

Historical-Memory Equity Selection: A Comparative Study of Retrieval, Neural Prediction, and Adaptive Mixing

Rohil Gujarathi

`Rohil.wrkgmail.com`

Vivekanand Education Society's Institute of Technology

Sangeeta Oswal

Vivekanand Education Society's Institute of Technology

Vaibhav Goyal

Predixion AI

Anjali Yeole

Vivekanand Education Society's Institute of Technology

Research Article

Keywords: Historical analogue retrieval, Equity selection, Time-series representation, Portfolio evaluation, Selective prediction, Financial machine learning

Posted Date: September 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11116893/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Competing interest reported. V.G. is Chief Executive Officer of Predixion AI. R.G. previously worked as a Engineering Intern at Predixion AI; that employment was unrelated to this study. Predixion AI provided no data, funding, proprietary models, or commercial infrastructure for this research. The authors declare no other relevant financial or non-financial interests.

Historical-Memory Equity Selection: A Comparative Study of Retrieval, Neural Prediction, and Adaptive Mixing

Rohil Gujarathi^{1*}, Sangeeta Oswal¹, Vaibhav Goyal²,
Anjali Yeole¹

^{1*}Department of Artificial Intelligence and Data Science, Vivekanand Education Society's Institute of Technology, Mumbai, 400074, Maharashtra, India.

²Predixion AI, Mumbai, Maharashtra, India.

*Corresponding author(s). E-mail(s): Rohil.wrk@gmail.com;
Contributing authors: sangeeta.oswal@ves.ac.in; vaibhav@predixion.ai;
anjali.yeole@ves.ac.in;

Abstract

Motivated by recent progress in long-context sequence modelling, this study examines whether modern forecasting methods can support intermediate-horizon equity selection and whether a bounded numerical memory can make their evidence more inspectable. We implement a memory-centered framework in which forecasts are constructed from eligible historical records, their completed 63-session returns, and an explicit aggregation rule; the memory can operate independently or alongside neural predictors. The evaluation compares 16 policies across 103 equities in six markets, represented by 204 continuous market/run accounts over 1,568 union sessions from 2020 through 2025. Similarity memory averages the outcomes of 25 precedents selected by Euclidean distance between standardized, six-session-pooled annual feature windows. Controls include random eligible precedents, an unconditional historical mean, plain nearest neighbours, ridge and neural predictors, momentum, and passive equal weight. Passive equal weight has the highest mean-market Sharpe (0.636) and six-year annualized return (11.91%). Similarity memory records Sharpe 0.245 and annualized return 2.99%, below random memory (0.423) and the historical prior (0.352). All six prespecified memory contrasts are negative, and neither these comparisons nor the two gate-versus-mixture contrasts reject at the 5% level after Holm adjustment. A complete six-fold execution generated more than four million policy predictions;

its summaries report full required-forecast coverage and reproduce the primary analysis. The framework makes the retrieval calculation decomposable, but the multi-year evidence does not show that similarity selection improves portfolio performance.

Keywords: Historical analogue retrieval, Equity selection, Time-series representation, Portfolio evaluation, Selective prediction, Financial machine learning

JEL Classification: C45 , C52 , C58 , G11 , G17

1 Introduction

Recent work on Transformers raised a practical question for this study: can sequence models that represent long contexts also support equity decisions held for more than a few sessions? Financial forecasting makes that extension difficult. Outcomes arrive slowly, market regimes change, and a forecast that appears accurate on average may still produce poor selections after trading costs and path-dependent exits. We therefore study a 63-exchange-session target and evaluate the resulting signals through a common portfolio engine rather than treating forecast accuracy alone as sufficient evidence.

A second motivation was to make the basis of a decision easier to inspect. Instead of adding a language model or an open-ended retrieval-augmented generation layer, we use a bounded, task-specific memory of numerical market states and completed outcomes. The memory runs alongside the forecasting models and can also be evaluated on its own. For any admitted query, its calculation is defined by the eligible records, their distances, their realized forward returns, and the aggregation rule. This construction does not make the forecast correct by design, but it makes the retrieval contribution identifiable and, when the query and neighbour records are retained, reviewable at case level.

The empirical design separates two questions that are sometimes conflated. First, does the framework provide a reproducible route from historical evidence to a forecast and then to a position? Second, does similarity-based selection add investment value relative to simpler alternatives? The memory bank stores pooled annual feature windows and completed outcomes, admits records only after their labels have matured, and excludes exact same-ticker matches. At each query it selects 25 precedents and averages their subsequent returns. The resulting forecast can stand alone or be combined with a temporal Transformer or multilayer perceptron (MLP) through a fixed mixture or an adaptive gate. Random retrieval, an unconditional historical prior, conventional predictors, momentum, and passive equal weight provide matched references under the same execution rules.

The final evaluation spans 2020–2025 and maintains separate portfolios for six markets. Its most important result is also the one that differed from the earlier single-year development evidence. Passive equal weight has the strongest mean-market Sharpe, while similarity memory trails random memory by 0.1777 and the historical prior by 0.1070. Neither difference is statistically resolved, and every prespecified memory contrast has a negative point estimate. The experiment does not identify one cause for

this reversal: broader regime coverage, the chosen pooled representation, Euclidean distance, and the corrected production protocol may all matter. We therefore report it as negative evidence about the implemented similarity rule, not as a general rejection of historical memory.

The main contribution is a specified interface between memory construction, neighbour selection, forecast aggregation, optional neural mixing, and portfolio execution. It allows the retrieval mechanism to be audited separately from the neural predictor and tested against controls that preserve the same eligible information set. The empirical contribution is a matched comparison of 16 policies with paired inference for the same mean-market Sharpe statistic shown in the performance table. The Transformer and MLP are comparison backbones rather than new neural architectures, and the paper does not claim that the memory layer guarantees reliability, correctness, or regulatory compliance.

2 Related work

Multivariate Transformers learn contextual representations from temporal inputs (Zerveas et al. 2021), while PatchTST uses patching and channel-independent processing for forecasting (Nie et al. 2023). These approaches motivate comparisons, not an assumption that a forecasting representation automatically defines a useful distance space. The mean-pooled multivariate Transformer evaluated here is not a PatchTST replication: it averages six daily rows before feature projection and uses a single multivariate stream.

Retrieval augmentation separates a parametric model from an external collection of examples. Language-model retrieval (Borgeaud et al. 2022) and multi-horizon forecasting with retrieval augmentation (Yang et al. 2022) demonstrate ways to combine these information sources. Their success does not establish the incremental value of equity analogues, where outcomes are delayed and market conditions change. We therefore compare similarity-based retrieval with random eligible records and a historical-prior policy rather than treat retrieval itself as evidence of useful information.

Momentum provides an established financial reference (Jegadeesh and Titman 1993), while its risk properties vary across episodes (Barroso and Santa-Clara 2015; Daniel and Moskowitz 2016). Here division by volatility is fixed across forecast-producing policies. It is an allocation transformation, not a security-level Sharpe estimate. A positive constant forecast already produces inverse-volatility ranking, making the historical-prior comparator important.

Model selection and backtest inspection can exaggerate apparent performance differences (Cawley and Talbot 2010; Bailey and López de Prado 2014). Serial dependence complicates Sharpe comparisons (Ledoit and Wolf 2008). The fixed-length moving-block construction follows the dependent-observation bootstrap framework of Künsch (1989); Politis and Romano (1994) provides related stationary-bootstrap context. The reported reanalysis uses synchronized moving calendar-week blocks and symmetric intervals, without establishing exact finite-sample coverage.

3 Data, predictions, and historical memory

3.1 Universe, observations, and chronology

The panel contains 103 retained equities: United States 18, India 18, China 18, Brazil 15, France 17, and United Kingdom 17. It was selected from 108 requested liquid securities using continuity and venue criteria. The accompanying identifier manifest records the retained and excluded securities. The panel is survivorship-conditioned rather than a point-in-time historical constituent universe.

Inputs comprise 23 technical price–volume features covering returns and momentum, volatility, volume, moving-average relations, drawdown, and price location. The scaler is fitted once per fold on all admitted training rows pooled across the six markets from 2013 through the fold cutoff, uses population standard deviations floored at 10^{-4} , and clips standardized values to $[-5, 5]$. Online Resource 1 provides the exact feature dictionary. The prediction window for a close- t decision excludes the current session and ends at $t - 1$:

$$X_{i,t} = [\tilde{x}_{i,t-252}, \dots, \tilde{x}_{i,t-1}]^\top \in \mathbb{R}^{252 \times 23}. \quad (1)$$

The pooled-window systems construct 42 non-overlapping six-session means,

$$\bar{X}_{i,t,p,f} = \frac{1}{6} \sum_{a=0}^5 X_{i,t,6p+a,f}, \quad p = 0, \dots, 41, \quad (2)$$

Here the row index of X is zero-based and $f = 1, \dots, 23$, yielding 966 values after flattening. This spans the full year but removes within-patch daily order. We therefore call it a *pooled input window*, not an uncompressed daily window. Let $C_{i,t}^{\text{model}}$ denote the causal model-price close used by feature and label construction, and let $O_{i,t}^{\text{trade}}$ and $C_{i,t}^{\text{trade}}$ denote the raw as-traded open and close used by the portfolio engine. The stored target is

$$r_{i,t}^{63} = \frac{C_{i,t+63}^{\text{model}}}{C_{i,t}^{\text{model}}} - 1 \quad (3)$$

on the native exchange calendar. This is a forecasting label and ranking proxy, not the payoff directly traded by the portfolio. A close- t signal is executed at $O_{i,t+1}^{\text{trade}}$, pays entry costs, and may exit before the 63-session ceiling through the trailing rule; consequently the label includes a close-to-next-open interval that is not executable after signal formation and ends at a different price and potentially different date from the realized trade. The reported portfolio returns are calculated from the actual simulated fills and exits, but the study does not claim that minimizing label error is identical to maximizing net traded return.

The final release uses six annual walk-forward folds. For evaluation year $y \in \{2020, \dots, 2025\}$, neural training records and memory-bank origins end by 31 December of year $y - 3$, validation queries use year $y - 2$, and mixture/gate development uses year $y - 1$. Neural checkpoints use seeds 7, 17, and 37. Completed 63-session labels and 126-session memory-maturity windows must be available by the applicable cutoff. Account state is continuous across fold boundaries rather than restarted annually. The release contains 204 market/run accounts and 1,568 union sessions from 1 January

2020 through 31 December 2025. The repaired query identities restore admission after the fold boundary, but the protocol and evaluation history were inspected during development; the comparison is not presented as an untouched confirmation test.

The model-price adapter defines a base-100 causal series by

$$C_{i,t}^{\text{model}} = C_{i,t-1}^{\text{model}} \frac{S_{i,t} C_{i,t}^{\text{trade}} + D_{i,t}}{C_{i,t-1}^{\text{trade}}}, \quad (4)$$

where $S_{i,t}$ is the split ratio and $D_{i,t}$ is the cash distribution per pre-action share. Features and targets use the corresponding model-price OHLC and split-normalized volume, whereas fills, valuation, and stops use raw as-traded prices. In the public production runner, this adapter is called with an empty action list and quote-unit multiplier 1.0; the tracked release therefore does not demonstrate ingestion of nontrivial split, dividend, or quote-unit records. The vendor cache was assembled retrospectively rather than preserved as contemporaneous point-in-time snapshots. The universe is not a historical constituent panel and no point-in-time-universe claim is made. The results remain research simulations with data-vintage, corporate-action, and survivorship limitations.

3.2 Memory-only prediction and selection controls

A historical record contains an identifier, market, origin, pooled window, and completed forward outcomes. Each fold constructs its memory bank only from history permitted by that fold’s chronology. Admission retains a completed 126-session outcome window on the precedent’s exchange calendar, even when the predictor uses only the 63-session return. Required outcome availability must precede the query; an origin-date cutoff alone is insufficient.

The global eligible pool excludes exact same-ticker records. Selected origins from the same ticker must be at least 21 native sessions apart, with at most three selected precedents per ticker. These controls do not establish issuer/affiliate-wide isolation or independence of overlapping 63-session outcomes. Let $u_{i,t} = \text{vec}(\bar{X}_{i,t})$. Similarity retrieval scans records in increasing distance

$$d(i, t; j) = \|u_{i,t} - u_j\|_2^2 \quad (5)$$

and accepts the first $k = 25$ satisfying the restrictions. Its forecast is

$$\hat{r}_{i,t}^{\text{mem}} = \frac{1}{25} \sum_{j \in \mathcal{N}_{i,t}} r_j^{63}. \quad (6)$$

For every admitted query, Equation 6 is exactly decomposable into the eligibility checks, 25 selected record identifiers, their distances and stored returns, and the uniform aggregation weights. These objects provide a citation-like audit trail that documents the empirical basis of the forecast and permits the signal-formation process to be inspected and reconstructed. The sealed pipeline reports complete required-forecast coverage and fallback status, but the coverage manifest and query-level neighbor records are omitted

from the tracked compact release and therefore cannot be independently inspected there. Exhausting a provisional candidate buffer is not equivalent to exhausting the eligible pool; expansion and insufficient-neighbor handling are part of the fixed retrieval configuration.

Random memory replaces distance order with random priority and applies the same finite-neighbor restrictions. Seeds 7, 17, and 37 produce separate executed policies, not a forecast-averaged ensemble. Its pseudorandom stream is derived from the bank identity, fold year, corrected query identifier, and master seed; the final recovery regenerated this control where the earlier fold-prefix defect changed that seed. This is constrained random-priority retrieval, not necessarily uniform sampling over all feasible sets.

Plain KNN uses the same 966-dimensional standardized pooled input, squared Euclidean distance, $k = 25$, arithmetic target mean, exact same-ticker exclusion, record-ID tie break, and unconditional-mean fallback as similarity memory, but removes the three-per-ticker cap and 21-session within-ticker spacing rule. Annual ridge is refit once per fold on admitted training records through 31 December of $y - 3$. It predicts the 63-session model-price return from the 966-dimensional pooled vector using deterministic float64 ridge regression, equal total weight for each represented market, penalty $0.001\|\beta\|_2^2$, and an unpenalized intercept; it has no seed realization.

The historical-prior policy supplies the unconditional memory-bank target mean to every query without nearest-neighbor selection. With a fixed fold-specific bank, its constant forecast has undefined Spearman correlation and, after volatility normalization, induces inverse-volatility ranking when positive. It is an unconditional economic baseline that differs from similarity memory in localization and query-ticker eligibility; it is not an isolating test of local conditioning. Figure 1 summarizes the comparison.

3.3 Neural predictors and mixtures

The frozen temporal Transformer accepts $(B, 42, 23)$ pooled windows, projects each patch to width 64, adds sinusoidal positions, and applies two encoder layers with four heads, feed-forward width 128, GELU, and dropout 0.1. Attention pooling is followed by a 128-dimensional linear projection, LayerNorm, and a scalar head. The MLP consumes the row-major flattened annual representation in $\mathbb{R}^{966}$ and uses $966 \rightarrow 64 \rightarrow 128 \rightarrow 1$: LayerNorm follows each hidden linear map, with a single GELU after the first LayerNorm. Training, bank construction, validation, and query evaluation share the identical standardized-and-clipped 42-patch representation. Both backbones output decimal-return predictions; frozen checkpoints use seeds 7, 17, and 37. Online Resource 1 specifies checkpoint selection and distinguishes the final 966-input MLP from the legacy single-patch branch.

A legacy input diagnostic found maximum batch-dependent latent and prediction differences of 1.46×10^{-6} and 1.19×10^{-7} in evaluation mode. This supports the tested tensor implementation, not predictive generalization. Transformer latents varied across observations, but their centered leading component accounted for approximately 64–75% of variance. Such concentration motivates testing the metric, not declaring that attention necessarily destroyed information.

Fixed mixtures use

$$\hat{r}_{i,t}^\lambda = (1 - \lambda)\hat{r}_{i,t}^{\text{base}} + \lambda\hat{r}_{i,t}^{\text{mem}}. \quad (7)$$

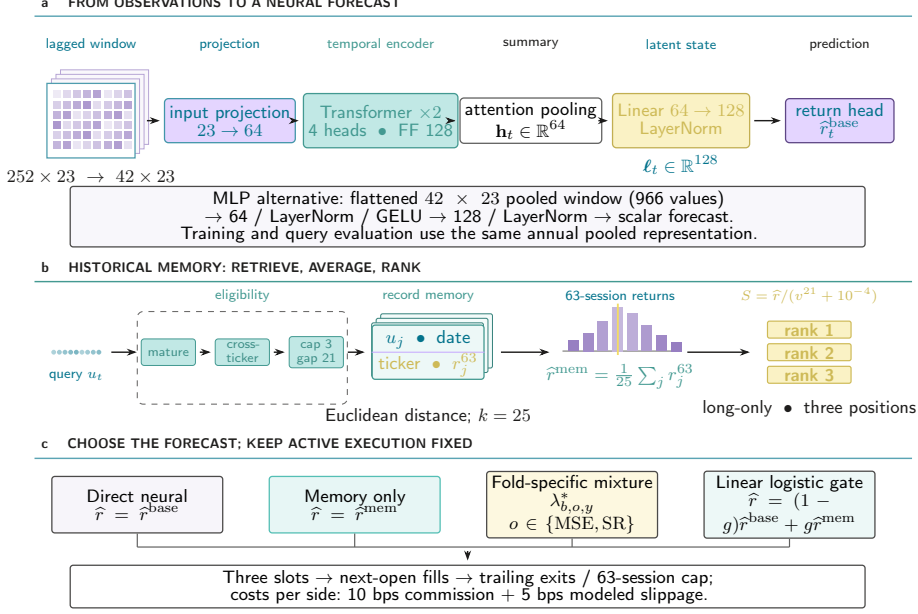


Fig. 1 From observations to forecasts and positions. (a) Both backbones use the 42-patch annual representation; the MLP flattens it to 966 values. (b) Memory retrieves the same pooled input and takes a uniform return mean; no learned latent is required. (c) Four alternative forecasts share active execution. The histogram is schematic

The development grid is $\{0, 0.10, 0.25, 0.50, 1\}$. For backbone b , objective o , and evaluation year y , write the selected coefficient as $\lambda_{b,o,y}^*$. The MSE objective is

$$J_{b,y}^{\text{MSE}}(\lambda) = \frac{1}{6} \sum_{m=1}^6 \frac{1}{3} \sum_{s \in \{7, 17, 37\}} \frac{1}{|\mathcal{D}_{m,y-1}|} \sum_{q \in \mathcal{D}_{m,y-1}} [(1 - \lambda) \hat{r}_{b,s,q}^{\text{base}} + \lambda \hat{r}_q^{\text{mem}} - r_q^{63}]^2, \quad (8)$$

where $\mathcal{D}_{m,y-1}$ contains admitted development queries whose labels mature by the development cutoff. The Sharpe objective analogously averages development-market net portfolio Sharpe across seeds and markets. The smallest grid value wins ties within tolerance 10^{-12} . Coefficients are frozen before evaluation. The compact tracked release does not export the numerical $\lambda_{b,o,y}^*$ table, so the figures and text do not assign single-year scalar values to the six-fold policies; exact coefficient recovery requires the excluded fold-level prediction/selection artifacts.

The selective gate substitutes $g_{i,t} = \text{sigmoid}(a^\top z_{i,t} + b)$ for λ , where $z_{i,t} \in \mathbb{R}^3$ contains standardized gate diagnostics, not the neural latent. Its ordered raw vector is $(|\hat{r}_{\text{base}}|, |\hat{r}_{\text{base}} - \hat{r}_{\text{mem}}|, |\hat{r}_{\text{mem}}|)^\top$. Each channel is normalized using fold-specific development means and population standard deviations floored at 10^{-4} before the linear logistic map. Trailing asset volatility enters allocation separately. These magnitude and disagreement diagnostics are not calibrated epistemic uncertainty. Gate fitting minimizes equal-market weighted mixed-forecast squared error on the development year immediately preceding each evaluation fold. Initialization is $a = 0$, $b = -1.0986$; Adam

uses learning rate 0.05, weight decay 10^{-3} , and 50 steps. Parameters and transformations remain fixed during the associated evaluation year. Gate fitting targets prediction error, whereas the Sharpe-selected constant mixture targets net Sharpe: these are implemented alternatives, not capacity comparisons under identical objectives.

3.4 Shared allocation and execution

Forecast-producing arms use

$$S_{i,t} = \frac{\hat{r}_{i,t}}{v_{i,t}^{21} + 10^{-4}}, \quad (9)$$

Here v^{21} is the unannualized sample standard deviation (ddof = 1) of 21 consecutive model-price close returns. Evaluation queries require a valid 252-session feature window, so undefined volatility values are excluded rather than filled at execution. No memory-tail penalty is added in the final comparison. Active portfolios have at most three slots, with daily close signals executed at the next recorded as-traded open. Only finite positive scores are eligible; candidates are ranked by descending score with security identifier as the deterministic secondary key. Holdings are not continuously rebalanced. Online Resource 1 specifies slot funding, affordability, holding-age boundaries and accounting measures.

Each market account starts with nominal capital of 100,000 account units. The common engine uses a close-triggered percentage trailing exit, a one-session minimum stop hold, and a 63-session holding ceiling. With current peak P_t and ATR-to-price ratio a_t , the stop condition is

$$\frac{C_{i,t}^{\text{trade}} - P_{i,t}}{P_{i,t}} < -\max(0.10, 2.5a_{i,t}). \quad (10)$$

This is the rule labeled “Chandelier” in the reports, not an intraday guaranteed-price stop. Each side applies 10 basis points commission and 5 basis points modeled slippage. Spreads beyond that slippage allowance, taxes, nonlinear impact, and realistic capacity are not separately modeled.

Final open positions are marked to the final close for NAV metrics. Synthetic calendar-end trade records, where retained for trade summaries, deduct hypothetical exit fees without cash-settled liquidation. Their P&L differs from terminal NAV by those hypothetical fees. The accounting identity reconciles these conventions; it is not an independent performance test. We emphasize NAV-based portfolio outcomes rather than interpret synthetic closeouts as actual fills.

The conventional controls include direct 21-session momentum, volatility-scaled 21-session momentum, and passive equal-weight buy-and-hold. Momentum uses $C_t^{\text{model}}/C_{t-21}^{\text{model}} - 1$ directly as its score; the volatility-scaled variant divides that score by $v_t^{21} + 10^{-4}$. They are economic references, not exact architecture ablations. A positive constant prior under Equation 9 ranks stocks by inverse volatility; it does not size every holding inversely to volatility. Online Resource 1 gives the complete 16-policy contract, preprocessing, missing-query behavior, realization count, ranking, and execution rules.

3.4.1 A worked positive-signal example

Figure 2 follows one hypothetical query from eligible records to a positive ranking score. Suppose 25 eligible precedents comprise five records each with 63-session returns of -4% , 2% , 6% , 14% , and 22% . Their uniform mean is 8% . With a base forecast of 3% and a given gate output of 0.60 , the mixed forecast is 6% . Allocation volatility of 0.015 gives $S = 0.06 / (0.015 + 10^{-4}) \approx 3.974$. This can earn an entry only relative to competing scores and available slots. The example is arithmetic, not a fitted gate replay or an observed profitable trade; a positive forecast is not a known future return.

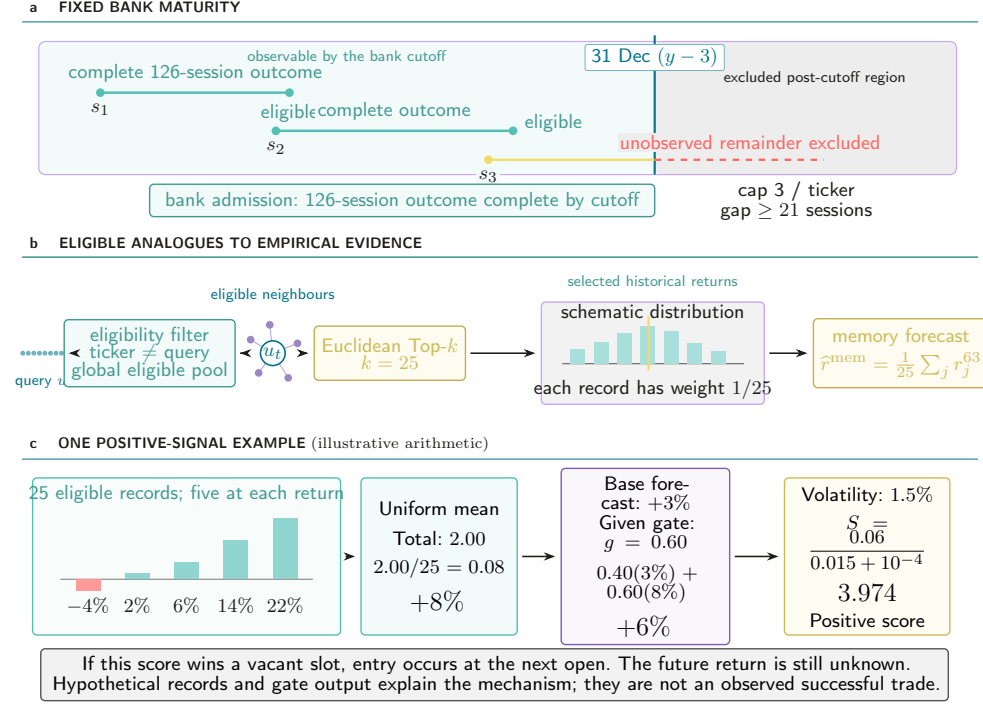


Fig. 2 From eligible history to a positive signal. Outcome maturity governs bank admission (a); ticker and spacing controls constrain Euclidean neighbor selection (b). The displayed cutoff is the fold-specific bank cutoff. Panel (c) is the canonical illustrative example: memory 8% , base 3% , and gate 0.60 yield a 6% forecast, not a known future profit. Distribution shapes are schematic and example values are hypothetical

4 Evaluation and statistical comparison

4.1 Portfolio and forecast criteria

Net Sharpe is the primary portfolio endpoint; annualized return, maximum drawdown, and trade activity describe economic behavior. Metrics are calculated within each continuous local-market run. Neural policies have three checkpoint realizations, random

memory has three retrieval realizations, and deterministic policies are reported once per market. The final report contains 204 continuous market/run accounts across 16 policies, not 204 independent market histories.

For policy A , let $R_{A,m}$ be its number of realizations in market m . Displayed mean Sharpe is

$$T(A) = \frac{1}{6} \sum_{m=1}^6 \left[\frac{1}{R_{A,m}} \sum_{r=1}^{R_{A,m}} \text{SR}_{A,m,r} \right]. \quad (11)$$

Run-level Sharpe uses daily net returns, a zero return benchmark, population standard deviation ($\text{ddof} = 0$), and the common annualization factor $\sqrt{252}$. The displayed return is the compound annual growth rate over the fixed six-calendar-year evaluation span. For N observed sessions with mean return $\bar{r}$ and $\sigma_r^2 = N^{-1} \sum_t (r_t - \bar{r})^2$,

$$\text{SR} = \sqrt{252} \frac{\bar{r}}{\sigma_r + 10^{-9}}, \quad \text{CAGR}_6 = \exp\left(\frac{1}{6} \sum_{t=1}^N \log(1 + r_t)\right) - 1. \quad (12)$$

The canonical run-summary function sets Sharpe to zero when $\sigma_r \leq 10^{-8}$; Online Resource 1 distinguishes the development-engine and bootstrap numerical conventions. Daily win rate is $N^{-1} \sum_t \mathbf{1}\{r_t > 0\}$, not the fraction of profitable closed trades. The weights in Equation 11 are statistical: there is no common-currency global portfolio or FX hedge behind the mean. Return and drawdown are likewise summaries of run-level metrics, not outcomes of an averaged NAV path.

Legacy forecast criteria are mean squared error (MSE) in squared decimal-return units and Spearman correlation pooled across asset-date predictions within each market/run. The 2020–2025 release used here provides the continuous portfolio return series and primary portfolio contrasts; it does not promote the earlier 2024 forecast summaries to six-year validation. Portfolio performance therefore should not be interpreted as a corresponding ranking of forecast accuracy.

4.2 Paired inference for the displayed statistic

The eight primary contrasts are similarity memory versus random memory (P1), historical prior (P2), plain nearest neighbors (P3), annual ridge (P4), direct MLP (P5), and direct Transformer (P6), followed by MLP gate versus its Sharpe-selected mixture (P7) and Transformer gate versus its Sharpe-selected mixture (P8). Each original-sample estimate is $\hat{\Delta}_{A,B} = T(A) - T(B)$ on the same sample.

Original-sample and bootstrap contrasts use Equation 11, not the Sharpe of cross-market averaged returns. The latter depends on cross-market covariance and cannot be converted using a universal $\sqrt{6}$ factor. The reported unrounded identity check is $|\hat{\Delta} - [T(A) - T(B)]| \leq 10^{-10}$.

The reanalysis uses 10,000 paired moving-block draws with four consecutive calendar weeks per block, following fixed-length moving-block resampling for dependent observations (Künsch 1989). Returns from 1 January 2020 through 31 December 2025 are grouped into 314 Monday-boundary calendar weeks. For each account and week, the analysis stores the sum of returns, squared returns, log gross returns, and observed

session count, producing a $204 \times 314 \times 4$ sufficient-statistics array. Sampled week indices are synchronized across policies, markets, and realizations. Counts retain each exchange’s observed sessions rather than assuming five observations per week. Blocks do not wrap; concatenated blocks are truncated to 314 sampled weeks. The saved random seed is 42.

Each draw recomputes run-level Sharpe and six-year CAGR from the sampled sufficient statistics and aggregates them as in Equation 11. Write $e^{(b)} = \Delta^{(b)} - \hat{\Delta}$ for the centered difference in draw b , with $B = 10,000$. The first-order centered symmetric marginal interval and two-sided p-value are

$$c_{0.95} = Q_{0.95}(\{|e^{(b)}|\}_{b=1}^B), \quad \text{CI}_{0.95} = [\hat{\Delta} - c_{0.95}, \hat{\Delta} + c_{0.95}], \quad (13)$$

$$p = \frac{1 + \sum_{b=1}^B \mathbf{1}\{|e^{(b)}| \geq |\hat{\Delta}|\}}{B + 1}. \quad (14)$$

Here $Q_{0.95}$ denotes the empirical 95th percentile with linear interpolation. These are not studentized or simultaneous intervals. Holm adjustment covers the eight primary Sharpe p-values P1–P8 jointly (Holm 1979). Neither these intervals nor resampling makes previously inspected periods an untouched holdout or corrects survivorship-conditioned universe construction.

5 Results

5.1 Final policy comparison

Table 1 reports the production-verified 2020–2025 configurations. Passive equal weight has the highest mean-market Sharpe, 0.636, and annualized return, 11.91%. Annual ridge follows at Sharpe 0.444. Pure pooled-window memory records annualized return 2.99%, Sharpe 0.245, and mean maximum drawdown -35.31% . These are point-estimate rankings, not established expected-performance differences.

Random memory has Sharpe 0.423 and the historical prior 0.352, both above similarity memory at 0.245. Memory’s differences are -0.1777 versus random memory and -0.1070 versus the prior. Plain nearest neighbors, annual ridge, direct MLP, and direct Transformer also have higher mean-market Sharpe point estimates. This ordering indicates that the implemented similarity rule did not add portfolio value over its matched controls across the production-verified period; it is not a causal decomposition of returns into risk and predictive components.

Memory is also not best on the reported risk descriptors. Its mean maximum drawdown is -35.31% , compared with -34.93% for passive equal weight, -33.64% for random memory, and -34.37% for the historical prior. The Sharpe-selected MLP mixture has the smallest drawdown magnitude among the active policies in Table 1, but the table does not establish a statistically superior expected drawdown.

5.2 Incremental performance remains uncertain

All primary intervals span zero (Table 2; Figure 3). Holm-adjusted p-values range from 0.7487 to 1.0000. P4, memory minus annual ridge, is the closest unadjusted

Table 1 Final 2020–2025 multi-year comparison. Entries are equally weighted market summaries after averaging realizations within markets. Return is CAGR over the fixed six-calendar-year span; drawdown is the mean run-level maximum drawdown; and the final column is the mean number of ledger trade records, including synthetic terminal closeouts, across market/run accounts. Returns and drawdowns are percentages. Ordering does not imply statistical superiority.

System	CAGR (%)	Sharpe	Max drawdown (%)	Ledger records
Passive equal weight	11.91	0.636	−34.93	33.0
Ridge annual	7.20	0.444	−39.01	180.0
MLP mixture (Sharpe)	6.16	0.425	−29.79	171.3
Random memory	6.21	0.423	−33.64	161.1
MLP gate	6.41	0.416	−34.42	169.2
Transformer gate	5.96	0.412	−36.64	171.9
Transformer mixture (MSE)	5.19	0.393	−36.32	171.9
MLP direct	6.21	0.383	−36.79	173.1
Historical prior	4.44	0.352	−34.37	157.3
Momentum 21	5.33	0.349	−35.42	173.0
Transformer mixture (Sharpe)	4.37	0.312	−36.36	180.6
KNN plain	4.01	0.295	−37.64	172.7
Transformer direct	3.82	0.274	−42.80	181.3
MLP mixture (MSE)	3.12	0.250	−36.42	170.6
Memory only	2.99	0.245	−35.31	172.0
Volatility-scaled momentum 21	1.26	0.146	−38.97	167.0

Notes. The six-year comparison contains 204 continuous market/run accounts over 1,568 union sessions. Neural and random-memory policies are averaged across their released realizations before equal market weighting. The 2024 component tables in Online Resource 1 are retained as explicitly legacy development evidence and do not replace these multi-year results.

Table 2 Primary differences in mean market Sharpe over 2020–2025. Marginal symmetric 95% intervals use four-week calendar blocks and 10,000 draws; Holm adjustment covers P1–P8 jointly.

ID	Candidate minus comparator	Δ SR	95% interval	Raw p	Holm p
P1	Memory – random	−0.1777	[−0.4156, 0.0602]	0.1416	0.9911
P2	Memory – prior	−0.1070	[−0.3901, 0.1761]	0.4508	1.0000
P3	Memory – KNN plain	−0.0497	[−0.3043, 0.2049]	0.7061	1.0000
P4	Memory – ridge annual	−0.1984	[−0.4318, 0.0350]	0.0936	0.7487
P5	Memory – MLP direct	−0.1380	[−0.3819, 0.1059]	0.2686	1.0000
P6	Memory – Transformer direct	−0.0290	[−0.2580, 0.2000]	0.7987	1.0000
P7	MLP gate – MLP mixture (Sharpe)	−0.0091	[−0.1940, 0.1759]	0.9200	1.0000
P8	Transformer gate – Transformer mixture (Sharpe)	0.0996	[−0.0689, 0.2681]	0.2415	1.0000

comparison to conventional significance, but its estimate is negative (−0.1984), its interval is [−0.4318, 0.0350], and its adjusted p -value is 0.7487. P8 is the only positive point estimate: the Transformer gate exceeds its Sharpe-selected mixture by 0.0996, with interval [−0.0689, 0.2681] and raw $p = 0.2415$. Neither similarity selection nor adaptive gating is established as superior on the primary endpoint. Conversely, the intervals do not establish equivalence or exclude economically relevant differences.

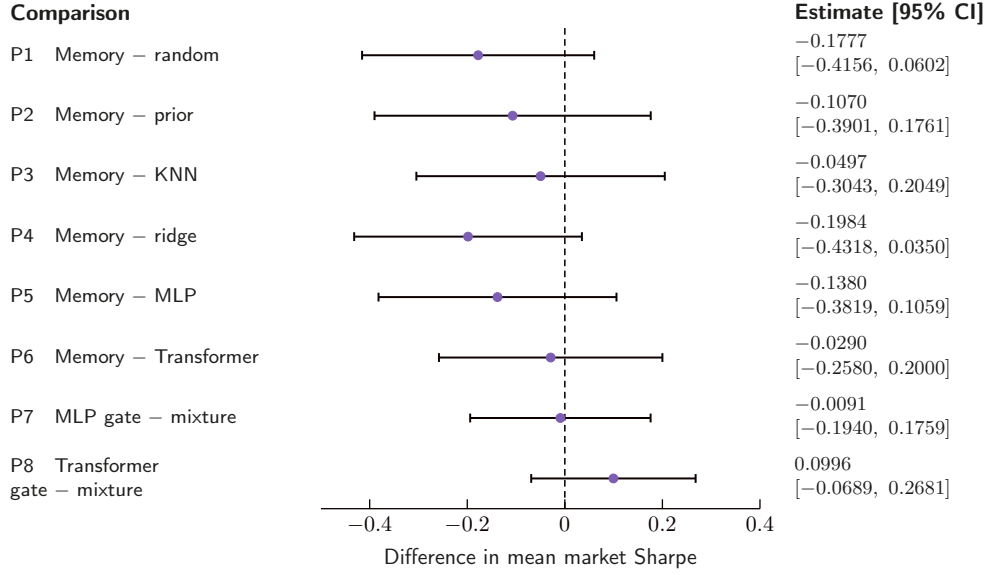


Fig. 3 Primary differences in mean market Sharpe over 2020–2025. Circles mark original-sample estimates; capped lines show marginal symmetric 95% block-bootstrap intervals; the dashed reference denotes zero. All intervals span zero; Holm-adjusted inference appears in Table 2

The final release reports the four-week primary specification for P1–P8. Earlier two- and eight-week sensitivity tables belong to the single-year legacy analysis and are not presented as sensitivity evidence for the production-verified six-year comparison.

5.3 Production verification

On 18 September 2026, the complete six-fold workflow was executed through the production runner using the authenticated 36 neural checkpoints and six ridge baselines, without neural retraining. The run generated 4,066,060 policy-prediction records across 119,590 fold-specific evaluation queries, with no missing required forecasts and no non-finite predictions. It then reconstructed all 204 continuous market/run accounts over the same 1,568-session horizon. The resulting P1–P8 point estimates, confidence intervals, and raw and Holm-adjusted probabilities reproduced the corrected recovery analysis; the archived-return replay matched the released analysis at tolerance 10^{-10} . This production execution is a computational confirmation of the reported comparison rather than an additional independent market sample.

5.4 Market heterogeneity

Table 3 distinguishes aggregate from uniform market leadership. Similarity memory has positive Sharpe in Brazil, India, the United Kingdom, and the United States, and negative Sharpe in China and France. It exceeds both the prior and random-memory

Table 3 Sharpe by market for final candidates and matched controls over 2020–2025, averaging realizations within markets. Positive Sharpe is not necessarily a benchmark win.

System	Brazil	China	France	India	UK	US
Passive equal weight	0.428	0.481	0.446	1.223	0.173	1.067
Ridge annual	0.375	−0.315	0.021	1.068	0.363	1.152
MLP mixture (Sharpe)	0.393	−0.116	0.292	0.796	0.426	0.758
Random memory	0.386	−0.317	0.523	0.600	0.561	0.785
MLP gate	0.147	−0.008	0.338	0.756	0.229	1.031
Transformer gate	0.403	−0.022	0.037	0.982	0.095	0.975
MLP direct	0.546	−0.179	0.236	0.657	0.101	0.938
Historical prior	0.466	−0.367	0.269	0.775	0.372	0.599
Momentum 21	0.512	−0.086	−0.044	0.715	−0.063	1.061
Transformer direct	−0.130	−0.087	−0.213	0.759	0.198	1.119
Memory only	0.341	−0.229	−0.086	0.336	0.236	0.874

control only in China and the United States. Positive levels are not benchmark wins or evidence of positive alpha.

The prior-relative difference is approximately +0.275 in the United States and +0.138 in China, but negative in the other four markets. Production accounting reports no missing required forecasts for admitted queries. The tracked bundle does not include the full coverage manifest, however, and the daily return series contains extended flat segments for some China and India market-years. Those segments can reflect cash-only portfolios or periods without admitted eligible queries; they are not interpreted as evidence of stable market-specific performance without the corresponding coverage record. No new market-specific selection rules are fitted to these outcomes.

5.5 Legacy development ablations

These development results precede the MLP representation correction and are retained as legacy exploratory evidence, not evaluations of the corrected MLP. Component studies are separate branches, not a cumulative stack retaining every module. The first mean-only comparison fixes mixing at 0.25. Transformer pooled-input memory yields Sharpe 0.410, versus 0.100 for uniform latent-cosine memory and 0.275 for direct prediction. MLP values are 0.659, 0.595, and 0.758. Neither latent retrieval nor unconditional mixing benefits every backbone.

A learned 32-dimensional key raises Transformer latent-retrieval Sharpe from 0.100 to 0.253, below the pooled-input reference. On MLP the learned-key estimate is 0.507, below random projection (0.694) and the original latent (0.595). This does not establish that projection resolves anisotropy or that all forecasting latents are unusable for retrieval.

Aggregation was tested on learned-key neighbors. Learned logit adjustment gives Transformer Sharpe 0.171, matching fixed softmax at displayed precision and below uniform weighting (0.253). Enriched values were instead tested with pooled-input neighbors. Additional 21-session returns leave the displayed portfolio metrics nearly

unchanged; drawdown enrichment does not improve Sharpe. These observations concern the tested configurations, not every aggregation model or additional outcome field.

In the legacy risk extension, a memory-tail coefficient of 0.20 reduces pure-memory Sharpe from 0.972 to 0.664 and increases drawdown magnitude; no such term enters the final comparison. Online Resource 1 preserves these branch tables descriptively. Their MLP results and earlier significance labels are not evidence for the corrected MLP configurations.

5.6 Legacy gate diagnostic

The release retains an unchanged pre-correction gate diagnostic of 25,753 observations. Online Resource 1 labels those bins as legacy evidence: they have not been regenerated for the corrected MLP gate and do not establish its mechanism. The corrected comparison supports only the aggregate forecast and portfolio results reported in Table 1.

6 Discussion

The study began from the possibility that longer-context sequence modelling could be extended to an intermediate investment horizon. The results do not show that using a Transformer, or placing historical memory beside it, is sufficient to produce a stronger portfolio. Across the production-verified multi-year matrix, memory-only selection does not lead the ranking and trails each of its six primary comparators. Similarity memory also underperforms random eligible retrieval and the unconditional historical prior on the displayed Sharpe point estimates. The additional local geometry therefore failed to improve the tested policy, even though it used the same matured information set as its closest controls.

This finding separates auditability from predictive value. A bounded numerical memory does not need a language-generation component to connect a forecast with historical cases. Where the query and neighbour records are retained, reviewers can examine the eligibility rule, selected episodes, distances, outcomes, and aggregation arithmetic. That evidence trail can support model review and governance, but it does not certify that the analogues are economically relevant or that the resulting decision is correct. In hybrid policies, the final account must also include the neural forecast, mixture or gate, volatility normalization, ranking, and execution rules.

The reversal from the earlier 2024 development ordering was the most consequential empirical outcome. The present design cannot attribute that reversal to a single mechanism. A fixed Euclidean metric may not remain meaningful across market regimes; six-session pooling removes within-patch order; completed 63-session returns may be weak labels for trades governed by next-open entry and path-dependent exits; and similarity in standardized technical features may not imply similarity in future returns. These are research hypotheses arising from the comparison, not explanations established by the reported experiments. They motivate testing alternative representations and retrieval metrics under the same matched-control design.

The production release establishes a portfolio ranking rather than a new six-year forecast-metric ranking. Legacy 2024 diagnostics suggested differences between

predictive and portfolio criteria, but those diagnostics should not be treated as multi-year confirmation. Portfolio outcomes depend on entry into a small selected set, holding paths, costs, and risk normalization, not simply average prediction error.

The Transformer comparison neither proves overparameterization nor establishes an architectural performance limit. Mean pooling, objective, regularization, and distance metric are specific to the implementation. Simple-model point estimates and anisotropy do not isolate these explanations. The Transformer gate has a higher Sharpe point estimate than its Sharpe-selected mixture, but P8 remains statistically unresolved; the MLP gate is slightly below its corresponding mixture in P7. Gate and selector objectives also differ.

6.1 Limitations

The production execution establishes computational completeness and numerical reproducibility for the reported comparison, but several features delimit external interpretation. The walk-forward years and protocol were inspected during development, and the curated universe is survivorship-conditioned rather than a point-in-time constituent panel. The 63-session close-to-close model-price label is a forecasting target rather than the exact next-open-to-exit realized payoff, and the retrospectively assembled cache does not demonstrate nontrivial corporate-action or quote-unit ingestion in the evaluated run. Uniform fees and omitted market impact, taxes, FX conversion, and capacity also simplify execution. Finally, complete forecast accounting applies to admitted queries; the untracked full coverage manifest limits external diagnosis of some flat market-year intervals, and the block bootstrap conditions on the observed markets and fitted policies. These qualifications define the scope of generalization and reconstruction rather than alter the internally reproduced portfolio comparison.

The public release separates tracked analytical results from the larger prediction and execution artifacts. The return-based replay recomputes return/Sharpe summaries and paired inference and matches the production analysis at tolerance 10^{-10} ; it is not an independent retraining exercise. Large prediction arrays, account ledgers, checkpoints, and the full coverage manifest remain outside the tracked publication-grade commit, while their production summaries and digests provide the documented audit trail.

Despite the development repository’s name, the evaluated method combines supervised/non-parametric prediction with a deterministic policy. It is not an offline reinforcement-learning agent, and comparisons with reinforcement-learning algorithms are not inferred.

7 Conclusion

We developed this study to test whether modern sequence models could support intermediate-horizon equity decisions and whether a task-specific historical memory could expose the evidence used alongside those models. The resulting framework links matured historical episodes, retrieval rules, neural forecasts, and portfolio execution without requiring a language model. Its components are separable, so the contribution of similarity retrieval can be reconstructed and tested rather than inferred from a final score alone.

The production-verified 2020–2025 comparison gives a cautious answer. Passive equal weight has the highest mean-market return and Sharpe point estimates. Pooled-input memory records Sharpe 0.245 and trails random memory, the historical prior, plain nearest neighbours, ridge, MLP, and Transformer controls. None of the eight primary contrasts establishes superiority at the 5% level after Holm adjustment. More than four million policy-prediction records and the archived-return replay reproduce this comparison, but they do not turn the inspected evaluation period into an independent confirmation sample.

The central lesson is that an inspectable memory mechanism and a profitable similarity signal are different achievements. The framework provides a disciplined way to study historical evidence and neural-memory interaction, whereas the current multi-year results provide no incremental portfolio advantage for the implemented analogue rule or adaptive gates. Future work should retain the same temporal controls and matched baselines while testing whether alternative state representations, distance measures, and outcome definitions make the retrieved cases economically informative.

Supplementary information. Online Resource 1 provides supplementary walk-forward fitting and execution methods, the final multi-year inference specification, the feature dictionary, release access instructions, and explicitly labeled legacy 2024 development evidence.

Statements and Declarations

Author contributions

Rohil Gujarathi: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, and writing—original draft.

Sangeeta Oswal: Supervision, project administration, methodology, and writing—review and editing.

Vaibhav Goyal: Methodology, supervision, validation, and writing—review and editing.

Anjali Yeole: Methodology, supervision, validation, and writing—review and editing.

Funding

No external, commercial, or private funding was received for this research. In-kind computational resources and institutional infrastructure were provided by Vivekanand Education Society’s Institute of Technology.

Competing interests

Vaibhav Goyal is Chief Executive Officer of Predixion AI. Rohil Gujarathi previously worked as a software developer at Predixion AI; that employment was unrelated to this study. Predixion AI provided no data, funding, proprietary models, or commercial infrastructure for this research, which was conducted independently of its corporate intellectual property. The authors declare no other competing financial or non-financial interests.

Use of AI-assisted tools

AI coding and writing assistants supported code refactoring, manuscript organization and editing, interpretation checks, bibliographic checks, and L^AT_EX preparation. AI assistance was not used to generate empirical observations or substitute for experimental validation. The authors reviewed the resulting material and take responsibility for the study design, implementation, results, interpretation, and final text.

Data and code availability

Online Resource 1 provides supplementary methods and results; it is distinct from the experimental archive. The final production code and tracked 2020–2025 analytical release are public in *Core-RL-Agent* at commit `fe56a39960c90ac7da3dce9973035f9b4f5baa80` (Gujarathi et al. 2026a). The release contains the production receipt, runtime configuration, six fold summaries and prediction manifests, portfolio manifest, P1–P8 outputs, replay report, pipeline-completion record, and release verification files. The run receipt identifies execution code revision `16ca345fd40a8cce76c9ac83ea7aef110c7c79a2`. The older `memory_study/run_final_comparison.py` and the *historical-memory-equity-data* v1.0.3 checkout describe upstream or legacy components rather than the final 204-account production contract (Gujarathi et al. 2026b). Large prediction arrays, account ledgers, decision logs, checkpoints, the full coverage manifest, and the fold-specific numerical mixture-coefficient table are excluded from the tracked commit. The archived-return replay verifies the published return analysis at tolerance 10^{-10} ; deeper fitting and execution reconstruction requires the excluded artifacts. The code is distributed under Apache-2.0; the data repository describes the scope of CC-BY-4.0 and source-data exclusions. Underlying vendor data remain subject to provider terms. No archival DOI is claimed.

Ethics approval

The study uses historical market data without human participants or animal experiments. Ethics approval is not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Acknowledgments

The authors acknowledge the Department of Artificial Intelligence and Data Science at Vivekanand Education Society’s Institute of Technology for institutional support and access to computational infrastructure.

References

- Bailey, D. H., & López de Prado, M. (2014). The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *The Journal of Portfolio Management*, 40(5), 94–107. <https://doi.org/10.3905/jpm.2014.40.5.094>
- Barroso, P., & Santa-Clara, P. (2015). Momentum has its moments. *Journal of Financial Economics*, 116(1), 111–120. <https://doi.org/10.1016/j.jfineco.2014.11.010>
- Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., ... Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning* (Vol. 162, pp. 2206–2240). PMLR. <https://proceedings.mlr.press/v162/borgeaud22a.html>
- Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11, 2079–2107. <https://www.jmlr.org/papers/v11/cawley10a.html>
- Daniel, K., & Moskowitz, T. J. (2016). Momentum crashes. *Journal of Financial Economics*, 122(2), 221–247. <https://doi.org/10.1016/j.jfineco.2015.12.002>
- Gujarathi, R., Oswal, S., & Goyal, V. (2026a). *Core-RL-Agent: 2020–2025 full production study and analytical release* [Software and derived results]. GitHub. Exact release revision: <https://github.com/Rohil72/Core-RL-Agent/tree/fe56a39960c90ac7da3dce9973035f9b4f5baa80>
- Gujarathi, R., Oswal, S., & Goyal, V. (2026b). *Historical-memory equity data* (release v1.0.3) [Upstream data and legacy study artifacts]. GitHub. Exact upstream revision: <https://github.com/Rohil72/historical-memory-equity-data/tree/0dba6f88504fffc802e915ff240177b515a2d8b9>
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
- Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1), 65–91. <https://doi.org/10.1111/j.1540-6261.1993.tb04702.x>
- Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3), 1217–1241. <https://doi.org/10.1214/aos/1176347265>
- Ledoit, O., & Wolf, M. (2008). Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance*, 15(5), 850–859. <https://doi.org/10.1016/j.jempfin.2008.03.002>
- Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=Jbdc0vTOcol>
- Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428), 1303–1313. <https://doi.org/10.1080/01621459.1994.10476870>

- Yang, S., Eisenach, C., & Madeka, D. (2022). MQ-ReTCNN: Multi-horizon time series forecasting with retrieval-augmentation. In *8th SIGKDD International Workshop on Mining and Learning from Time Series*. <https://assets.amazon.science/e2/de/b6d4e7dc4042b9f61ba5fed4d542/mq-retcnn-multi-horizon-time-series-forecasting-with-retrieval-augmentation.pdf>
- Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., & Eickhoff, C. (2021). A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 2114–2124). <https://doi.org/10.1145/3447548.3467401>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementarymaterial.pdf](#)