

# The Accuracy–Diversity Frontier in Adaptive Volatility Forecasting

---

Fuli Yang

GARCH Institute · Cambridge, Massachusetts, USA

Manuscript v1.3 · September 2026

## Abstract

Adaptive forecasting systems are usually evaluated by predictive loss. This paper asks a different question: what happens to collective model diversity when forecasting systems retain different amounts of historical information? Using four equity markets from 2000–2026, a common out-of-sample period beginning in 2015, four forecast horizons, and heterogeneous forecasting technologies, I construct frozen, finite-memory, and expanding-memory forecasting cohorts that differ only in their updating rule. Full-history expanding adaptation produces lower QLIKE loss and lower effective model diversity than every finite-memory specification throughout the 16-cell market–horizon robustness grid. The cells are dependent and are not interpreted as independent replications; when horizons are collapsed within market, all four market blocks retain the same sign, but four blocks provide limited standalone statistical power. An equal-sample-size exclusion test shows that the contrast survives when broad-history and matched 3-, 5-, and 10-year learners use the same number of training observations, indicating that sample size alone does not account for the frontier. Around 22 rule-based market-specific stress onsets, expanding adaptation attenuates the post-stress loss of forecast diversity relative to a frozen benchmark; the result also survives re-clustering temporally overlapping cross-market events into global episodes. A volatility-targeting mapping shows that forecast architecture can propagate to the correlation of position adjustments, while the paper deliberately stops short of claiming a causal effect on observed market volatility. The evidence implies that predictive accuracy and collective diversity are distinct design objectives for adaptive financial forecasting systems.

**Keywords:** volatility forecasting; adaptive learning; model diversity; QLIKE; market stress; continual learning

**JEL:** C22; C53; C58; G17

## 1. Introduction

Financial forecasting systems are increasingly adaptive. Models are retrained, recent observations are reweighted, and large-scale machine-learning pipelines can update more frequently than traditional econometric specifications. Evaluation, however, remains dominated by individual predictive loss. That criterion is natural for a forecaster, but it is incomplete for a system in which many models respond to the same data. A collection of individually accurate models may provide less diversification than its model count suggests if their outputs collapse onto a small number of common directions.

A related forecasting literature has long emphasized that the value of maintaining multiple forecasts depends not only on their individual accuracy but also on their dependence structure. Forecast combination can improve accuracy when constituent forecasts contain distinct information, while highly correlated forecast errors limit diversification gains (Clemen, 1989; Timmermann, 2006). The present paper shifts that idea from combination weights to system design: cross-model forecast dependence and effective diversity are treated as outcomes that can themselves change with the learning architecture.

Recent financial-AI evidence makes this distinction increasingly consequential. Lim and Kim (2026) show that LLM-based volatility estimates can generate covariance structures with a higher largest-eigenvalue share and lower effective rank than historical estimates, even while producing favorable portfolio outcomes. Saxena et al.

(2026) show in simulated asset markets that sophisticated LLM agents adapt to other agents and can raise profits while contributing to greater market volatility. These findings motivate a narrower empirical question that does not require treating any one technology as representative of AI: does the memory architecture of an adaptive forecasting system change the trade-off between predictive accuracy and cross-model diversity?

I answer this question with a controlled historical replay. Four equity markets—the S&P 500, FTSE 100, Nikkei 225, and Hang Seng—are observed from 2000 through June 2026. Forecasts are evaluated out of sample from January 2015. For each market and forecast horizon, the same feature set and model family are subjected to five updating architectures: a frozen pre-2015 model, rolling 3-, 5-, and 10-year memories, and an expanding-history learner. Because the information set and forecasting technologies are held fixed, the comparison isolates the consequences of retaining or discarding historical observations rather than comparing unrelated model classes.

The central result is simple but not obvious. Expanding-history adaptation has lower average QLIKE than every finite-memory alternative in all 16 market–horizon cells. At the same time, expanding-history forecasts have lower effective diversity than every finite-memory alternative in the same 16 cells. The relationship need not be monotone among 3-, 5-, and 10-year windows; rather, full-history learning occupies a distinct low-loss, low-diversity corner of the feasible set. I therefore describe the object as an accuracy–diversity frontier rather than a mechanical monotonic-memory law.

The paper contributes to three literatures. First, it adds a system-level criterion to volatility forecast evaluation, complementing broad model-comparison evidence and work on forecast loss and task-specific volatility measurement (Hansen and Lunde, 2005; Corsi, 2009; Patton, 2011; Patton and Zhang, 2026). Second, it connects forecast-combination ideas about dependence and diversification to adaptive model ensembles. Third, it contributes to financial-AI evaluation by distinguishing the predictive performance of individual learners from the diversity of the forecasting system that deploys them.

## 2. Relation to Recent Work

The paper is closest to three emerging strands. Regime-aware LLM forecasting shows that conditioning on state information can materially improve high-volatility performance, implying that adaptation design—not simply model class—matters under nonstationarity (Asaad, Hamidi, and Bereyhi, 2026). LLM-based volatility estimation has also been associated with more concentrated covariance geometry (Lim and Kim, 2026). At the market-interaction level, Machine Spirits documents adaptation across heterogeneous LLM agents and the possibility that individual strategic success coexists with greater aggregate instability (Saxena et al., 2026). The present paper differs by using real historical market paths, holding the forecasting technologies fixed, and making retained memory itself the treatment of interest.

The memory treatment also has a classical forecasting interpretation. Estimation-window choice under structural breaks trades bias from stale observations against variance from discarding usable history (Pesaran and Timmermann, 2007), while the concept-drift literature studies how online learners should adapt when predictive relationships change over time (Gama et al., 2014). Those literatures focus primarily on forecast performance. Here the same memory choice is evaluated jointly against an ensemble-level diversity criterion.

The design also responds to concerns about reproducibility in agentic trading research. Xia et al. (2026) find substantial protocol incomparability across the existing literature. Here all forecasting dates, data splits, refit frequencies, model hyperparameters, horizons, and stress definitions are fixed in code and can be replayed deterministically from the same inputs.

### 3. Data and Experimental Design

#### 3.1 Markets and targets

The dataset contains daily observations for the S&P 500, FTSE 100, Nikkei 225, and Hang Seng from 2000-02-02 through 2026-06-30. The base volatility state is a 22-trading-day realized-volatility measure. The rich feature panel additionally contains precomputed location, scale, distributional-shape, lower- and upper-tail shape, and speed coordinates from the author's broader distributional-volatility data construction, together with lagged and rolling transformations of realized volatility. Because reliance on precomputed coordinates would impede stand-alone replication, Section 4.5 repeats the full memory grid with a transparent RV-only feature set defined entirely in this paper. Forecast horizons are 5, 10, 22, and 44 trading days.

The common calibration sample ends on 2014-12-31. All observations after that date form the out-of-sample evaluation period. Forecast targets are estimated in logs and transformed back to levels. The primary loss measure is QLIKE, which is well suited to volatility forecast evaluation in the presence of noisy volatility proxies; MSE is retained as a secondary metric.

The transparent replication feature set is  $\{RV_t, RV_{\{t-1\}}, RV_{\{t-5\}}, RV_{\{t-22\}}, RV_{\{t-66\}}, \text{mean}_5(RV), \text{mean}_{22}(RV), \text{sd}_{22}(RV)\}$ . All rolling moments use trailing observations only. This ablation removes every distributional coordinate and therefore makes the central frontier reproducible from the realized-volatility series alone. The richer coordinates are retained as the baseline because they were available at every historical forecast date, but they are not necessary for the principal result.

#### 3.2 Forecasting cohort and memory treatment

The baseline cohort consists of Ridge, Elastic Net, k-nearest-neighbor, and bounded-depth regression-tree forecasts. A 22-day robustness cohort adds histogram gradient boosting. The central treatment is not model identity but memory architecture. Frozen models are fit only to the pre-2015 sample. Rolling learners retain the most recent 3, 5, or 10 years at each refit. Expanding learners retain the entire available history. Adaptive models are refit every 252 trading days. All architectures use the same contemporaneously available feature space.

Within any architecture and refit date, all forecasting technologies are trained on exactly the same rows. Pairwise training-set overlap across models is therefore 100% for rolling and expanding cohorts alike. The memory treatment changes the number and historical span of retained observations, not the fraction of observations shared across model classes. Section 4.5 separately equalizes training-set size to distinguish a pure sample-size explanation from a history-span/composition effect.

#### 3.3 Effective model diversity

Let  $R_t$  denote the correlation matrix of model forecasts evaluated over a common window, with eigenvalues  $\lambda_1, \dots, \lambda_N$ . Effective model diversity is defined as

$$D_{\text{eff}} = (\sum \lambda_i)^2 / \sum \lambda_i^2.$$

The measure equals  $N$  when eigenvalues are evenly dispersed and approaches one when forecasts become effectively one-dimensional. I also report average pairwise forecast correlation, the leading-eigenvalue share, and analogous measures for forecast errors. These objects are deliberately distinguished: forecasts can become similar without necessarily making the same errors.

#### 3.4 Stress onsets and position mapping

Stress is defined from the pre-2015 distribution of realized volatility. A stress onset is a transition into the upper-tail volatility state after a clean pre-event interval, with overlapping episodes removed mechanically within each market. This yields 22 market-specific onsets across the four markets. Because global stress can

trigger several markets within days of one another, these 22 onsets are not treated as 22 economically independent shocks. Event statistics compare pre-event and post-event diversity; inference is reported both at the market-event level and after temporally adjacent cross-market onsets are collapsed into global episode clusters. To assess a plausible transmission mechanism, forecasts are also mapped into a common volatility-targeting rule. The resulting correlation of position changes is interpreted as an implied synchronization channel, not as an estimate of historical market impact.

## 4. Results

### 4.1 The accuracy-diversity frontier

Across the 16 market-horizon cells, expanding-history learning has lower QLIKE than every finite-memory specification in 16/16 cells. The same expanding specification has lower effective model diversity than every finite-memory specification in 16/16 cells. These cells are strongly dependent because four horizons share each market and the markets co-move. I therefore use 16/16 as a direction-of-effect summary, not as a claim of 16 independent validations. Collapsing the four horizons to a single average within each market leaves the same sign in all four markets for both outcomes; with only four market blocks, an exact one-sided sign test is  $p=0.0625$ . The evidential weight therefore comes from the uniform direction, effect sizes, exclusion tests, and architecture robustness rather than from a small cell-level p-value.

Cell-level signed-rank calculations are retained only as descriptive sensitivity checks. They are not used to infer an effective sample size of 16. The 3-, 5-, and 10-year windows also do not form a monotone diversity sequence in most cells, which is inconsistent with a simple law that more observations must mechanically produce lower diversity. The more specific empirical regularity is that full-history adaptation occupies a distinct low-loss, low-diversity corner relative to every finite-memory rule tested.

| Memory      | Mean QLIKE | 95% market-block CI | Mean D_eff | 95% market-block CI |
|-------------|------------|---------------------|------------|---------------------|
| Frozen      | 0.0582     | [0.0462, 0.0729]    | 1.234      | [1.163, 1.317]      |
| Rolling 3Y  | 0.1496     | [0.0714, 0.2515]    | 1.641      | [1.453, 1.817]      |
| Rolling 5Y  | 0.0861     | [0.0547, 0.1296]    | 1.474      | [1.400, 1.549]      |
| Rolling 10Y | 0.0654     | [0.0496, 0.0843]    | 1.505      | [1.422, 1.600]      |
| Expanding   | 0.0580     | [0.0456, 0.0724]    | 1.229      | [1.159, 1.310]      |

The expanding-versus-10-year QLIKE gap is numerically modest: the mean difference is -0.0074, equivalent to about 11.3% of the rolling-10-year mean QLIKE. The diversity difference is larger in scale (mean D\_eff 1.229 versus 1.505). I therefore distinguish statistical directional consistency from economic magnitude rather than treating every contrast as equally consequential.

The confidence intervals resample the four markets as blocks, keeping all horizons from a sampled market together. With only four market blocks, they are intentionally conservative and should be read as uncertainty summaries rather than high-powered cross-sectional inference.

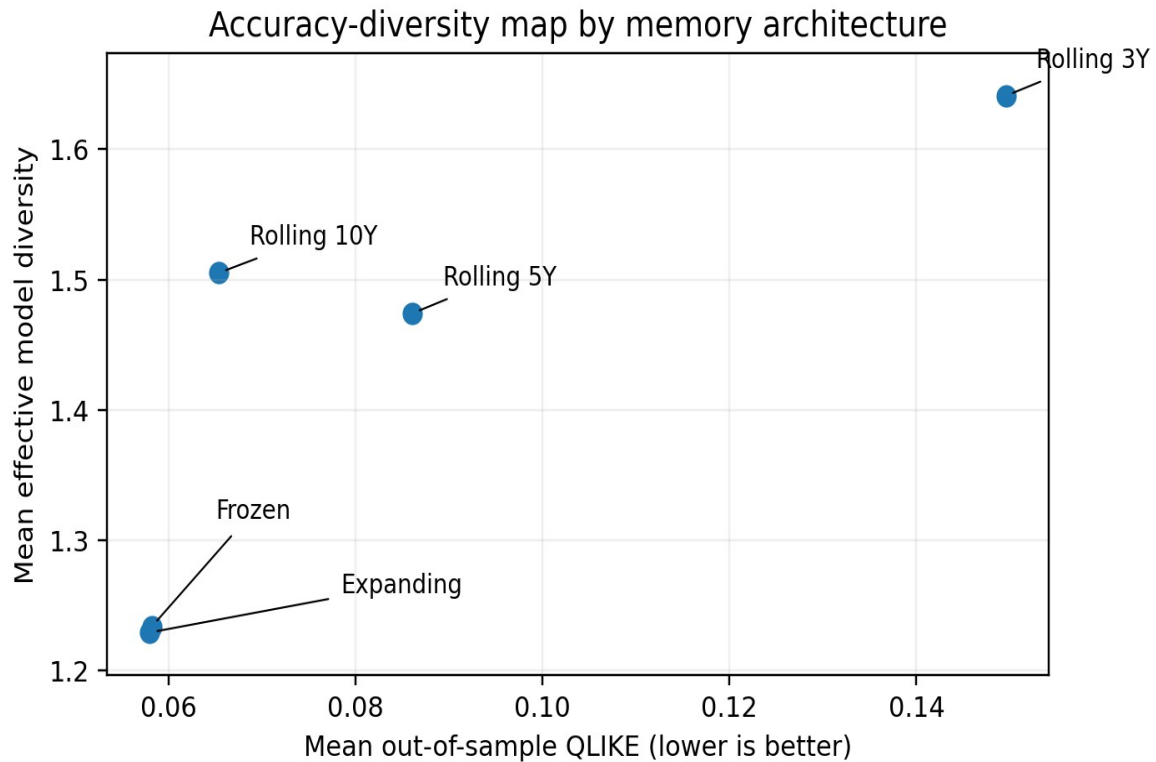


Figure 1. Mean out-of-sample QLIKE and effective model diversity by memory architecture. Each point averages the 16 market-horizon cells.

#### 4.2 Stress transitions

Stress itself often compresses forecast diversity, but the architecture determines the magnitude of that compression. Define the event change as post-event  $D_{\text{eff}}$  minus pre-event  $D_{\text{eff}}$ . Relative to frozen models, expanding-history adaptation makes this change 0.061 less negative on average. In other words, expanding learners lose less diversity around stress onsets; the estimate does not mean that diversity necessarily rises after stress begins. A market-event cluster bootstrap gives a 95% interval of [0.009, 0.115].

Cross-market event dependence does not overturn this result. I repeat global-episode clustering over linkage widths from 5 to 30 calendar days. The 22 market-specific onsets collapse to between 16 and 12 global episodes. Across all six linkage widths, the expanding-versus-frozen event-change estimate remains 0.061 and the global-cluster bootstrap interval remains above zero; 95% lower bounds range from 0.0068 to 0.0109 and upper bounds from 0.118 to 0.128. The effective number of stress shocks is therefore clearly below 22, but the conclusion is not tied to a single arbitrary linkage width. The linkage is transitive rather than capped by total span; Appendix A reports the realized maximum episode span, which never exceeds 18 calendar days over the 5–30-day sensitivity grid.

The position mapping yields a second, narrower result. Five-year rolling adaptation reduces the post-stress change in the correlation of position adjustments relative to frozen forecasts by 0.097 (95% CI [-0.177, -0.016],  $p=0.019$ ). The sign is consistent with finite-memory heterogeneity dampening synchronized risk adjustment. Other position-mapping coefficients are less precisely estimated, so this channel is treated as supporting rather than primary evidence.

#### 4.3 Robustness to a more modern nonlinear cohort

Adding histogram gradient boosting at the 22-day horizon does not eliminate the central contrast. In all four markets, the 5-year rolling cohort has substantially greater effective diversity than the frozen or expanding

cohort, while QLIKE is generally worse. The robustness exercise therefore supports the interpretation that the frontier is not an artifact of a single linear model family.

#### 4.4 Alternative diversity definitions and error dependence

The frontier is not specific to the participation-ratio definition of effective diversity. Relative to each 3-, 5-, and 10-year rolling cohort, expanding-history forecasts have higher average pairwise forecast correlation in 16/16 cells and a higher leading-eigenvalue share in 16/16 cells. These six output-concentration checks, together with three error-diversity comparisons, form a separate nine-test robustness family. Market-block bootstrap intervals for all six output-concentration effect sizes exclude zero in the concentration direction, and Holm adjustment across the nine-test family leaves all six adjusted p-values below 0.001. The effect-size intervals are reported in Table 2 so that economic magnitude, not only statistical direction, is visible.

Error diversity is deliberately separated from output diversity. Expanding history lowers effective diversity of forecast errors relative to the 3-year and 10-year cohorts, with market-block intervals that exclude zero, but the 5-year comparison is unresolved and its interval crosses zero. Holm-adjusted p-values are approximately 0.005, 0.353, and 0.029 against 3Y, 5Y, and 10Y, respectively. This mixed result narrows the interpretation: the paper identifies concentration in forecast outputs, not a universal tendency for adaptive models to make the same mistakes.

| Metric         | Comparison      | Mean effect | 95% market-block CI | Holm p   |
|----------------|-----------------|-------------|---------------------|----------|
| Forecast corr. | Expanding - 3Y  | +0.195      | [+0.100, +0.311]    | 0.000137 |
| Forecast corr. | Expanding - 5Y  | +0.118      | [+0.079, +0.158]    | 0.000137 |
| Forecast corr. | Expanding - 10Y | +0.130      | [+0.068, +0.193]    | 0.000137 |
| PC1 share      | Expanding - 3Y  | +0.141      | [+0.073, +0.222]    | 0.000137 |
| PC1 share      | Expanding - 5Y  | +0.086      | [+0.057, +0.114]    | 0.000137 |
| PC1 share      | Expanding - 10Y | +0.095      | [+0.048, +0.142]    | 0.000137 |
| Error D_eff    | Expanding - 3Y  | -0.155      | [-0.290, -0.056]    | 0.00504  |
| Error D_eff    | Expanding - 5Y  | +0.034      | [-0.071, +0.138]    | 0.353    |
| Error D_eff    | Expanding - 10Y | -0.164      | [-0.344, -0.017]    | 0.029    |

Table 2. Alternative concentration and error-diversity effects. Effects are expanding minus rolling; horizons are averaged within market before the four market blocks are resampled. Positive forecast correlation and PC1 effects indicate greater output concentration; negative error D\_eff indicates lower error diversity. With only four market blocks, intervals are diagnostic rather than high-power cross-sectional inference.

#### 4.5 Mechanical-overlap and transparent-feature exclusion tests

Two exclusion exercises address whether the frontier is merely a mechanical consequence of using more observations or of relying on the author-specific distributional feature construction. First, I equalize training-set size separately against the 3-, 5-, and 10-year rolling learners. At each refit, an equal-N broad-history diagnostic selects exactly N observations from the full eligible history, where N equals the corresponding rolling-window sample count. The selected row indices are deterministic and approximately equally spaced from the first eligible observation to the most recent eligible observation, including the endpoints when feasible. Thus sample count is held fixed while historical span and composition change. This is a diagnostic exclusion design, not a proposed deployable forecasting architecture.

The result holds for every window comparison. Broad equal-N history has lower QLIKE, lower D\_eff, and higher average forecast correlation in 16/16 cells relative to each matched rolling window. Market-block bootstrap intervals, formed after averaging horizons within each market, also exclude zero in the predicted direction for all nine equal-N contrasts; Appendix B reports the point estimates and intervals. Equal observation count therefore does not remove the low-loss, low-diversity contrast associated with broader historical span. As elsewhere, only four market blocks are available, so the intervals are best read as uncertainty diagnostics rather than as large-cross-section inference.

Second, I discard all location, scale, shape, tail-shape, and speed coordinates and rerun the complete grid using only the transparent RV-based feature set defined in Section 3.1. Expanding history again has lower QLIKE than each 3Y, 5Y, and 10Y alternative in 16/16 cells and lower D\_eff in 16/16 cells. Mean expanding-minus-

rolling differences in QLIKE are -0.0206, -0.0102, and -0.0058 against 3Y, 5Y, and 10Y; the corresponding  $D_{\text{eff}}$  differences are -0.601, -0.428, and -0.338. The central result therefore does not depend on the unpublished distributional-coordinate construction.

## 5. Interpretation and Identification Limits

The findings distinguish two notions often conflated in financial AI. Predictive adaptation is a private objective: a learner minimizes its own forecast loss. Model diversity is a collective property: it describes the effective dimensionality of the system of forecasts. An updating rule can improve the first while weakening the second. Conversely, selective forgetting can restore heterogeneity while paying a forecast-loss penalty. Neither objective dominates without an explicit welfare or risk criterion.

External validity is intentionally narrow. All four empirical markets are major equity indices, so the results should not be extrapolated mechanically to rates, commodities, single stocks, or crypto assets. Cross-market equity co-movement does not enter  $D_{\text{eff}}$  directly because diversity is computed within each market across forecasting models, not across markets. It can nevertheless reduce the effective independence of the market blocks and can synchronize stress dates, which is why the stress analysis reports global-episode clustering. Extension across asset classes is a separate empirical test, not an implication of the current sample. Equal- $N$  and alternative-diversity confidence intervals are likewise formed at the market-block level; their role is to display effect-size uncertainty under the available four-market design, not to manufacture additional independent observations.

The four-market exact sign result ( $p=0.0625$ ) is also a useful limit on the strength of inference. It shows that the current equity-index sample has little power for a genuinely independent-replication claim even though every market points in the same direction. A decisive test of cross-market generality requires additional independent asset classes and instruments, not additional horizons from the same four indices.

The historical replay does not identify the causal effect of model diversity on observed market volatility. Historical prices were generated under the actual market ecology, not under the counterfactual forecast ensembles constructed here. The portfolio mapping therefore establishes only a mechanically defined channel from forecasts to hypothetical risk adjustment. A causal market-impact claim would require either deployment data with identified adoption shocks or a structural model of order flow and price impact.

## 6. Conclusion

Adaptive volatility forecasting has an architecture problem as well as an accuracy problem. Across four markets and four horizons, full-history learning systematically produces the lowest QLIKE among the tested memory rules, but it also produces the lowest effective model diversity. Finite-memory adaptation restores cross-model heterogeneity at a measurable predictive cost. Around stress transitions, adaptation architecture further determines how much diversity survives and how synchronized implied position changes become. The practical implication is not that long memory or short memory is universally preferable. It is that forecast loss and system diversity are separate design objectives, and an adaptive financial forecasting system should be evaluated on both.

### Appendix A. Global-event linkage sensitivity

Onsets are sorted by calendar date across markets. Consecutive market-specific onsets separated by no more than the stated linkage width are chain-linked transitively into one global episode; there is no additional ex ante cap on total episode span. A chain can therefore span slightly more than  $W$  days if each adjacent gap is within  $W$ . In the observed 22-onset set this does not create rolling “snowball” episodes: across  $W=5\text{--}30$  days, the

maximum total span of any linked episode is 18 calendar days. Bootstrap resampling is then performed at the global-episode level.

| Link width (days) | Global episodes | Max span | Estimate | 95% CI         | Bootstrap p |
|-------------------|-----------------|----------|----------|----------------|-------------|
| 5                 | 16              | 6        | 0.061    | [0.007, 0.126] | 0.026       |
| 10                | 14              | 18       | 0.061    | [0.007, 0.128] | 0.024       |
| 15                | 13              | 18       | 0.061    | [0.010, 0.125] | 0.022       |
| 20                | 12              | 18       | 0.061    | [0.011, 0.120] | 0.019       |
| 25                | 12              | 18       | 0.061    | [0.011, 0.119] | 0.019       |
| 30                | 12              | 18       | 0.061    | [0.010, 0.118] | 0.020       |

## Appendix B. Equal-N broad-history diagnostic

At each refit, let  $N$  be the number of eligible observations in the matched  $W$ -year rolling window and  $M$  the number in all history then observable. The equal- $N$  diagnostic selects  $N$  deterministic order-statistic positions  $\text{round}(\text{linspace}(0, M-1, N))$ ; any rounding duplicate is filled from the nearest unused chronological index. The same features, models, targets, and refit dates are then used. This intentionally nonlocal training set is only a diagnostic for separating sample count from historical span/composition.

| Matched window | $\Delta$ QLIKE [95% CI]    | $\Delta$ D_eff [95% CI] | $\Delta$ forecast corr. [95% CI] | Directional cells |
|----------------|----------------------------|-------------------------|----------------------------------|-------------------|
| 3Y             | -0.0919 [-0.1783, -0.0228] | -0.424 [-0.624, -0.244] | +0.201 [+0.120, +0.294]          | 16/16 each        |
| 5Y             | -0.0284 [-0.0565, -0.0082] | -0.276 [-0.329, -0.202] | +0.135 [+0.100, +0.159]          | 16/16 each        |
| 10Y            | -0.0076 [-0.0110, -0.0048] | -0.301 [-0.414, -0.187] | +0.143 [+0.090, +0.197]          | 16/16 each        |

## Declarations

**Declaration of Interest.** The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Funding.** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**Data Availability.** The replication package will provide the historical-replay code, transparent RV-only features, memory-grid specifications, equal- $N$  sampler, event definitions, 5–30-day linkage sensitivity code, and model outputs. Source market data subject to redistribution restrictions will be documented with acquisition instructions. The central frontier is reproducible without author-specific distributional coordinates using the RV-only specification in Section 4.5.

**AI Assistance.** Generative AI tools were used for research assistance, code review, and editorial support. All research design choices, statistical results, interpretations, and final manuscript text were reviewed and approved by the author.

## References

- Asaad, S., Hamidi, S. M., & Bereyhi, A. (2026). Regime-aware financial volatility forecasting via in-context learning. arXiv:2603.10299.
- Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4), 559–583. [https://doi.org/10.1016/0169-2070\(89\)90012-5](https://doi.org/10.1016/0169-2070(89)90012-5).
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), 174–196. <https://doi.org/10.1093/jjfinec/nbp001>.
- Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), Article 44, 1–37. <https://doi.org/10.1145/2523813>.
- Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? *Journal of Applied Econometrics*, 20(7), 873–889. <https://doi.org/10.1002/jae.800>.
- Lim, T., & Kim, J. H. (2026). Measuring volatility with LLMs: Implications for covariance structure and portfolio construction. *Finance Research Letters*, 106, 110293. <https://doi.org/10.1016/j.frl.2026.110293>.
- Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1), 246–256. <https://doi.org/10.1016/j.jeconom.2010.03.034>.
- Patton, A. J., & Zhang, H. (2026). Bespoke realized volatility: Tailored measures of risk for volatility prediction. *Journal of Econometrics*, 254, 106122. <https://doi.org/10.1016/j.jeconom.2025.106122>.

- Pesaran, M. H., & Timmermann, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics*, 137(1), 134–161. <https://doi.org/10.1016/j.jeconom.2006.03.010>.
- Saxena, M., Pangallo, M., Hommes, C., Caccioli, F., & del Rio-Chanona, R. M. (2026). Machine Spirits: Speculation and adaptation of LLM agents in asset markets. arXiv:2604.18602.
- Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (Vol. 1, pp. 135–196). Elsevier. [https://doi.org/10.1016/S1574-0706\(05\)01004-9](https://doi.org/10.1016/S1574-0706(05)01004-9).
- Xia, Y., You, P., Wang, T., Liu, F., Qi, H., Wu, X., & Zhang, S. (2026). Agentic Trading: When LLM agents meet financial markets. arXiv:2605.19337.