolic decay of dependence. In the weak-form EMH benchmark, raw returns should not contain persistent directional memory. In volatility and risk applications, however, absolute returns may exhibit long memory without contradicting weak-form efficiency in levels. A model that confuses these two objects risks producing attractive forecast metrics and misleading financial interpretation.

The paper is written for an international financial-markets audience for three reasons. First, the empirical setting compares a U.S. benchmark market, the S&P 500, with a large European benchmark, the DAX30, and examines whether the directional-efficiency/volatility-memory distinction is visible across both markets. Second, the paper connects market-efficiency testing with model-risk diagnostics for institutions that increasingly use machine-learning forecasts in market-risk, asset-allocation and volatility-monitoring systems. Third, it shows that the commonly reported criteria in ML finance - mean squared error and directional accuracy - are not sufficient for judging whether a model has learned an economically meaningful dependence structure.

The empirical design is deliberately cumulative. We first estimate the memory properties of daily log returns and absolute log returns using the feasible exact local Whittle estimator of Shimotsu and Phillips (2005); Shimotsu (2010). We then conduct two tests. The Type I test asks whether the predicted series produced by a trained model preserves a known target memory $d = 0.3$ under open-loop and closed-loop evaluation. The Type II test asks whether a model can estimate d directly from an unseen series, using the FELW estimate as the benchmark. Finally, the leading architectures are applied to the held-out S&P 500 absolute-return series to test whether synthetic performance transfers to real equity-market data.

The results are unambiguous. Raw log returns for all eight series have d estimates close to zero or negative, consistent with no persistent directional predictability. Absolute returns have positive and statistically meaningful long memory. For the S&P 500, $d(|r|)$ equals 0.305, 0.446

and 0.498 for bandwidth exponents $\delta = 0.5, 0.6, 0.7$; for the DAX30, the corresponding values are 0.447, 0.412 and 0.413. This is the empirical regularity that the ML models must reproduce.

They mostly do not. In the Type I experiment, linear regression and ridge deliver the lowest open-loop MSE values, but estimate d at approximately 0.486, outside the acceptance band around the target. Regression trees and standalone attention pass in open loop but collapse in closed loop. Recurrent and hybrid attention-LSTM models are much more stable across modes, but they systematically overstate persistence, with d estimates generally between 0.78 and 0.93. In the Type II experiment, the best ML estimator is a four-layer DNN with an acceptance rate of 46.3%, followed by a five-layer DNN and an Attention-LSTM, far below the approximate 82% FELW self-consistency benchmark. In the real-market application, however, the embedding-attention LSTM stands out: its average absolute error across FELW bandwidths is only 0.048.

The contribution is threefold. First, the paper separates two ideas often conflated in ML-finance applications: pointwise forecasting accuracy and memory fidelity. Second, it proposes a direct diagnostic framework for evaluating whether ML forecasts preserve the long-memory structure of financial data. Third, it provides a finance interpretation of architectural behaviour: shallow and linear models smooth without remembering, recurrent models remember too much, and embedding-attention hybrids appear best suited to balancing selective memory with volatility-regime adaptation.

The remainder of the paper is organised as follows. Section 2 develops the economic mechanism and hypotheses. Section 3 describes the data and empirical design. Section 4 reports the long-memory evidence and Type I and Type II tests. Section 5 reports the real-market application. Section 6 discusses implications for international financial institutions and model-risk governance. Section 7 concludes.

2. Literature Review and Hypothesis Development

2.1. Market efficiency, fractional integration and the economic meaning of volatility memory

The first strand of literature motivating the paper is the market-efficiency literature. In the weak-form formulation, historical prices should not contain exploitable information about future returns once information processing, trading costs and risk compensation are taken into account (Fama, 1970). The strong version of this argument is deliberately austere: if return signs were persistently predictable from their own past, informed traders should arbitrage away that predictability. At the same time, the Grossman–Stiglitz critique implies that informational efficiency cannot be understood as a frictionless and perfectly absorbing state. Markets can be close to efficient in return direction while still displaying slow adjustment, heterogeneous information processing and persistent risk states (Grossman and Stiglitz, 1980). This distinction is central to the present paper. The empirical question is not whether all aspects of equity markets are unpredictable, but which aspect of the return process remembers the past.

Fractional integration provides the natural language for this question. In an ARFIMA-type process, the fractional parameter d governs the speed at which dependence decays. When $d = 0$, shocks have short memory; when $0 < d < 0.5$, the process is covariance-stationary but persistent; when d approaches or exceeds 0.5, persistence becomes so strong that the process behaves as a near-nonstationary risk state (Granger and Joyeux, 1980; Hosking, 1981; Baillie, 1996). For raw equity returns, positive long memory is difficult to reconcile with the weak-form EMH because it would imply persistent directional predictability. For return magnitudes, however, positive long memory is not an anomaly in the same sense. It indicates that volatility shocks, liquidity conditions and market uncertainty dissipate slowly even when the sign of the next return remains hard to forecast.

This separation between the mean and magnitude of returns is a robust fact in financial econometrics. The early evidence on speculative prices documented heavy tails and non-Gaussian return behaviour (Mandelbrot, 1963). Subsequent work showed that absolute and squared returns display much stronger dependence than returns themselves (Ding et al., 1993). Fractionally integrated volatility models formalised this insight by allowing volatility shocks to decay at a hyperbolic rather than exponential rate (Bollerslev and Mikkelsen, 1996). More recent evidence emphasises that persistence can vary with market states, data frequency and structural breaks, so a measured d should be interpreted as a risk-state parameter rather than as a purely mechanical statistic (Perron and Qu, 2010). In international markets, this is especially important because cross-market stress episodes often manifest first through volatility co-movement and tail dependence rather than through simple return autocorrelation (Longin and Solnik, 2001; Bartram et al., 2007).

The economic implication is direct. If raw returns are short-memory but absolute returns are long-memory, then the relevant inefficiency is not a simple trading rule based on yesterday's sign. It is the persistence of market risk, the slow normalisation of volatility after shocks and the possibility that institutions underestimate how long stress remains embedded in the market. For an international financial-markets journal, this distinction matters because it connects the EMH debate to risk management. A market can be informationally efficient in direction and still require models that remember volatility for many days or weeks. In that case, the object that financial institutions must forecast is not only $E(r_{t+1} \mid \mathcal{F}_t)$, but the persistence of $|r_t|$ and of the risk state implied by it.

H1 (directional efficiency with volatility memory). The S&P 500, DAX30 and large equity constituents display little or no fractional long memory in raw daily returns, while absolute returns display positive and economically meaningful fractional integration. Formally, $d(r_t) \approx 0$ and $d(|r_t|) > 0$.

H1 follows directly from the literature above. It does not treat volatility memory as a contradiction of weak-form efficiency. Instead, it predicts a dual structure: directional information is rapidly absorbed, while risk information persists.

2.2. Machine-learning forecasting, memory fidelity and institutional model risk

The second strand of literature concerns machine learning in empirical finance. Machine-learning models are increasingly used because they can approximate nonlinear relations, extract interactions from large predictor sets and improve out-of-sample forecasting in asset-pricing and volatility applications (Gu et al., 2020). Deep learning architectures extend this logic to sequential data. LSTM networks were designed to retain information through gated recurrent states (Hochreiter and Schmidhuber, 1997), attention mechanisms allocate weights to relevant parts of the input history (Vaswani et al., 2017), and hybrid ARFIMA–LSTM or attention-LSTM models explicitly combine time-series structure with nonlinear representation learning (Qiu et al., 2020; Bukhari et al., 2020; Lei et al., 2021). Recent work also shows that neural volatility models can resemble parsimonious fractional or rough-volatility mechanisms when trained on broad equity panels (Rosenbaum and Zhang, 2024).

The literature nevertheless leaves an important validation gap. Most ML-finance papers rank models by mean squared error, mean absolute error, directional accuracy, R^2, Sharpe ratios or portfolio returns. These criteria are informative but they do not answer whether the generated series has the correct dependence structure. A model may win a one-step forecasting contest by smoothing noise, but smoothing can attenuate low-frequency dependence and understate persistence. A recurrent model may be stable in recursive simulation, but its hidden state can mechanically propagate shocks for too long and overstate persistence. In both cases the forecast can look acceptable under standard metrics while being economically wrong for risk applications.

This is a model-risk problem rather than a purely statistical problem. Banks, asset managers, clearing houses and supervisors do not use volatility forecasts only to predict tomorrow's number. They use them to set margins, size positions, allocate capital, monitor stress and decide how quickly a volatility episode is expected to decay. If an ML model underestimates d, it implicitly tells the institution that the market will forget a shock too quickly. If it overestimates d, it makes stress appear excessively permanent and can trigger unnecessary de-risking. In both directions, the error is not captured by a one-day MSE. The relevant validation question is whether the model reproduces the market's memory of risk.

We therefore define *memory fidelity* as the ability of a model-generated or model-implied series to preserve the fractional integration parameter of the target process within a transparent tolerance. This criterion complements standard forecast accuracy. It is stricter than MSE because it evaluates the path property of forecasts, not only pointwise deviations. It is also more economically interpretable than a black-box accuracy ranking because it asks whether the model has learned the same risk horizon as the market. In this sense, fractional econometrics can serve as an external diagnostic for financial AI: the econometric benchmark tells us whether the machine has learned market memory or merely produced a numerically convenient forecast.

H2 (forecast accuracy is not memory fidelity). Conventional machine-learning accuracy metrics do not reliably identify models that preserve or estimate the fractional integration

parameter. Models ranked highly by MSE or directional accuracy differ from models ranked highly by d-preservation, and sequence architectures may either improve memory replication or impose excessive persistence.

H2 follows from the ML and model-risk literature. If financial AI is used for volatility and risk monitoring, the relevant validation layer must include the memory of the forecasted process, not only the average size or sign of forecast errors.

3. Data, Variable Construction and Empirical Strategy

3.1. Market sample, transformations and economic identification

The empirical design deliberately separates three objects that are often conflated in ML-finance applications: the market object, the risk object and the model-output object. The market object is the daily log return on an international equity benchmark or large constituent stock. The risk object is the absolute return, used as a transparent proxy for volatility and market stress. The model-output object is either a predicted path whose memory can be re-estimated or a direct scalar estimate of d. This separation is necessary because a model that predicts returns and a model that reproduces volatility memory are not solving the same economic problem.

The market sample covers daily adjusted closing prices from January 2010 to December 2025. The two benchmark markets are the S&P 500 and the DAX30. The S&P 500 represents the deepest U.S. equity benchmark and a global risk-pricing reference. The DAX30 represents a major euro-area equity benchmark exposed to a different monetary-policy environment, industrial structure and international trade cycle. The cross-section is extended with six large U.S. constituents: AAPL, MSFT, JPM, GS, XOM and CVX. These firms are not intended to form a representative equity universe. They are included as economically recognisable sectoral checks: technology, banking and energy stocks should all obey the same return-versus-risk-memory distinction if H1 is a general market property rather than an index-specific artefact.

Prices are converted into continuously compounded returns,

$$r_{i,t} = \ln(P_{i,t}) - \ln(P_{i,t-1}), \tag{1}$$

where $P_{i,t}$ is the adjusted closing price of asset i. The absolute-return series $|r_{i,t}|$ is then constructed from the same observations. This transformation is intentionally nonparametric. We do not estimate a GARCH model first because the purpose of the first stage is to measure memory in the observed return magnitude itself. Parametric volatility models enter the literature motivation, but the empirical benchmark must remain simple enough to be used as a validation diagnostic for arbitrary ML outputs.

3.2. Distributional properties and why they matter for ML validation

Table 2 reports the descriptive statistics. The return series have small daily means relative to volatility, negative skewness in most cases and excess kurtosis far above the Gaussian benchmark.

Table 1: Data architecture and empirical role of each variable block

Block	Series	Transformation	Empirical role
International benchmark indices	S&P 500, DAX30	Daily log returns and absolute log returns	Core test of directional efficiency versus volatility memory across U.S. and European markets
Large U.S. stocks	AAPL, MSFT, JPM, GS, XOM, CVX	Daily log returns and absolute log returns	Cross-sectional validation across technology, banking and energy exposures
Synthetic fractionally integrated series	Simulated ARFIMA-type processes	Target memory $d = 0.3$ for Type I; varying $d \in [-0.4, 1.0]$ for Type II	Controlled environment where the memory parameter is known or benchmarked
Model outputs	Predicted series and direct d estimates	FELW re-estimation of predicted paths; scalar prediction of d	Memory-fidelity validation of ML forecasts and ML estimators

The design separates the financial object of interest (returns versus absolute returns) from the machine-learning object of evaluation (point forecasts versus preserved memory).

This matters economically because the relevant market states are not average days. Equity risk is concentrated in clusters of large moves, and those clusters are precisely where volatility persistence affects margins, risk limits and de-risking decisions. It matters statistically because squared-error losses place high weight on large observations but do not reveal whether the model has learned their temporal persistence. The Student-t comparison in Figure 1 therefore supports the later synthetic design: heavy-tailed innovations are used so that the artificial series resemble the empirical environment more closely than Gaussian simulations.

Table 2: Descriptive statistics of daily log returns

Series	N	Mean (x100)	Std (x100)	Skewness	Kurtosis	Min (x100)	Max (x100)
S&P 500	3917.0000	0.0474	1.0983	-0.6254	13.7049	-12.7652	9.0895
DAX30	3995.0000	0.0332	1.2196	-0.4618	7.6967	-13.0549	10.4143
AAPL	3917.0000	0.0951	1.7748	-0.1392	6.2553	-13.7708	14.2617
MSFT	3917.0000	0.0762	1.6117	-0.1334	8.0951	-15.9453	13.2929
JPM	3917.0000	0.0679	1.7349	-0.1228	9.7884	-16.2106	16.5620
GS	3917.0000	0.0551	1.8026	-0.2688	8.5741	-13.6863	16.1951
XOM	3917.0000	0.0340	1.5680	-0.2186	7.2412	-13.0391	11.9442
CVX	3917.0000	0.0377	1.6886	-0.9465	26.5821	-25.0062	20.4904

Returns are continuously compounded. Mean, standard deviation, minimum and maximum are multiplied by 100. Kurtosis is reported in the empirical output convention used in the replication files.

3.3. FELW estimation, bandwidth design and interpretation of d

The baseline long-memory estimator is the feasible exact local Whittle estimator of Shimotsu and Phillips (2005); Shimotsu (2010). Let $I_x(\lambda_j)$ denote the periodogram of x_t at Fourier frequency $\lambda_j = 2\pi j/T$. The estimator uses the low-frequency part of the spectrum, where

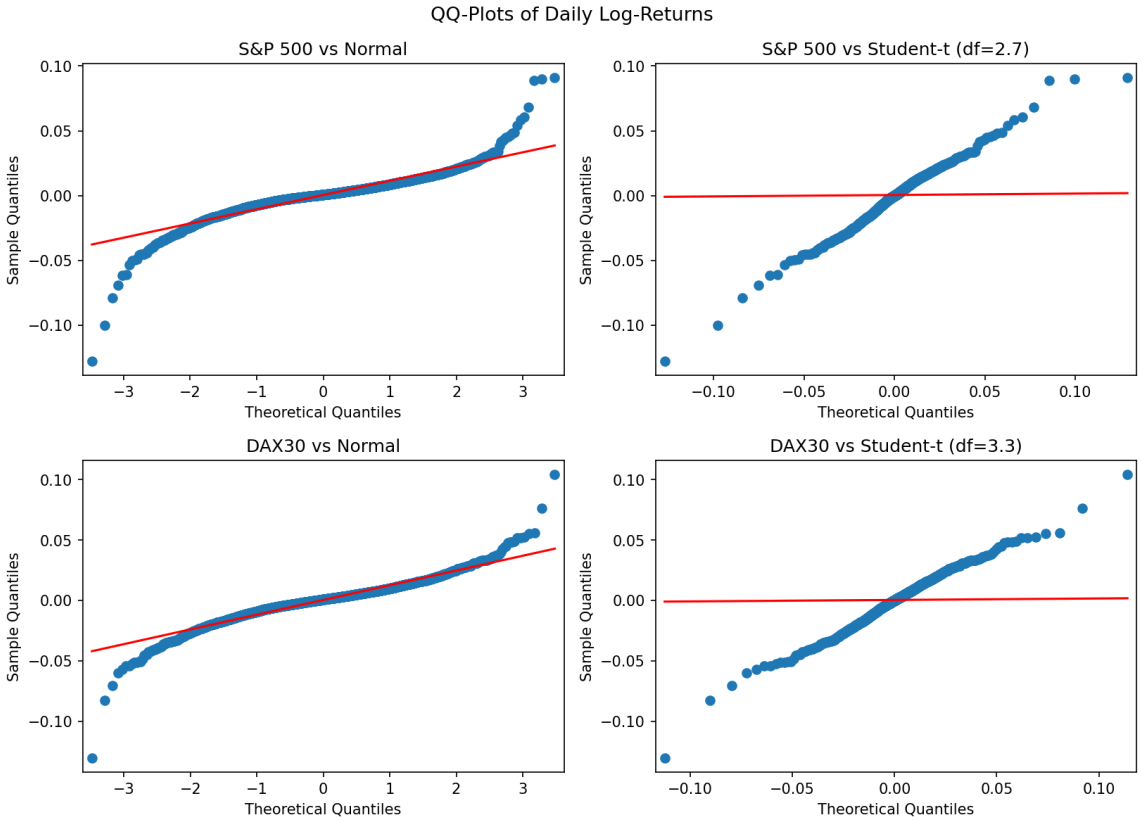

Figure 1: QQ-plots of daily log returns against Normal and Student-t benchmarks for the S&P 500 and DAX30. The Student-t benchmark captures the tail behaviour more closely than the Gaussian benchmark.

fractional integration is identified. In simplified form,

$$\widehat{d} = \arg\min_{d} \left[\log\left(\frac{1}{m} \sum_{j=1}^{m} \lambda_j^{2d} I_x(\lambda_j) \right) - \frac{2d}{m} \sum_{j=1}^{m} \log \lambda_j \right], \tag{2}$$

where $m = T^{\delta}$ is the bandwidth. The paper reports $\delta \in \{0.5, 0.6, 0.7\}$. A smaller bandwidth focuses more tightly on the lowest frequencies and is less sensitive to short-run dynamics; a larger bandwidth uses more ordinates and can reduce variance at the cost of greater contamination by short-memory components. Reporting all three values is therefore a robustness device rather than a mechanical sensitivity check. Economically, a stable positive estimate across bandwidths means that volatility memory is not confined to a narrow frequency band. A large disagreement across bandwidths would suggest that the result may be driven by short-run dynamics, jumps or breaks rather than by genuine low-frequency persistence.

Figure 2 adds a dynamic interpretation to the static FELW estimates. The relevant economic object is not only whether d is positive on average, but whether persistence rises precisely when institutions most need reliable risk horizons. The rolling estimates show that the memory of absolute returns is not constant: it moves with market conditions and becomes more pronounced around periods in which volatility shocks are likely to trigger portfolio rebalancing, margin pressure, deleveraging and changes in risk appetite. This pattern is important for JIFMIM-

Table 3: FELW estimates of d for returns and absolute returns, $\delta = 0.6$

Series	$d(r_t)$	$d(\lvert r_t \rvert)$
S&P 500	-0.1276	0.4463
DAX30	-0.1109	0.4118
AAPL	0.0594	0.3361
MSFT	-0.0366	0.3257
JPM	-0.0756	0.4527
GS	-0.0254	0.3774
XOM	-0.0288	0.4313
CVX	-0.1300	0.4098

The table uses $m = T^{0.6}$ as the central bandwidth. The empirical interpretation is based on the full bandwidth set $\delta \in \{0.5, 0.6, 0.7\}$.

Figure 2: Rolling-window FELW estimates for the S&P 500 and DAX30. The estimates show that volatility memory is time-varying and intensifies around major market disruptions.

type audiences because it links long memory to international market functioning rather than to a purely statistical anomaly. If persistence strengthens in stress states, then a one-day-ahead forecasting model that ignores memory can understate the duration of risk, even if it forecasts tomorrow's absolute return reasonably well. Figure 2 therefore motivates the ML validation strategy: the model is not asked only to forecast a value; it is asked to preserve the market's risk-memory horizon.

3.4. Machine-learning models and the two-stage validation protocol

The model set is designed to create an economically interpretable contrast between models that smooth, models that represent nonlinear features and models that carry state through time. It contains classical ML, feed-forward neural networks and sequential neural architectures. The classical block includes linear regression, Lasso, Ridge, regression trees, random forests, radial-basis-function support-vector regression and Gaussian-process regression. These models are useful benchmarks because they can fit local nonlinearities or regularised linear mappings, but they do not possess an explicit recurrent state. The neural block includes multilayer perceptrons, LSTM, bidirectional LSTM, GRU, standalone attention, Attention-LSTM, Embedding-Attention-LSTM and MLP-Embedding-Attention-LSTM variants.

The evaluation is deliberately split into two tests because the paper asks two different validation questions. The first is whether a forecasted path has the right memory. The second is whether a network can estimate memory directly. The **Type I test** evaluates memory preservation in forecasted paths. Synthetic fractionally integrated series are generated with target memory $d = 0.3$, and each model produces out-of-sample predictions. The predicted series is then re-estimated with FELW. A model passes when

$$\widehat{d}(\hat{y}_t) \in [0.2, 0.4]. \tag{3}$$

The same model is evaluated in open-loop and closed-loop mode. In open loop, true lagged observations are supplied as inputs. In closed loop, the model recursively uses its own forecasts. This distinction is essential: open-loop performance can exploit the memory already present in the realised input sequence, whereas closed-loop performance tests whether the model can sustain the correct memory structure endogenously.

The **Type II test** treats machine learning as a direct estimator of the memory parameter. The input is a series of length 512 and the output is a scalar estimate,

$$f_\theta : \{x_1, \ldots, x_{512}\} \mapsto \widehat{d}_\theta. \tag{4}$$

The benchmark is the FELW estimate. A prediction is counted as accurate when

$$|\widehat{d}_\theta - \widehat{d}_{\mathrm{FELW}}| < 0.1. \tag{5}$$

The training set contains 8,000 synthetic fractionally integrated series with d values drawn from $[-0.4, 1.0]$ and heavy-tailed innovations. The independent test set contains 512 series. This design evaluates whether ML can learn a mapping from the shape of a time series to its memory parameter, not merely whether it can predict the next observation.

3.5. Real-market transfer test: from controlled memory to equity-market risk

Synthetic experiments are necessary because the target memory parameter is controlled, but they are not sufficient for financial interpretation. A model can perform well on an artificial fractionally integrated process and still fail on real market data where volatility contains jumps, leverage effects, structural breaks and changing liquidity. The final layer therefore trains selected neural architectures on the first 80% of the S&P 500 absolute-return series and evaluates them on the remaining 20%. The true and predicted held-out absolute-return series are then re-estimated using FELW at all three bandwidths. This transfer test asks whether a model that behaves well in controlled experiments can reproduce the memory of an actual equity-market risk proxy.

4. Empirical Results

4.1. Directional efficiency but persistent risk: what the markets remember

The FELW estimates support H1 and give the paper its first economic message. Raw daily returns for the S&P 500, DAX30 and the six large U.S. stocks have d estimates close to zero or below zero. At the central bandwidth $\delta = 0.6$, the S&P 500 has $d(r_t) = -0.1276$ and the DAX30 has $d(r_t) = -0.1109$. This is not a trivial statistical detail. It means that the sample does not contain persistent directional memory of the type that would directly challenge weak-form efficiency through simple return-sign predictability.

The same markets look very different once the object is changed from direction to magnitude. At $\delta = 0.6$, $d(|r_t|)$ equals 0.4463 for the S&P 500 and 0.4118 for the DAX30. The large-stock estimates are also positive: 0.3361 for AAPL, 0.3257 for MSFT, 0.4527 for JPM, 0.3774 for GS, 0.4313 for XOM and 0.4098 for CVX. Across bandwidths, the S&P 500 estimate ranges from 0.305 to 0.498 and the DAX30 estimate ranges from 0.412 to 0.447. These values are economically large because they imply that the market's memory of risk decays slowly. A volatility shock is not absorbed in the way a short-memory ARMA disturbance would be absorbed; it remains visible in the dependence structure of return magnitudes.

The "so what" is therefore clear. The data do not say that an investor can forecast tomorrow's sign from yesterday's return. They say that risk managers, clearing houses and portfolio managers should not treat volatility shocks as short-lived disturbances. In both the U.S. and European benchmark markets, the market forgets return direction quickly but remembers risk. This is precisely the distinction that an ML model used for risk forecasting must preserve.

4.2. Accurate forecasts with the wrong risk horizon: memory preservation in predicted paths

The first part of H2 is evaluated through the Type I memory-preservation test. The target process has $d = 0.3$, and a forecasted path passes only if the re-estimated memory lies in $[0.2, 0.4]$. The design is deliberately simple because the economic implication is sharp: if the true risk process has moderate long memory and the predicted path has either short memory, anti-persistence or near-unit-root persistence, then the model has changed the horizon over which a volatility shock is expected to matter. That is not a cosmetic error. It changes portfolio turnover, volatility-targeting leverage, option hedging frequency, margin forecasts and stress-test persistence.

Table 4: Type I test: memory preservation of predicted series, target $d = 0.3$

Model	d_{OL}	OL pass	d_{CL}	CL pass	MSE OL	DA OL	DA CL
Linear Regression	0.4853	no	0.3837	yes	0.9002	73.0469	68.3594
Lasso	0.4930	no	0.6220	no	0.9030	72.2656	73.8281
Ridge	0.4864	no	0.3850	yes	0.8997	73.0469	68.3594
Regression Tree	0.3516	yes	-0.4985	no	0.9898	74.0234	73.8281
Random Forest	0.4102	no	0.3305	yes	0.9399	72.8516	73.8281
SVR (RBF)	0.6844	no	0.5812	no	0.9612	73.8281	72.6562
Gaussian Process (Matern)	1.1115	no	1.0993	no	1.1370	71.2891	73.8281
MLP (sklearn)	0.4476	no	0.3128	yes	1.0994	72.6562	56.4453
LSTM	0.9095	no	0.9304	no	0.9674	73.6328	73.8281
BiLSTM	0.8942	no	0.9972	no	0.9656	73.6328	73.8281
GRU	0.8601	no	0.8888	no	0.9672	72.6562	73.8281
Attention (Transformer)	0.3457	yes	0.9017	no	1.0490	73.2422	44.3359
Attention-LSTM	0.8268	no	0.8625	no	1.0057	72.4609	73.8281
Emb-Attention-LSTM	0.7826	no	0.8741	no	0.9973	71.4844	73.8281
MLP-Emb-Attention-LSTM	0.8213	no	0.8645	no	1.0217	71.6797	61.5234

OL = open loop; CL = closed loop. Pass means $\widehat{d} \in [0.2, 0.4]$. Directional accuracy is in percent.

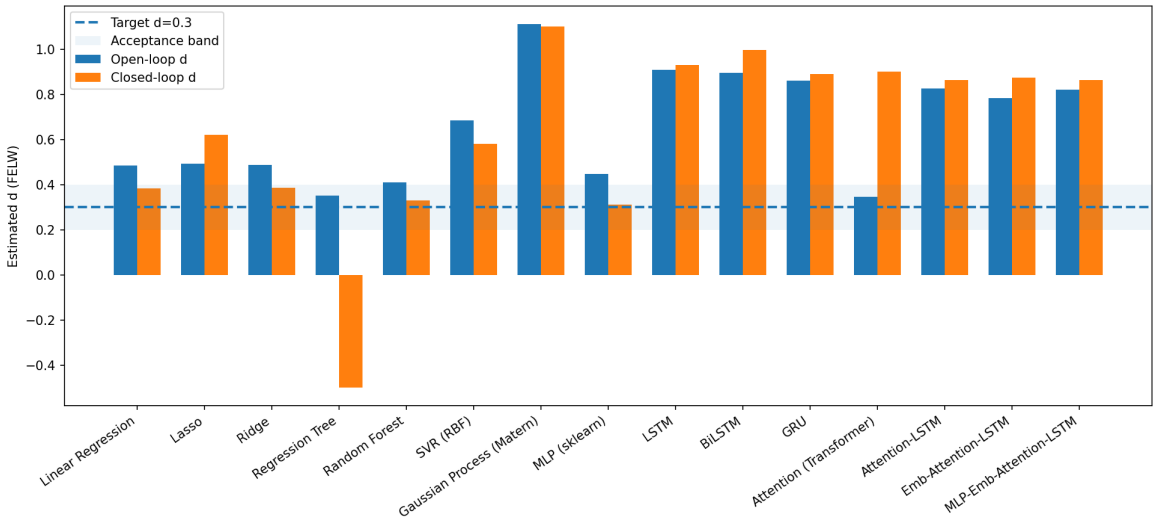

Figure 3: Memory preserved by forecast paths in the Type I experiment. The shaded band marks the economically admissible interval $[0.2, 0.4]$ around the target $d = 0.3$; values above the band imply excessive risk persistence, while values below the band imply that volatility shocks disappear too quickly or become anti-persistent.

Table 4 shows three economically distinct failure modes. First, the regularised linear models produce low one-step errors but do not reproduce the correct persistence in open loop. Linear

regression and Ridge have the best open-loop MSE values, but their open-loop d estimates are around 0.49, above the target band. In an institutional setting, such a model would not merely make a small statistical mistake; it would extend the perceived duration of a risk shock. A volatility-targeting strategy would deleverage for too long, while a stress-testing engine would treat a moderate-persistence shock as more persistent than it really is.

Second, models that pass in open loop can fail in recursive use. The regression tree and standalone attention model pass the open-loop memory criterion, but once they must feed on their own forecasts, their memory structure becomes economically implausible. The tree moves to strong anti-persistence in closed loop, while standalone attention jumps to excessive persistence. This distinction matters because many operational forecasting systems are recursive: a one-day model is repeatedly used to construct multi-day risk scenarios. Table 4 therefore shows why open-loop validation is too weak for risk applications. It tests interpolation around observed histories rather than the model's ability to generate a path with the right dependence structure.

Third, the recurrent and hybrid neural models reveal that architectural memory is not the same as market memory. LSTM, BiLSTM, GRU and attention-LSTM variants maintain high directional accuracy in closed loop, yet their estimated d values are generally between 0.78 and 1.00. Economically, they remember too much. This is the central model-risk result of Table 4: a neural architecture designed to carry state can convert moderate volatility persistence into near-permanent persistence unless the validation loss explicitly penalises memory distortion. The implication is that recurrent ML forecasts can be attractive for short-horizon accuracy but dangerous when they are used to infer the persistence of market stress.

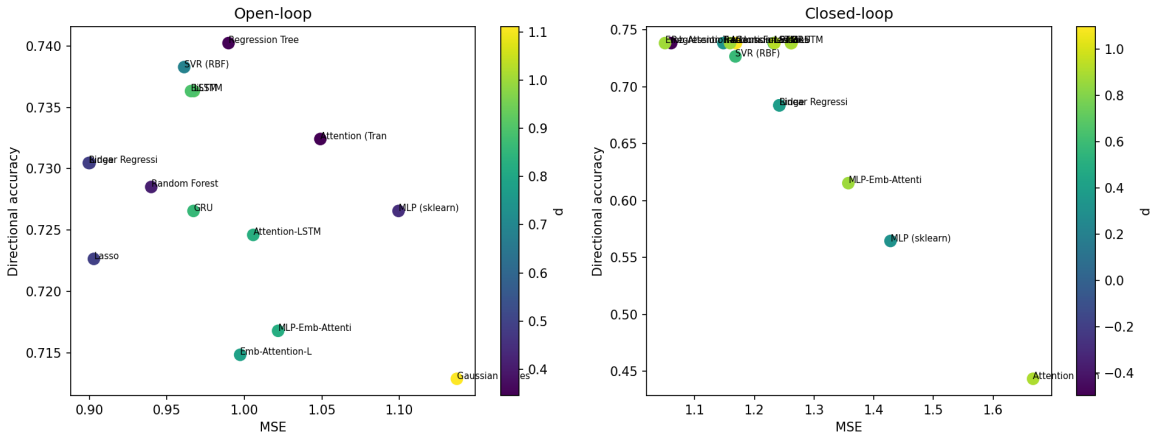

Figure 4: Point-forecast accuracy versus memory fidelity. The figure contrasts conventional validation criteria with the estimated fractional integration parameter of the predicted path.

Figure 4 gives the central empirical warning of the paper. The conventional accuracy frontier and the memory-fidelity frontier are not the same frontier. The lowest-MSE models are not the models that best preserve d, and high directional accuracy does not protect against an economically wrong dependence structure. This is why the result matters beyond a technical comparison of algorithms. In financial markets, point forecasts are often downstream inputs into risk allocation, VaR scaling, margin models and scenario design. When a model has a

good MSE but a biased d, the institution receives the right short-run signal with the wrong risk horizon. In calm markets this may be nearly invisible. In stressed markets it can become costly, because the institution either exits risk too late, re-enters too early or calibrates stress persistence to the architecture rather than to the market. Figure 4 therefore converts H2 into a governance claim: ML validation in finance should report memory fidelity whenever the target variable is a volatility or absolute-return proxy.

4.3. Learning the persistence parameter: direct estimation of market memory

The Type II experiment asks whether machine learning can estimate d directly. This is a harder task than ordinary forecasting because the model must infer a low-frequency property of the whole path from finite data. The economic motivation is different from Type I. Type I asks whether a forecast is safe to use as a generated risk path. Type II asks whether ML can become a diagnostic instrument for persistence itself: a fast, nonlinear estimator that screens many assets, regimes or rolling windows for long-memory risk.

Table 5: Type II test: machine learning as a direct estimator of the fractional integration parameter d

Model	Accept rate	MSE vs FELW
DNN depth=4	46.2891	0.0364
DNN depth=5	45.5078	0.0431
Attention-LSTM [T2]	43.5547	0.0340
MLP (1 layer)	37.8906	0.0556
MLP-Emb-Attention-LSTM [T2]	37.3047	0.0402
Random Forest	34.9609	0.0422
MLP depth=3	33.9844	0.0630
MLP depth=2	27.9297	0.1005
LSTM [T2]	27.7344	0.0684
Emb-Attention-LSTM [T2]	26.9531	0.0846

The table reports the ten best models. Acceptance rate is measured in percent. A prediction is accepted if $|\widehat{d}_\theta - \widehat{d}_{FELW}| < 0.1$.

Table 5 shows that the best direct estimator is a four-layer DNN, with an acceptance rate of 46.3% and MSE against FELW of 0.0364. A five-layer DNN and an Attention-LSTM follow closely, while linear and shallow models perform much worse in the full output files. The ranking is economically informative because it reverses a common intuition from financial sequence modelling. The best direct estimator of long memory is not necessarily the model with the most explicit recurrent structure. Instead, a sufficiently deep feed-forward network can learn a useful representation of the spectral and autocorrelation shape of the series. This supports a more precise interpretation of H2: memory can be learned by ML, but it should be learned as a target property, not assumed to emerge automatically from next-step prediction.

Figure 5 makes the same point visually. The leading models form a relatively tight group, but no architecture approaches a perfect estimator. This is expected because FELW itself is a semiparametric finite-sample benchmark rather than an observable truth. The economic lesson

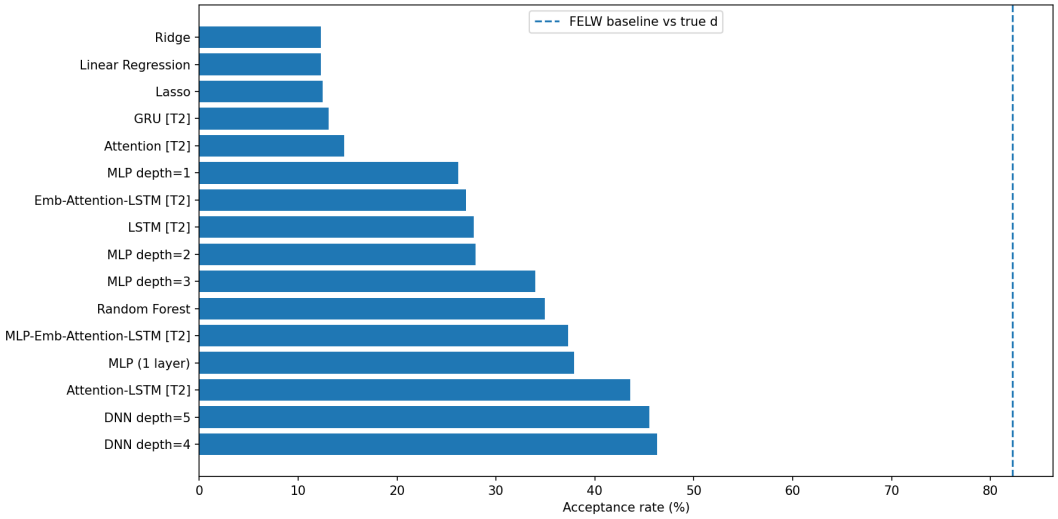

Figure 5: Acceptance rates in the Type II direct-estimation experiment. The figure ranks models by the share of test paths for which the ML estimate lies within 0.1 of the FELW benchmark.

is therefore not that ML should replace fractional econometrics. The lesson is that ML can be used as a complementary screening tool: it can flag which assets or windows display materially different persistence and then send those cases to more formal econometric validation. For asset managers or risk departments monitoring many markets, that role is practically meaningful.

The depth experiment in Figure 6 clarifies why the direct-estimation task is different from ordinary forecasting. Acceptance rises gradually from shallow to three-layer MLPs and then jumps at depth four, suggesting that the model needs several layers before it can combine local autocorrelation, tail behaviour and low-frequency spectral shape into an estimate of global power-law decay.

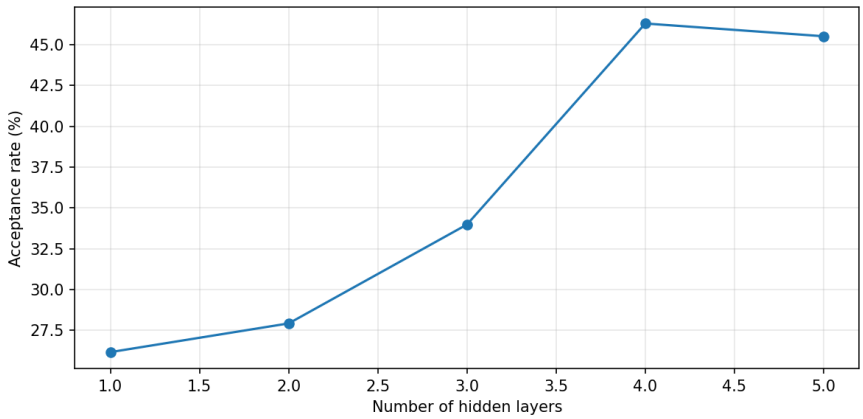

Figure 6: Depth and memory estimation in Type II. The strongest gain occurs between three and four hidden layers, indicating that estimating d requires nonlinear feature composition rather than simple smoothing.

Figure 6 is especially important for the "so what?" issue because it tells us what kind of complexity is useful. Additional complexity is valuable only when it helps the model form a representation of persistence. The improvement at four layers suggests that the network begins to combine short-run volatility clustering, distributional tail information and low-frequency

dependence in a way that resembles the information used by a semiparametric memory estimator. This supports H2 from a second angle. ML can learn memory-related features, but the model that estimates d best is not automatically the model that forecasts point values best. Memory should therefore be treated as a separate supervised target or validation constraint.

5. From Synthetic Memory to Market Risk: The S&P 500 Transfer Test

The final test asks whether synthetic memory performance transfers to an actual market-risk series. The held-out target is the S&P 500 absolute-return series. This is the economically most relevant test because $|r_t|$ is the object that carries volatility memory in H1 and the object that institutional risk systems would actually monitor.

Table 6: Real-market application: S&P 500 held-out absolute returns

Market	Model	d true .5	d pred .5	diff .5	d true .6	d pred .6	diff .6	d true .7	d pred .7	diff .7	Avg |diff|
S&P 500	Emb-Attention-LSTM	0.3051	0.1890	0.1161	0.4463	0.4332	0.0132	0.4977	0.5137	0.0160	0.0484
S&P 500	Attention-LSTM	0.3051	0.5863	0.2812	0.4463	0.6557	0.2093	0.4977	0.4661	0.0316	0.1741
S&P 500	MLP-Emb-Attention-LSTM	0.3051	0.6061	0.3010	0.4463	0.8568	0.4105	0.4977	0.7227	0.2250	0.3122
S&P 500	LSTM	0.3051	0.5227	0.2176	0.4463	0.8555	0.4092	0.4977	0.9678	0.4701	0.3656

The table compares FELW estimates of the true held-out absolute-return series and predicted absolute-return series.

Table 6 is the economically decisive transfer test. The embedding-attention LSTM is the only architecture that reproduces the held-out S&P 500 memory with a small error across bandwidths. Its average absolute d error is 0.048, compared with 0.174 for Attention-LSTM, 0.312 for MLP-Embedding-Attention-LSTM and 0.366 for the standard LSTM. The result is especially strong at the intermediate and high bandwidths: at $\delta = 0.6$, the true d is 0.4463 and the predicted d is 0.4332; at $\delta = 0.7$, the true d is 0.4977 and the predicted d is 0.5137.

The comparison is not only a ranking of neural networks. It shows that the same asset, the same training window and the same target variable can imply very different persistence estimates depending on architecture. A standard LSTM substantially overstates memory at $\delta = 0.6$ and $\delta = 0.7$, which would translate into an exaggerated duration of market stress. The MLP-Embedding-Attention-LSTM also overstates persistence, suggesting that adding components does not automatically improve financial validity. The embedding-attention LSTM is economically preferable because it approximates the risk horizon of the actual S&P 500 absolute-return series rather than merely generating smooth predictions.

Figure 7 explains why the table matters. A useful risk model must reproduce not only the scale of absolute returns but also the autocorrelation pattern that determines how long a shock remains relevant. The embedding-attention LSTM is not simply smoother than the target; it preserves enough of the dependence structure to keep the FELW estimate close to the observed one. This result supports H2 with an important qualification. The best architecture is not the model with the strongest recurrent memory. It is the architecture that balances embedding, attention and recurrent dynamics.

The economic content of this result is therefore not that complexity wins mechanically. The standard LSTM is worse than the embedding-attention LSTM because it over-imposes

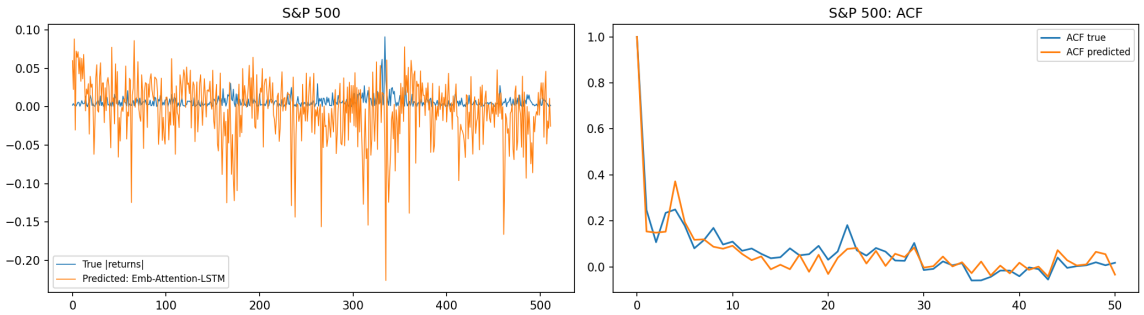

Figure 7: Real-market transfer to S&P 500 absolute returns. The figure compares the held-out absolute-return path and autocorrelation structure with the embedding-attention LSTM prediction, showing whether the model preserves the empirical risk-memory pattern rather than only the level of volatility.

persistence. The hybrid model works because it appears to balance three tasks: embedding stabilises local return-magnitude representation, attention selects relevant historical states, and recurrence carries volatility-regime information. In financial terms, the model behaves less like a mechanical smoother and more like a selective risk-memory filter. That is the practical answer to the "so what?" question: the architecture matters because different architectures imply different risk horizons.

6. Discussion: Why the Results Matter for Markets and Institutions

The paper's main implication is that market efficiency and risk persistence should be discussed together rather than separately. The evidence supports a market in which directional return predictability is weak, but volatility memory is strong. This is not a contradiction; it is a realistic description of liquid equity markets. Prices may incorporate directional information quickly, while volatility, liquidity conditions and uncertainty regimes decay slowly. For international investors, this means that diversification and risk budgeting depend less on forecasting the sign of tomorrow's return and more on understanding how long risk states persist across markets.

The second implication is that long memory is an economic risk-horizon parameter. When $d(|r_t|)$ is around 0.4, as in the S&P 500 and DAX30 results, volatility shocks do not behave like short-lived noise. They create a persistent environment in which risk limits, margins, leverage constraints and portfolio overlays remain affected after the initial shock. A model that underestimates this persistence will release risk capital too quickly; a model that overestimates it will keep institutions in a defensive posture for too long. Both errors are costly, but they are invisible if validation stops at one-day MSE.

The third implication concerns financial AI governance. In many institutional workflows, ML forecasts are not consumed as isolated point predictions. They feed into VaR engines, volatility targeting, stress dashboards, collateral models and portfolio construction. Those applications require path realism. A predicted volatility path must have a plausible memory structure because the persistence of the path determines the persistence of constraints. The

paper therefore proposes memory fidelity as a model-risk diagnostic: every ML model used for risk forecasting should be asked not only "how large is the forecast error?" but also "does the generated path have the correct d?"

The fourth implication concerns architecture selection. Linear and regularised models can look good under one-step error metrics but fail to represent low-frequency dependence. Pure recurrent models can solve the stability problem and create a new one: excessive persistence. Standalone attention can select useful history in open loop but may be fragile in recursive use. The best real-market result comes from an embedding-attention LSTM because it combines selective history with recurrent state. This suggests that financial ML architecture should be evaluated by economic behaviour, not only by statistical ranking.

The fifth implication is methodological. The paper shows how semiparametric econometrics can discipline black-box financial AI. FELW is not a competitor to ML in the usual forecasting horse race. It is an external measuring instrument. It tells us whether a forecast path preserves the market-memory property that the econometric literature has shown to be economically meaningful. This combination is useful for journals such as JIFMIM because it links modern prediction tools with interpretable financial mechanisms.

The limitations also matter. The real-market transfer test is deepest for the S&P 500, while a full international extension should include a broader panel of developed and emerging markets, exchange rates, bond volatility and crisis subsamples. The synthetic design isolates fractional memory, whereas real markets combine memory with jumps, breaks, leverage effects and market-microstructure frictions. The paper also evaluates memory fidelity rather than trading profitability. These are not weaknesses of the contribution but boundaries of the claim: the paper does not say that memory fidelity alone generates abnormal returns. It says that without memory fidelity, ML-based risk forecasts can be economically misleading.

Overall, the answer to "so what?" is that the paper identifies a missing validation layer for financial AI. Equity markets may be efficient in direction but persistent in risk. Models used by institutions must therefore remember the right object, by the right amount, for the right economic purpose.

7. Conclusion

This paper asks whether machine-learning models can remember financial markets in the way that matters for market efficiency and institutional risk control. The answer is conditional. Machine-learning models can reproduce parts of the memory structure of equity-market risk, and the embedding-attention LSTM performs strongly in the held-out S&P 500 absolute-return application. But conventional accuracy metrics are not sufficient, and many models either fail to preserve the fractional integration parameter or impose too much persistence.

The evidence supports two conclusions. First, major equity markets in the sample exhibit directional efficiency with volatility memory: raw daily returns have d estimates close to zero or negative, while absolute returns display positive long memory. Second, forecast accuracy is not

memory fidelity. MSE and directional accuracy do not reliably select the models that preserve or estimate d. This is the main methodological contribution of the article.

The broader recommendation is straightforward. Empirical ML-finance papers should report not only how accurately a model predicts, but also whether the predicted process reproduces the relevant dependence structure. For financial institutions, this is a model-risk issue. A model that forecasts well but remembers incorrectly can distort risk horizons, stress persistence and the speed at which volatility shocks are assumed to decay.

Future research should extend the framework to a larger panel of international equity indices, emerging markets, exchange rates, bond-market volatility and cross-asset risk measures. A second extension should combine memory-fidelity diagnostics with economic loss functions, including VaR exceedances, margin backtesting and volatility-targeting performance. The central message, however, is already visible in the present evidence: financial AI needs econometric memory checks.

Data and code availability

The empirical tables and figures are generated from the authors' replication files. Raw market prices were transformed into daily log returns and absolute log returns before estimation. The article draft includes the compiled tables and figures needed to reproduce the reported conclusions.

Declaration of generative AI and AI-assisted technologies

During manuscript preparation, AI-assisted tools were used for language editing, restructuring and LaTeX drafting. The authors remain responsible for the content, empirical results and interpretation.

References

Bartram, S.M., Brown, G.W., Hund, J.E., 2007. Estimating systemic risk in the international financial system. Journal of Financial Economics 86, 835–869.

Ding, Z., Granger, C.W.J., Engle, R.F., 1993. A long memory property of stock market returns and a new model. Journal of Empirical Finance 1, 83–106.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica 50, 987–1007.

Fama, E.F., 1970. Efficient capital markets: A review of theory and empirical work. Journal of Finance 25, 383–417.

Granger, C.W.J., Joyeux, R., 1980. An introduction to long-memory time series models and fractional differencing. Journal of Time Series Analysis 1, 15–29.

Grossman, S.J., Stiglitz, J.E., 1980. On the impossibility of informationally efficient markets. American Economic Review 70, 393–408.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Review of Financial Studies 33, 2223–2273.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Computation 9, 1735–1780.

Hosking, J.R.M., 1981. Fractional differencing. Biometrika 68, 165–176.

Lo, A.W., 1991. Long-term memory in stock market prices. Econometrica 59, 1279–1313.

Longin, F., Solnik, B., 2001. Extreme correlation of international equity markets. Journal of Finance 56, 649–676.

Mandelbrot, B., 1963. The variation of certain speculative prices. Journal of Business 36, 394–419.

Qiu, J., Wang, B., Zhou, C., 2020. Forecasting stock prices with long-short term memory neural network based on attention mechanism. PLOS ONE 15, e0227222.

Shimotsu, K., 2010. Exact local whittle estimation of fractional integration with unknown mean and time trend. Econometric Theory 26, 501–540.

Shimotsu, K., Phillips, P.C.B., 2005. Exact local whittle estimation of fractional integration. Annals of Statistics 33, 1890–1933.

Ubal, C., Di-Giorgi, G., Contreras-Reyes, J.E., Salas, R., 2023. Predicting the long-term dependencies in time series using recurrent artificial neural networks. Machine Learning and Knowledge Extraction 5, 1340–1358.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need, in: Advances in Neural Information Processing Systems.

Baillie, R.T., 1996. Long memory processes and fractional integration in econometrics. Journal of Econometrics 73, 5–59.

Bollerslev, T., Mikkelsen, H.O., 1996. Modeling and pricing long memory in stock market volatility. Journal of Econometrics 73, 151–184.

Perron, P., Qu, Z., 2010. Long-memory and level shifts in the volatility of stock market return indices. Journal of Business & Economic Statistics 28, 275–290.

Papailias, F., Dias, G.F., 2015. Forecasting long memory series subject to structural change: A two-stage approach. International Journal of Forecasting 31, 1056–1066.

Bukhari, A.H., Raja, M.A.Z., Sulaiman, M., Islam, S., Shoaib, M., Kumam, P., 2020. Fractional neuro-sequential ARFIMA-LSTM for financial market forecasting. IEEE Access 8, 71326–71338.

Lei, B., Zhang, B., Song, Y., 2021. Volatility forecasting for high-frequency financial data based on web search index and deep learning model. Mathematics 9, 320.

Rosenbaum, M., Zhang, J., 2024. On the universality of the volatility formation process: When machine learning and rough volatility agree. Frontiers of Mathematical Finance 3, 1–24.