

MATHEMATICS FOR MACHINE LEARNING

**Decomposing Deep Momentum Performance:
When Does an LSTM Break or Shine Against Classical Trend-Following Rules?**

Author:

Cristóvão Matos (140686)

cristovao.matos@tecnico.ulisboa.pt

2025/2026 – 2nd Semester

2026/06/02

1 Introduction

Time-series momentum is one of the most widely studied systematic trading effects. Classical implementations, such as trend-following rules based on past returns or MACD signals, are attractive because they are simple, interpretable and easy to backtest. However, this simplicity also imposes rigidity: the trend estimator and position-sizing rule are fixed in advance, which may limit adaptation to regime changes, turning points or volatility shocks.

Deep learning offers a natural alternative: instead of specifying the trading rule manually, the model can learn a position function directly from historical market sequences. Recent work shows that LSTM-based momentum models can be trained to output trading positions and optimize a Sharpe-type objective directly [1]. More recent applied treatments of deep learning in quantitative trading also emphasize the move from return prediction toward direct trading-signal and portfolio-decision frameworks [5]. Wood et al. [4] extend this line of research with the Momentum Transformer, an attention-based architecture designed to improve adaptability and interpretability in deep momentum strategies.

This project does not attempt to reproduce the full Momentum Transformer architecture. Instead, it focuses deliberately on the LSTM component of the public *Trading with the Momentum Transformer* codebase and compares it with Long Only, classical time-series momentum (TSMOM), and MACD benchmarks. The objective is not architectural novelty, but empirical decomposition: does the LSTM provide a robust timing edge over classical momentum rules, or is its performance concentrated in particular regimes, random seeds and deployment windows?

This question matters because aggregate backtests can hide substantial dispersion. A strategy may appear attractive over a long sample while still producing unstable outcomes across realistic investment horizons. Following the general backtesting perspective that trading strategies should be evaluated through historical and walk-forward simulations rather than through a single aggregate path [3], we study many one-year deployment episodes across different market eras and random seeds.

The main contribution of the project is to decompose LSTM momentum performance rather than report only average results. First, we evaluate the LSTM and classical benchmarks on paired one-year episodes, separating central tendency from dispersion. Second, we study the left and right tails of LSTM relative performance to identify when the model breaks or shines. Third, we diagnose these tail episodes using position dynamics, classical-signal conflict, wrong-call rates, feature attribution, and instrument-level P&L contribution.

The central research question is therefore twofold: does the LSTM provide a robust risk-adjusted improvement over classical momentum strategies, and conditional on performance being regime-dependent, what distinguishes episodes where the LSTM breaks from episodes where it shines?

2 Methodology

We evaluate whether an LSTM-based momentum model adds value relative to classical momentum benchmarks, and under which conditions this edge appears or disappears. The analysis is out-of-sample and episode-based: instead of relying on a single aggregate backtest path, we evaluate many one-year deployment windows across four market eras.

2.1 Experimental Design

The empirical design follows a walk-forward backtesting logic. Models are trained only on information available before the evaluation period, validation is used for model selection, and the final test era is kept strictly out-of-sample. This separation is important in financial machine learning, where look-ahead bias and repeated testing on the same history can produce overstated results [2].

The test sample is divided into four non-overlapping market eras:

$$E_1 = 2005\text{--}2010, \quad E_2 = 2010\text{--}2015, \quad E_3 = 2015\text{--}2020, \quad E_4 = 2020\text{--}2025.$$

For each era, the LSTM is trained using an expanding historical window ending before the start of the corresponding test era. A validation split within the pre-test period is used for hyperparameter selection, early stopping and checkpoint selection. The test era is then used only to generate out-of-sample positions. No observations from the test era enter model training, validation, checkpoint selection or hyperparameter tuning.

To assess sensitivity to random initialization and training noise, five LSTM models are trained for each era using different random seeds. Each trained model generates daily positions over the full out-of-sample era. From these daily return series, we construct one-year deployment episodes of $T = 252$ trading days. For each era and seed, 200 episode start dates are sampled from the available out-of-sample window. Sampling is performed with replacement, and duplicate episode keys are removed before analysis. This yields approximately $M \approx 930$ unique episodes per era. The same episode windows are used for the LSTM and for all benchmarks, so all comparisons are paired.

2.2 Data and Instruments

The empirical analysis uses daily futures data obtained from Refinitiv. The starting point was the 50-instrument futures universe reported in Wood et al. [4]. In practice, only 40 instruments could be collected reliably from Refinitiv. From these, we selected a final subset of 26 instruments for the empirical study.

The selected subset preserves broad cross-asset exposure, covering equity indices, fixed income, energy, metals, grains, soft commodities and livestock. The same 26-instrument universe is used consistently across all four eras, avoiding changes in the cross-sectional universe across test periods. The full instrument list is reported in Appendix A.

2.3 Strategies

The LSTM strategy is the LSTM component of the public *Trading with the Momentum Transformer* codebase [4]. For each instrument i and decision date t , the model receives a rolling 63-day sequence of eight normalized momentum features:

$$X_{i,t} \in \mathbb{R}^{63 \times 8},$$

containing normalized daily, monthly, quarterly, biannual and annual returns, together with three MACD-style signals.

The model outputs a bounded position signal,

$$w_{i,t}^{LSTM} \in [-1, 1],$$

where $w_{i,t} = 1$ denotes full long exposure, $w_{i,t} = -1$ full short exposure, and $w_{i,t} = 0$ no exposure. The model therefore learns positions directly, rather than forecasting returns.

The benchmarks are Long Only, annual time-series momentum and MACD. Long Only keeps a constant long exposure. TSMOM uses the sign of annual momentum. MACD uses the average of three normalized MACD signals with the nonlinear position scaling from the reference implementation. Further architectural details, benchmark formulas and the LSTM training objective are reported in Appendices B and D.

2.4 Target Returns, Portfolio Construction and Instrument Allocation

All strategies are evaluated on volatility-scaled target returns. Let $r_{i,t+1}$ denote the next-day percentage return of instrument i , and let $\sigma_{i,t}^{ann}$ denote its estimated annualized volatility. The target return is

$$\tilde{r}_{i,t+1} = r_{i,t+1} \frac{\sigma_{\text{target}}}{\sigma_{i,t}^{ann}}, \quad \sigma_{\text{target}} = 15\%.$$

Thus, each instrument return is expressed in risk-normalized units. Price levels do not enter the strategy directly; only percentage returns and volatility-scaled momentum features are used.

Instrument-level captured returns are aggregated into an equal-weight cross-sectional portfolio. This construction is important for interpretation: the LSTM does not allocate capital across instruments. Each instrument receives the same base sleeve, and the LSTM controls only the signed exposure inside that sleeve. Therefore, a heatmap of LSTM positions should be interpreted as an instrument-level exposure signal, not as a cross-sectional capital-allocation map. Full construction details are reported in Appendix E.

2.5 Performance Metrics and Volatility Scaling

For each episode and strategy, we compute total return, annualized return, annualized volatility, Sharpe ratio, Sortino ratio, maximum drawdown, Calmar ratio, VaR, CVaR and hit rate. Since the objective is to compare timing skill rather than raw exposure to volatility, the main comparison uses risk-adjusted performance.

The main risk-adjusted metric is the annualized Sharpe ratio. Episode-level returns are also rescaled to a common 15% annualized volatility for ex-post comparability. This makes strategies more comparable across eras and models, since an apparently higher return may simply reflect higher realised volatility rather than better timing. This rescaling is used only as a diagnostic normalization of realised performance; it is not intended to represent an implementable ex-ante volatility-targeting rule. Full metric definitions are reported in Appendix E.

2.6 Relative Timing Edge and Tail Episode Selection

The main risk-adjusted comparison is the LSTM timing edge, defined as the LSTM Sharpe ratio minus the stronger of the two classical timing benchmarks, TSMOM and MACD, in the same one-year episode:

$$\Delta \text{Sharpe}_e^{\text{timing}} = \text{Sharpe}_e(\text{LSTM}) - \max\{\text{Sharpe}_e(\text{TSMOM}), \text{Sharpe}_e(\text{MACD})\}.$$

Positive values indicate that the LSTM outperformed both classical timing rules on a risk-adjusted basis, while negative values indicate that at least one classical timing rule performed better.

For the diagnostic tail analysis, episodes are selected using raw benchmark-relative total return rather than Sharpe ratio. This identifies economically material cases where the LSTM most clearly adds or destroys value

relative to feasible alternatives. Specifically, within each era we rank episodes by the LSTM’s raw return advantage over the best available benchmark, including Long Only, TSMOM and MACD.

From the full dataset of 3,722 unique one-year episodes, we select 160 diagnostic episodes: 20 breakdown and 20 outperformance episodes per era. This stratified selection ensures balanced regime coverage and prevents the diagnostic sample from being dominated by a single market period.

The selected episodes are then analysed using position dynamics, MACD–TSMOM signal conflict, wrong-call rates, feature-level attribution diagnostics and instrument-level P&L contribution. The exact tail-selection rule and diagnostic formulas are reported in Appendices F and G.

2.7 Code Availability

The code accompanying this report is available at:

<https://github.com/cristovaomatos/deep-momentum-lstm-diagnostics>

The repository contains the implementation used for data preparation, feature engineering, LSTM training, hyperparameter selection, episode construction, benchmark evaluation, tail diagnostics, Integrated Gradients attribution and figure generation.

3 Performance Analysis: Does the LSTM Improve Momentum Timing?

The first part of the empirical analysis evaluates whether the LSTM provides a robust risk-adjusted improvement over classical momentum timing rules. The relevant comparison is between the LSTM, TSMOM and MACD, since the LSTM is trained on normalized return and MACD-style features. Long Only is retained as a contextual benchmark, but not as the main measure of timing skill.

Equity-path dispersion. Figure 1 reports the median one-year equity path for each strategy across all unique deployment episodes in each era, together with dispersion bands for the LSTM. This figure shows that the LSTM has materially different behaviour across eras. In 2005–2010, Long Only delivers the strongest median path, while the LSTM remains below it despite positive performance. In 2010–2015, TSMOM is the strongest classical strategy, and the LSTM produces only modest median growth with a wide dispersion band. In 2015–2020, the LSTM becomes more competitive, clearly outperforming TSMOM and MACD and approaching Long Only. In 2020–2025, the LSTM has the strongest median path among all strategies, with substantial upside dispersion.

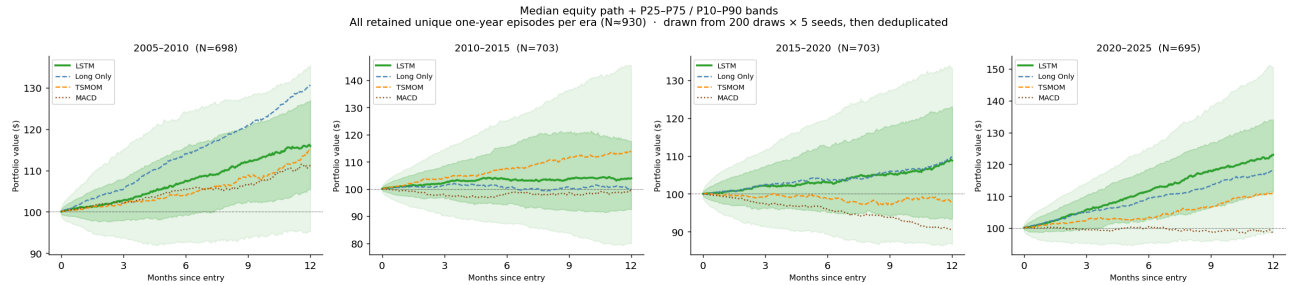


Figure 1: Median one-year equity path by era across all retained unique deployment episodes. The shaded bands report LSTM dispersion (P25–P75 and P10–P90).

The main insight from this artifact is that a single long-run average backtest would hide large entry-point variability. The LSTM does not produce a uniform edge across all periods; its performance depends strongly on the market era and the one-year deployment window.

Sharpe distribution across episodes. Figure 2 complements the equity-path analysis by showing the full distribution of episode-level Sharpe ratios. This confirms that the LSTM’s relative performance is regime-dependent. In 2005–2010, Long Only has the highest median Sharpe. In 2010–2015, TSMOM dominates, while the LSTM has a lower median and a wide distribution. In 2015–2020, the LSTM and Long Only are close, while TSMOM and MACD deteriorate. In 2020–2025, the LSTM achieves the highest median Sharpe.

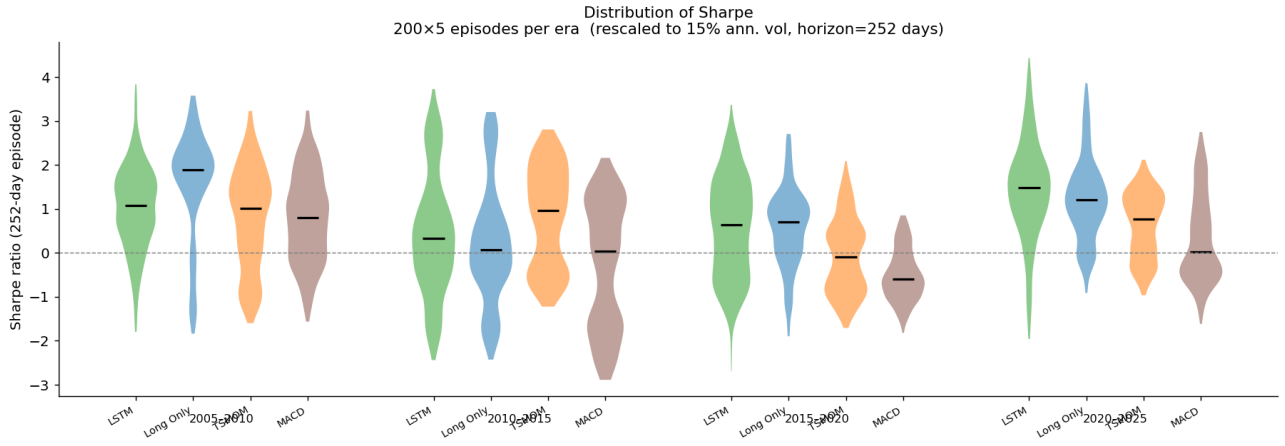


Figure 2: Distribution of one-year episode Sharpe ratios by era for the LSTM and benchmark strategies.

This distributional view is important because the LSTM often has wider tails than the classical benchmarks. Therefore, its advantage should not be assessed only through mean or median performance, but also through the stability and downside of its episode-level outcomes.

LSTM timing edge versus classical benchmarks. The central timing comparison is shown in Figure 3. Using the previously defined $\Delta\text{Sharpe}^{\text{timing}}$, each episode is compared against the stronger of the two classical timing benchmarks, TSMOM and MACD, over the same calendar window. Positive values indicate that the LSTM achieved a higher Sharpe ratio than both classical timing benchmarks; negative values indicate that at least one classical timing benchmark performed better.

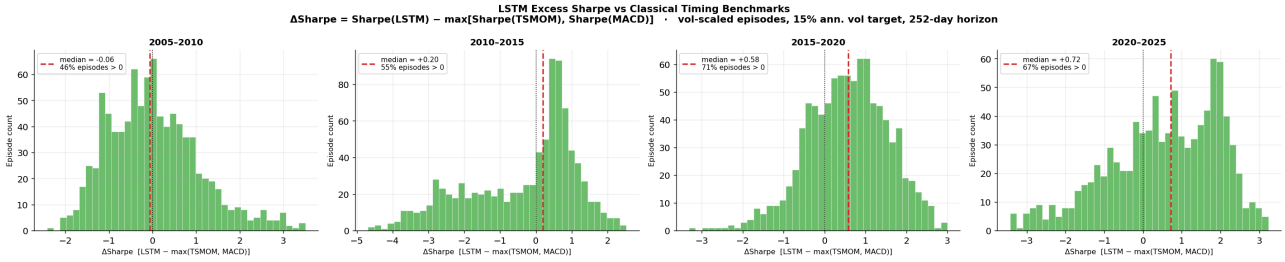


Figure 3: Distribution of $\Delta\text{Sharpe}^{\text{timing}}$ across one-year episodes in each era. Positive values indicate that the LSTM Sharpe ratio exceeds both TSMOM and MACD in the same episode.

Figure 3 shows a clear progression across eras. In 2005–2010, the median timing edge is slightly negative, indicating no robust LSTM improvement over the classical timing rules. In 2010–2015, the median becomes slightly positive, but the distribution still contains a long left tail. In 2015–2020 and 2020–2025, the LSTM shows a stronger timing advantage, with positive median excess Sharpe and a majority of episodes above zero. However, the left tail remains important: even in eras where the LSTM has a positive median edge, there are episodes where it materially underperforms MACD or TSMOM.

Performance summary. Table 1 summarizes the main performance metrics. The table confirms the figure-level conclusions. In 2005–2010, Long Only has the highest median Sharpe and final equity. In 2010–2015, TSMOM is the strongest strategy, with higher median Sharpe and annual return than the LSTM. In 2015–2020, the LSTM is competitive with Long Only and clearly stronger than TSMOM and MACD. In 2020–2025, the LSTM dominates the timing benchmarks and also improves on Long Only in median Sharpe, annual return, drawdown, hit rate and final equity.

Overall, the performance analysis suggests that the LSTM does not provide a universal momentum edge. Its advantage appears conditional. It is weak or absent in the earlier eras, becomes more visible after 2015, and is strongest in 2020–2025. The next part of the analysis therefore focuses on selected tail episodes: breakdown episodes where the LSTM materially underperforms the best benchmark, and outperformance episodes where it adds the most value. The objective is to understand not only where the LSTM fails, but also what distinguishes the episodes where it shines.

4 When Does the LSTM Break or Shine?

The previous section showed that the LSTM does not deliver a uniform improvement over classical momentum rules. Its relative performance varies substantially across regimes and across one-year deployment episodes. We

Era	Strategy	N	Sh. med	Sh. [P5,P95]	Ret. med (%)	DD med (%)	Hit (%)
2005–2010	LSTM	929	1.08	[-0.64,2.48]	16.2	-12.5	83.7
	Long Only	929	1.89	[-1.31,3.09]	31.3	-8.1	83.5
	TSMOM	929	1.02	[-1.14,2.50]	15.2	-13.4	71.7
	MACD	929	0.80	[-0.84,2.40]	11.5	-14.9	72.1
2010–2015	LSTM	932	0.33	[-1.73,2.90]	4.0	-14.3	58.2
	Long Only	932	0.07	[-1.88,2.87]	-0.1	-16.2	49.9
	TSMOM	932	0.96	[-0.90,2.48]	14.1	-11.2	67.1
	MACD	932	0.03	[-2.50,1.69]	-0.6	-15.2	48.1
2015–2020	LSTM	932	0.65	[-1.15,2.42]	8.9	-14.6	63.8
	Long Only	932	0.70	[-1.10,2.08]	9.9	-13.7	72.2
	TSMOM	932	-0.09	[-1.31,1.39]	-2.5	-17.8	45.6
	MACD	932	-0.60	[-1.37,0.57]	-9.6	-16.7	19.2
2020–2025	LSTM	929	1.48	[-0.78,3.28]	23.4	-9.7	88.8
	Long Only	929	1.20	[-0.23,3.07]	18.4	-10.4	86.1
	TSMOM	929	0.77	[-0.53,1.66]	10.9	-12.0	71.9
	MACD	929	0.02	[-0.95,2.09]	-0.9	-12.4	48.7

Table 1: Performance summary across one-year deployment episodes. Each era is generated from 200 sampled starts across five seeds, with duplicate episode keys removed. Metrics are computed from volatility-scaled episode returns. Sh. denotes Sharpe ratio, Ret. annualized return, DD maximum drawdown, and Hit the percentage of positive-return episodes.

therefore move from average performance to tail behaviour: when does the LSTM materially underperform, and when does it add value?

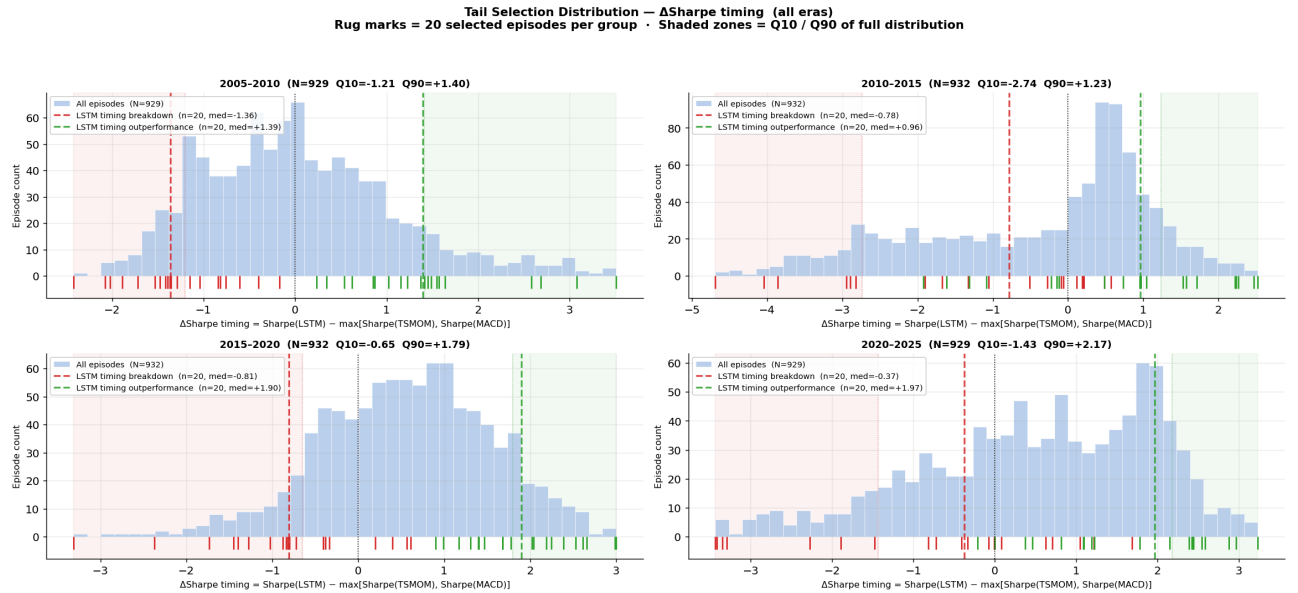


Figure 4: Distribution of $\Delta\text{Sharpe}^{\text{timing}}$ by era, with selected diagnostic episodes marked by rug lines. Histograms show all retained deployment episodes; rug marks show breakdown and outperformance episodes selected using raw excess return relative to the best benchmark, Δ_{bestBM} . Shaded regions mark the lower and upper deciles of the full $\Delta\text{Sharpe}^{\text{timing}}$ distribution within each era.

Figure 4 shows that the distribution of LSTM timing edge is regime-dependent. The location and dispersion of the distribution change materially across eras, and each era contains both positive and negative tails. This motivates a direct investigation of the selected diagnostic episodes: the left tail, where the LSTM breaks relative to classical timing rules, and the right tail, where it shines.

4.1 Visual motivation and diagnostic propositions

Before running formal diagnostics, we first inspect representative tail episodes. The objective is not to draw conclusions from individual examples, but to identify plausible mechanisms that can later be tested systematically.

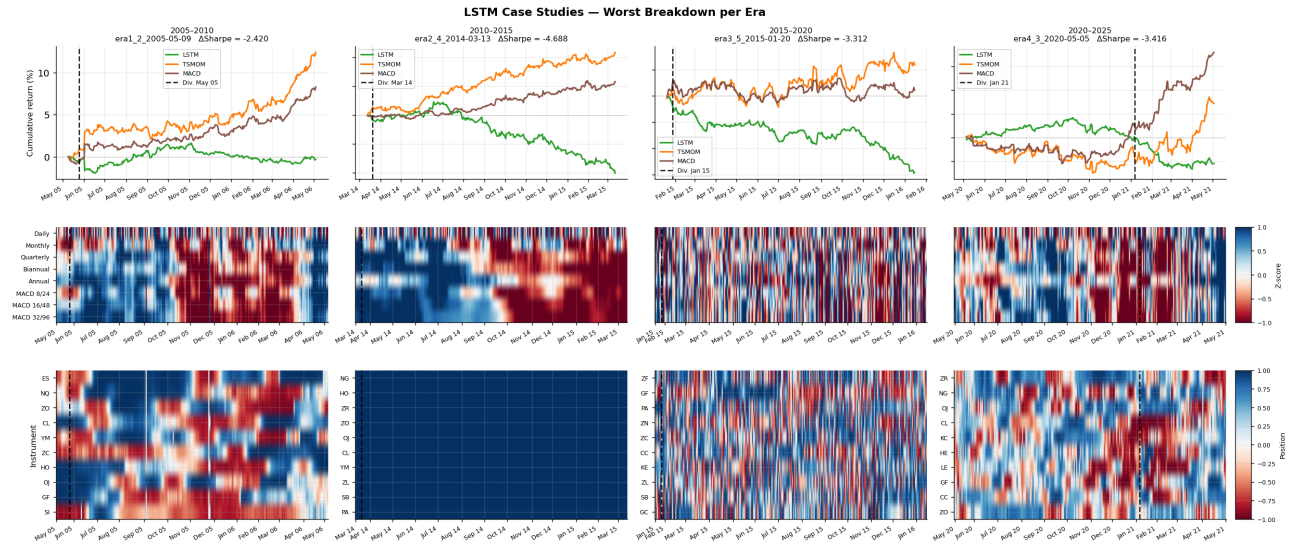


Figure 5: Worst LSTM timing breakdown episode in each era. Each column shows cumulative returns, position-weighted feature signals, and LSTM instrument-level position signals for the worst selected breakdown episode.

Figure 5 shows that LSTM breakdowns do not all have the same visual signature. In some cases, most clearly in 2010–2015, the model saturates into near-uniform extreme positions, effectively losing its ability to discriminate across instruments. In other cases, the position heatmaps show frequent changes in sign and intensity, raising the possibility that unstable switching may also contribute to failure.

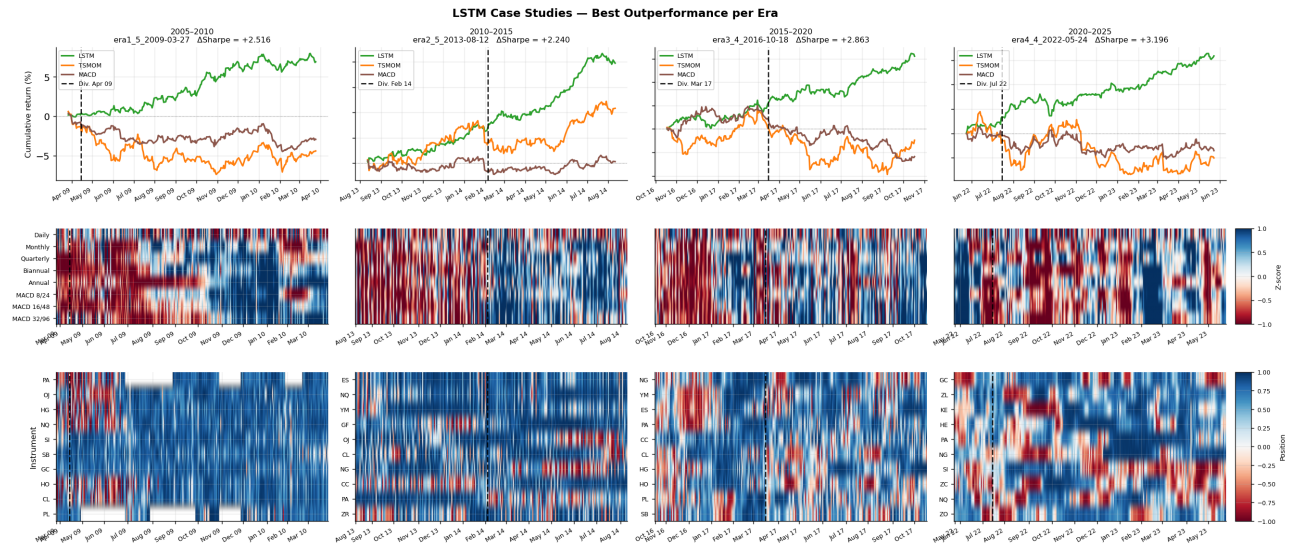


Figure 6: Best LSTM timing outperformance episode in each era. Each column shows cumulative returns, position-weighted feature signals, and LSTM instrument-level position signals for the strongest selected outperformance episode.

Figure 6 provides an important counterpoint. Some of the strongest outperformance episodes also display active switching across instruments and signs. This suggests that position instability may not be intrinsically harmful. Instead, switching may be valuable when it is directionally useful, while failure may occur when the model saturates or switches in the wrong direction.

Taken together, Figures 5 and 6 motivate five diagnostic propositions. These propositions should be read as diagnostic hypotheses: they structure the investigation of the selected tails, but they are not intended as definitive causal claims. Each proposition links a visible pattern from the case studies to an episode-level diagnostic that can be compared across eras. The purpose is to ask whether the same mechanisms appear repeatedly, and whether they differ between breakdown and outperformance episodes.

- H1. Saturation as loss of discrimination.** The LSTM may break when its tanh output collapses toward ± 1 , reducing its ability to discriminate across instruments and time.
- H2. Switching as instability or adaptation.** Frequent position changes may indicate noisy instability, but they may also be the mechanism through which the LSTM adapts successfully.
- H3. Classical-signal disagreement as ambiguity.** Disagreement between MACD and TSMOM may either confuse the LSTM or create an opportunity for nonlinear timing to add value.
- H4. Wrong calls and feature channels.** Severe breakdowns may involve repeated wrong directional calls,

potentially associated with greater reliance on particular return horizons or MACD-type features.

H5. Instrument-level concentration. Some breakdowns may be concentrated in specific wrong-sided contracts rather than reflecting broad market noise.

We next test these propositions using episode-level diagnostics based on LSTM position dynamics, MACD–TSMOM signal conflict, wrong-call rates, feature-attribution summaries, and instrument-level P&L decomposition.

4.2 H1–H2: Saturation versus switching

The first two propositions concern the behaviour of the LSTM output itself. Since the model outputs a bounded position signal $w_{i,t} \in [-1, 1]$, we examine whether breakdowns are associated with loss of discrimination through saturation, or with excessive position switching. The diagnostics used here are position turnover, sign-flip frequency and output saturation, defined formally in Appendix G.

Figure 7 gives the first important result. The evidence reverses the initial concern in **H2**. Episodes with higher turnover and more frequent sign flips tend to have higher, not lower, LSTM timing edge. This suggests that frequent switching is not a generic failure mode. Instead, active position adjustment appears to be part of the mechanism through which the LSTM shines when its directional changes are useful. This result is before transaction costs; therefore, the positive association between switching and timing edge should be interpreted as a gross-performance result.

The evidence for **H1** is more cautious. Output saturation is negatively associated with $\Delta \text{Sharpe}^{\text{timing}}$ in the pooled sample, but the visual relationship is weaker than for turnover and sign-flip frequency. Many observations are clustered near low saturation, while a smaller number of episodes are close to full saturation. Saturation is therefore better interpreted as a warning sign in specific breakdown episodes, rather than as the main general explanation for LSTM failure.

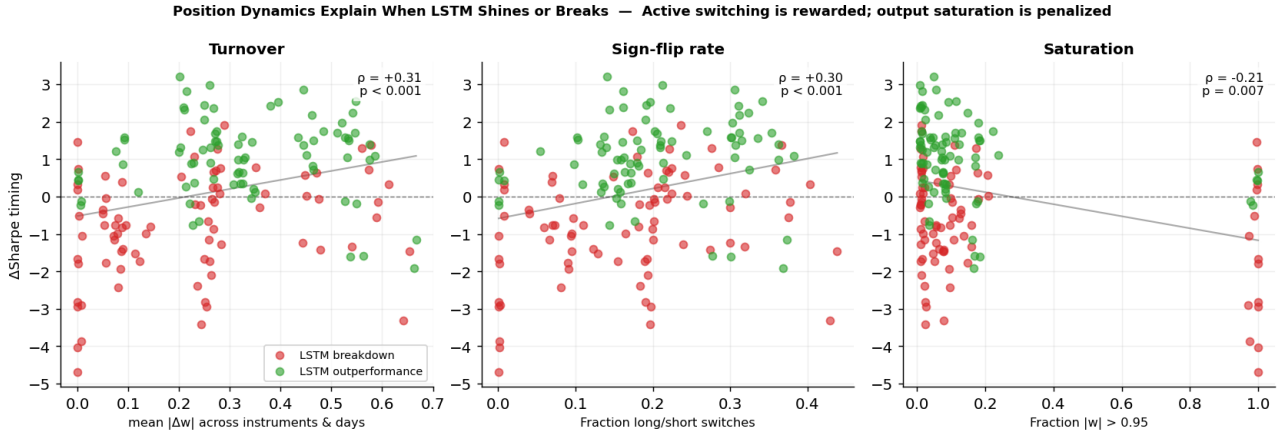


Figure 7: Position dynamics and LSTM timing edge. Turnover and sign-flip frequency are positively associated with $\Delta \text{Sharpe}^{\text{timing}}$, while output saturation is negatively associated with performance.

The evidence for **H1** is more cautious. Output saturation is negatively associated with $\Delta \text{Sharpe}^{\text{timing}}$ in the pooled sample, but the visual relationship is weaker than for turnover and sign-flip frequency. Many observations are clustered near low saturation, while a smaller number of episodes are close to full saturation. Saturation is therefore better interpreted as a warning sign in specific breakdown episodes, rather than as the main general explanation for LSTM failure.

Table 2: Pooled associations between LSTM position dynamics and timing edge.

Diagnostic relation	Spearman ρ	p -value	Interpretation
Turnover TO_e vs ΔS	+0.31	< 0.001	Higher turnover is associated with stronger LSTM timing edge.
Sign-flip rate SF_e vs ΔS	+0.30	< 0.001	More long/short switching is associated with stronger LSTM timing edge.
Output saturation SAT_e vs ΔS	−0.21	0.007	Saturation is negatively associated with timing edge, but the visual relation is weaker.

ΔS denotes $\Delta \text{Sharpe}^{\text{timing}}$. Because deployment episodes overlap in calendar time, the reported p -values should be interpreted as descriptive evidence rather than as fully independent-sample inference.

Table 2 confirms that switching is not a failure signature: both turnover and sign-flip frequency are positively associated with LSTM timing edge. This weakens **H2** as a failure explanation and suggests that active adjustment is part of the LSTM’s adaptive timing mechanism.

To assess whether these relationships are concentrated in a single period, Table 3 reports the same diagnostics by era.

Table 3: By-era robustness of turnover and saturation diagnostics.

Era	Turnover ρ	Turnover p	Saturation ρ	Saturation p
2005–2010	+0.47	0.002	−0.23	0.151
2010–2015	+0.27	0.092	−0.34	0.032
2015–2020	+0.12	0.447	+0.16	0.323
2020–2025	−0.20	0.213	−0.29	0.071
Pooled	+0.31	< 0.001	−0.21	0.007

The by-era results reinforce the regime-dependent nature of the mechanism. Turnover is most clearly rewarded in 2005–2010, while saturation is most clearly penalised in 2010–2015. In later eras, the saturation relation is weaker. Thus, the pooled evidence should be interpreted as a recurring diagnostic pattern, not as a universal law.

4.3 H3: Classical-signal disagreement

The third proposition is that LSTM breakdowns may occur when classical momentum rules disagree. We test this by measuring the fraction of instrument-days where MACD and TSMOM have opposite directional signs. The formal definition of this conflict measure is reported in Appendix G.

Figure 8 changes the interpretation of **H3**. Classical-signal disagreement is not, by itself, a source of LSTM breakdown. The scatter plot shows a positive association between MACD–TSMOM conflict and LSTM timing edge, with Spearman correlation +0.27 and $p = 0.0006$. This suggests that disagreement between classical rules creates an environment where the LSTM has more room to add value, consistent with the idea that nonlinear timing may be useful when fixed rules give mixed signals.

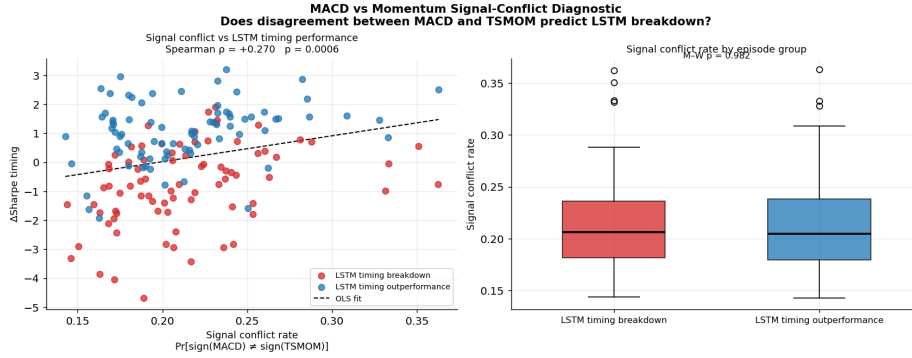


Figure 8: MACD–TSMOM signal conflict and LSTM timing edge. The scatter plot shows that higher disagreement between classical timing signals is positively associated with LSTM relative performance. The boxplot shows that signal conflict alone does not separate breakdown from outperformance episodes.

However, the boxplot gives an important qualification. The distribution of signal conflict is very similar for breakdown and outperformance episodes, with a Mann–Whitney p -value of 0.982. Therefore, classical-signal disagreement is not a classifier of success or failure. It appears in both strong breakdown and strong outperformance episodes. The better interpretation is that disagreement defines an ambiguous regime: the LSTM can shine if its learned position signal resolves the ambiguity correctly, but it can also break if it resolves the ambiguity in the wrong direction.

4.4 H4: Wrong-call diagnostics and feature channels

The fourth proposition is that severe LSTM breakdowns involve repeated wrong directional calls, and that these errors may be associated with greater reliance on particular input features. We therefore separate **H4** into two diagnostics. First, we test whether more severe breakdowns involve more frequent negative captured-P&L decisions. Second, conditional on wrong-call frequency, we examine whether these errors are associated with greater relative reliance on specific momentum horizons or MACD-type features. The wrong-call and feature-attribution measures are defined formally in Appendix G.

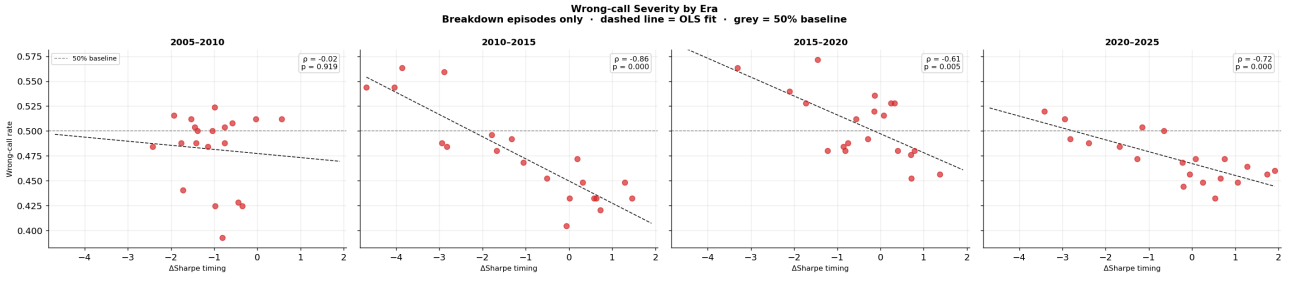


Figure 9: Wrong-call severity by era. Within breakdown episodes, each panel relates the LSTM wrong-call rate to $\Delta\text{Sharpe}^{\text{timing}}$. The dashed line is an OLS fit and the grey horizontal line marks the 50% wrong-call baseline.

Figure 9 provides the strongest evidence for the first part of **H4**. In 2005–2010, there is essentially no relation between wrong-call rate and breakdown severity. In the remaining eras, however, the relation is clearly negative: the more negative the LSTM timing edge, the higher the wrong-call rate. Given the small by-era sample of 20 breakdown episodes, the exact correlations should be interpreted cautiously, but the pattern is economically meaningful and consistent across three of the four eras.

This suggests that severe LSTM failures are not simply caused by a few isolated shocks. In most eras, the worst breakdowns involve repeated directional errors across many instrument-days. Since more severe breakdowns have more negative $\Delta\text{Sharpe}^{\text{timing}}$, the negative relationship indicates that deeper relative underperformance is associated with more frequent wrong calls by the LSTM.

Having established wrong-call frequency as a breakdown diagnostic, we next examine whether wrong calls are associated with greater relative reliance on particular input features.

Figure 10 shows that the feature-channel part of **H4** is exploratory and regime-specific. The dominant features associated with wrong-call frequency differ across eras. In 2010–2015, wrong calls are most positively associated with the short MACD feature $\text{MACD}_{8,24}$. In 2015–2020, the strongest positive association is with the long MACD feature $\text{MACD}_{32,96}$, while biannual momentum is negatively associated with wrong-call frequency. In 2020–2025, quarterly momentum is positively associated with wrong calls, whereas annual momentum is strongly negatively associated. The 2005–2010 panel is less conclusive.

The main conclusion for **H4** is therefore two-layered. The wrong-call diagnostic provides the stronger evidence: severe LSTM breakdowns often involve repeated directional errors, especially after 2010. The feature-channel evidence is weaker and more exploratory: there is no single input feature that explains all failures. Since Integrated Gradients explains the model’s position output rather than causally decomposing realised P&L, these feature results should be interpreted as diagnostic rather than definitive.

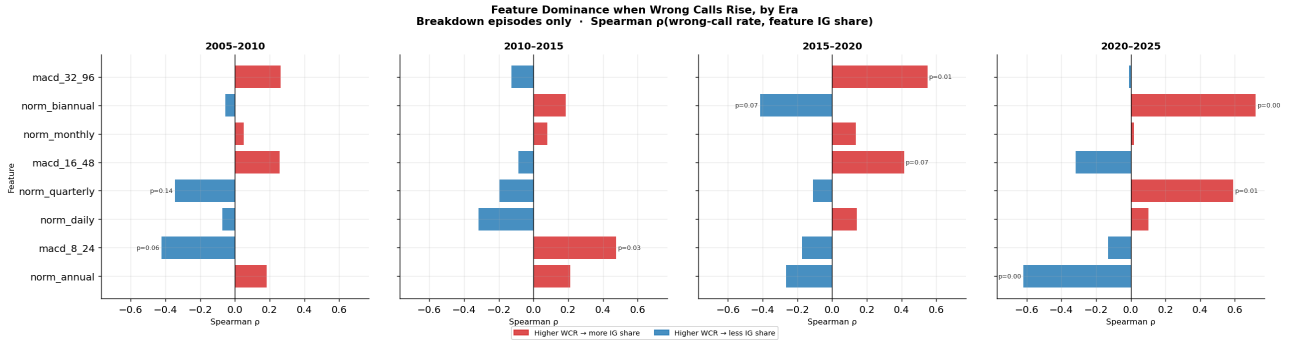


Figure 10: Feature dominance when wrong calls rise, by era. Within breakdown episodes, each panel reports the Spearman correlation between wrong-call rate and feature-level Integrated Gradients attribution share. Positive values indicate that higher wrong-call frequency is associated with greater relative reliance on that feature.

4.5 H5: Instrument-level diagnostics and wrong-sided contracts

The fifth proposition is that severe LSTM breakdowns may be driven by wrong directional bets in particular instruments rather than by broad market noise. This diagnostic asks whether the LSTM loses money diffusely across the portfolio, or whether losses are concentrated in specific contracts where classical momentum rules perform better. The instrument-level P&L decomposition is defined in Appendix G.

Figure 11 provides an illustrative mechanism check for **H5**. In the worst breakdown episodes shown, LSTM losses are concentrated in specific contracts where TSMOM and MACD often generate positive P&L. This indicates that the LSTM is not merely suffering from broad market noise or from an adverse regime common to all strategies. In these cases, the model is directionally wrong on instruments where classical timing rules capture profitable trends.

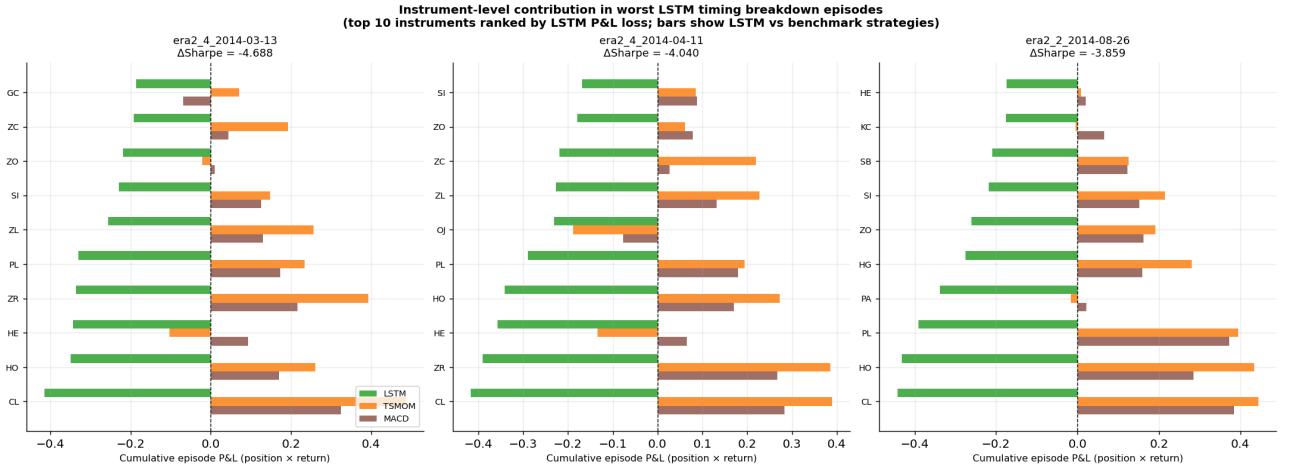


Figure 11: Instrument-level P&L contribution in severe LSTM breakdown episodes. Bars compare cumulative episode contribution from LSTM, TSMOM and MACD for the instruments with the largest LSTM losses.

This evidence is not a formal statistical test because the figure focuses on the worst inspected breakdowns and the instruments with the largest LSTM losses. Nevertheless, it supports the interpretation that some severe LSTM failures are wrong-sided instrument-level timing errors, not only portfolio-level volatility or random adverse shocks.

4.6 Hypothesis-testing summary

Table 4: Diagnostic propositions and empirical interpretation.

Proposition	Evidence from selected tail episodes	Interpretation
H1 Saturation	Saturation is negatively correlated with $\Delta Sharpe^{timing}$, but visual separation is weak and the effect is concentrated in some regimes.	Partial support: saturation can be a warning sign, but not a universal failure mechanism.
H2 Switching	Turnover and sign-flip rate are positively associated with $\Delta Sharpe^{timing}$.	Reinterpreted: switching is more consistent with adaptive timing than with failure.
H3 Classical disagreement	MACD–TSMOM conflict is positively associated with LSTM timing edge, but does not separate breakdown from out-performance groups.	Reinterpreted: disagreement is an opportunity condition, not a failure condition.
H4 Wrong calls / features	Severe breakdowns are associated with higher wrong-call rates in most eras; feature associations vary by era.	Supported for wrong calls; feature channels are exploratory and regime-specific.
H5 Wrong instruments	Worst inspected failures show concentrated losses in particular contracts where classical rules often perform better.	Illustrative support: some failures are wrong-sided instrument-level timing errors.

The most robust result is not that the LSTM fails because it is active. The opposite is closer to the evidence: active switching is positively associated with LSTM timing edge. Turnover and sign-flip frequency are higher in stronger LSTM episodes, suggesting that switching is part of the model’s adaptive timing mechanism when it is directionally useful.

The evidence for saturation is more cautious. Saturation appears in some visually important breakdowns and is negatively associated with LSTM timing edge in the pooled sample, but it does not provide a universal explanation. It is better interpreted as a regime-specific warning sign of lost discrimination.

The evidence also clarifies when the LSTM shines. MACD–TSMOM disagreement is not a breakdown condition; it is an opportunity condition. The LSTM tends to add value when classical rules give conflicting information, provided that its learned position signals remain directionally useful.

The failure diagnostics point toward repeated wrong directional calls rather than excessive activity. In most eras, severe breakdowns are associated with higher wrong-call frequency, and the worst inspected cases show concentrated losses in specific instruments where classical rules often perform better. Overall, the LSTM shines when active switching resolves ambiguous regimes, and it breaks when its learned signals become repeatedly wrong-sided or, in some regimes, lose discrimination through saturation.

5 Conclusion

This project investigated whether an LSTM-based momentum model provides a robust improvement over classical timing rules. The main contribution is empirical decomposition: instead of relying on a single aggregate backtest, we evaluate performance across four eras, five random seeds and many one-year deployment episodes. This exposes both the average behaviour and the dispersion of LSTM performance across realistic deployment windows.

The results show that the LSTM does not deliver a universal momentum edge. Its performance is strongly regime-dependent. It is weak or inconsistent in 2005–2010 and 2010–2015, but becomes more competitive after 2015, especially in 2020–2025, where it frequently outperforms TSMOM and MACD on a risk-adjusted basis. The relevant conclusion is therefore not that the LSTM dominates classical momentum rules, but that its edge is conditional.

The tail analysis clarifies this conditionality. Selected breakdown and outperformance episodes show that activity itself is not the failure mechanism. Turnover and sign-flip frequency are positively associated with LSTM timing edge, suggesting that active adjustment is often part of the model’s value mechanism. The LSTM shines when switching is directionally useful, especially when MACD and TSMOM give conflicting signals.

The failure diagnostics point instead toward wrong-sided decisions and, in some regimes, loss of discrimination. Severe breakdowns are often associated with repeated wrong calls: instrument-days where the LSTM position generates negative captured P&L. Saturation appears in some important breakdowns and is negatively associated with timing edge in the pooled sample, but it is better interpreted as a regime-specific warning sign than as a universal explanation. Feature attribution and instrument-level P&L decomposition provide exploratory evidence that some failures are concentrated in specific signals and contracts, rather than reflecting only broad adverse market conditions.

This project is intentionally limited in scope. Given the time and computational constraints of a short course project, the diagnostic analysis focuses on the 20 worst and 20 best episodes per era. A natural extension would be to test whether the conclusions remain stable when expanding the diagnostic sample to a larger fraction of the distribution, rather than only the most extreme tails. The analysis could also be extended to a larger futures universe, ideally closer to the full universe used in the reference literature, to test whether the observed regime dependence survives broader cross-asset coverage.

Several modelling extensions would also strengthen the research design. First, the feature set could be expanded beyond normalized momentum and MACD signals. Candidate additions include realized volatility, volatility-of-volatility, term-structure variables, carry, cross-asset signals, ARX-type lagged predictors, VIX or other risk-aversion proxies, interest-rate variables and macro-financial regime indicators. Second, the model could be tested explicitly around major market transitions, such as the 2007–2008 financial crisis, the COVID crash and rebound, and the inflationary/rate-hiking regime of 2022. This would help assess whether the LSTM retains useful “memory” after large market shocks, or whether its learned dynamics become stale after structural breaks.

A further promising direction is to make the regime dependence explicit. The results suggest that the LSTM performs better in some environments than others, but the current project treats regime dependence mostly *ex post*. Future work could build a market-regime indicator using variables such as interest rates, yield-curve slope, realized volatility, VIX, trend dispersion, cross-asset correlations or macroeconomic state variables. Such an indicator could be used either to condition the LSTM’s signals, to decide when to trust the LSTM versus classical rules, or to build an ensemble strategy that switches between LSTM, TSMOM, MACD and Long Only depending on market conditions.

Finally, future work should incorporate transaction costs explicitly, since active switching may be costly in practice. The model could also be trained with penalties against persistent saturation near ± 1 , extended with a true portfolio-allocation layer, or compared with Transformer-based architectures to test whether attention mechanisms improve robustness across regimes.

Overall, the LSTM is not a universally superior replacement for classical momentum rules. Its advantage is conditional, regime-dependent and sensitive to the quality of its position dynamics. It shines when active switching adds useful nonlinear timing information, and breaks when active decisions become repeatedly wrong-sided or, in some regimes, when its signals saturate and lose discrimination. The main research implication is that deep momentum models should not be evaluated only by average backtest performance: they should also be decomposed by regime, by tail behaviour and by the mechanisms through which they add or destroy value.

Appendices

A Data Universe

The empirical analysis uses a fixed 26-instrument futures universe collected from Refinitiv. The starting point was the 50-instrument futures universe reported in Wood et al. [4]. In practice, only 40 instruments could be collected reliably from Refinitiv, and a final subset of 26 instruments was selected for the empirical analysis.

The final universe preserves broad cross-asset exposure across equity indices, fixed income, energy, metals, grains, soft commodities and livestock. The same 26 instruments are used consistently across all four eras, avoiding changes in the cross-sectional universe across test periods.

Identifier	Description	From	Identifier	Description	From
CC	Cocoa	Jan 1991	OJ	Orange Juice	Jan 1991
CL	Crude Oil	Jan 1991	PA	Palladium	Jan 1991
ES	S&P 500	Sep 1998	PL	Platinum	Jan 1991
GF	Feeder Cattle	Jan 1991	SB	Sugar	Jan 1991
GC	Gold	Jan 1991	SI	Silver	Jan 1991
HE	Lean Hogs	Jan 1991	YM	Dow Jones	Apr 2003
HG	Copper	Jan 1991	ZB	30Y T-Bond	Jan 1991
HO	Heating Oil	Jan 1991	ZC	Corn	Jan 1991
KC	Coffee	Jan 1991	ZF	5Y T-Note	Jan 1991
KE	Wheat	Jan 1991	ZL	Soybean Oil	Jan 1991
LE	Live Cattle	Jan 1991	ZN	10Y T-Note	Jan 1991
NG	Natural Gas	Apr 1991	ZO	Oats	Jan 1991
NQ	Nasdaq	Jun 2000	ZR	Rough Rice	Jan 1991

Table 5: Final 26-instrument Refinitiv futures universe used in the empirical analysis. The universe is a diversified subset of the futures universe reported in Wood et al. [4], selected subject to Refinitiv availability and computational feasibility.

B LSTM Architecture and Training

The neural model used in this project is the LSTM component implemented in the public *Trading with the Momentum Transformer* codebase [4]. Although the repository also contains Transformer-based architectures, this project deliberately restricts the deep learning model to the LSTM and compares it against Long Only, TSMOM and MACD benchmarks.

For each instrument i and decision date t , the model receives a rolling input window

$$X_{i,t} = (x_{i,t-L+1}, \dots, x_{i,t}) \in \mathbb{R}^{L \times F},$$

where $L = 63$ is the lookback length and $F = 8$ is the number of input features. Each vector $x_{i,\tau}$ contains normalized momentum variables: daily, monthly, quarterly, biannual and annual returns, together with three MACD-style signals.

The architecture is

$$X_{i,t} \longrightarrow \text{LSTM} \longrightarrow \text{Dropout} \longrightarrow \text{TimeDistributed}(\text{Dense} + \tanh) \longrightarrow w_{i,t} \in [-1, 1].$$

The LSTM layer is configured with `return_sequences=True`, so it produces a hidden state at each step of the 63-day sequence. The time-distributed dense layer maps each hidden state into a scalar position signal. In the trading implementation, the final time-step output is used as the position for date t :

$$w_{i,t} = \tanh(z_{i,t}), \quad w_{i,t} \in [-1, 1].$$

Thus, $w_{i,t} = 1$ corresponds to a full long position, $w_{i,t} = -1$ to a full short position, and $w_{i,t} = 0$ to no exposure.

At each internal time step $\ell = 1, \dots, L$, the LSTM updates its memory through the standard gate equations:

$$\begin{aligned} f_\ell &= \sigma(W_f x_\ell + U_f h_{\ell-1} + b_f), \\ i_\ell &= \sigma(W_i x_\ell + U_i h_{\ell-1} + b_i), \\ \tilde{c}_\ell &= \tanh(W_c x_\ell + U_c h_{\ell-1} + b_c), \\ c_\ell &= f_\ell \odot c_{\ell-1} + i_\ell \odot \tilde{c}_\ell, \\ o_\ell &= \sigma(W_o x_\ell + U_o h_{\ell-1} + b_o), \\ h_\ell &= o_\ell \odot \tanh(c_\ell). \end{aligned}$$

Here f_ℓ is the forget gate, i_ℓ is the input gate, $\tilde{c}_\ell$ is the candidate memory update, c_ℓ is the cell state, o_ℓ is the output gate, and h_ℓ is the hidden representation passed to the output layer. The Keras implementation uses sigmoid recurrent gates and tanh state transformations.

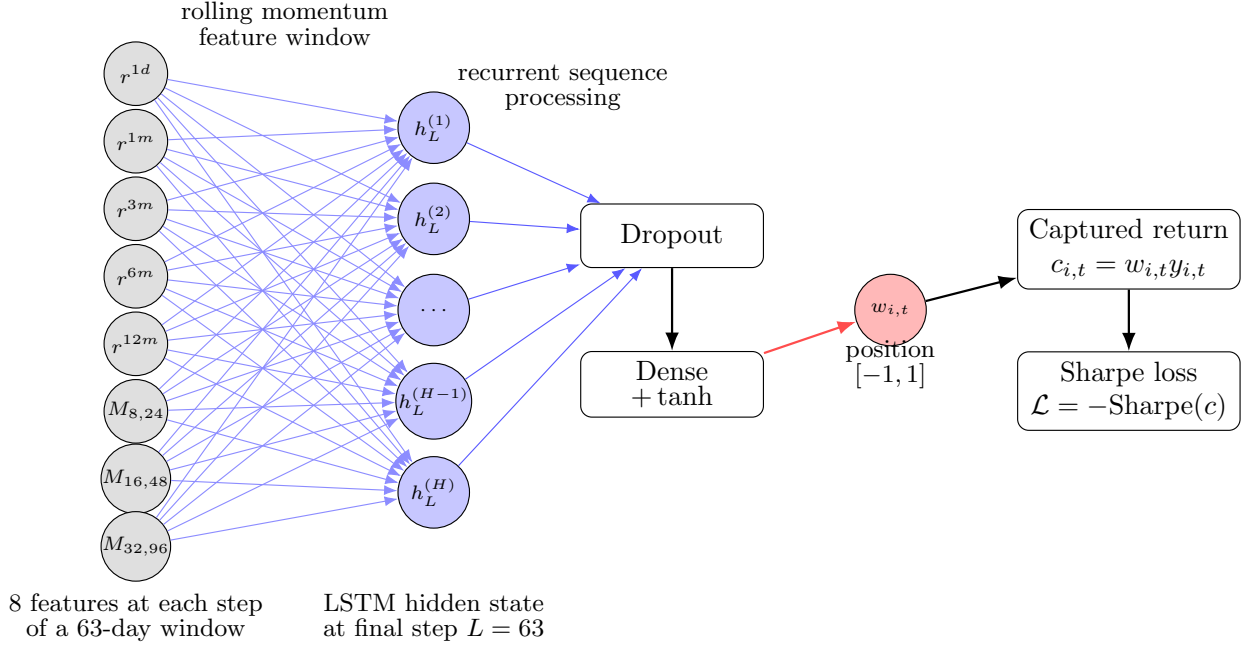


Figure 12: Schematic representation of the LSTM model used in the project. The input is a 63-day sequence of 8 normalized momentum features. The LSTM processes the sequence and its final hidden state is passed through dropout and a dense layer with tanh activation to produce a bounded trading position $w_{i,t} \in [-1, 1]$. The position is multiplied by the next-day target return to obtain captured return, and the model is trained using a Sharpe-based loss.

The model differs from a traditional forecasting LSTM in both its output and its objective. A standard forecasting LSTM would typically predict a next-period return and minimize a pointwise loss such as mean squared error. In contrast, the model used here outputs a trading position directly. For each instrument i and decision date t , the network produces a bounded exposure

$$w_{i,t} \in [-1, 1].$$

This position is then multiplied by the next-day volatility-scaled target return $y_{i,t}$, producing the captured return

$$c_{i,t} = w_{i,t} y_{i,t}.$$

Thus, the model is not rewarded for predicting returns accurately in the usual forecasting sense. It is rewarded for choosing positions that generate positive realised captured returns.

The training loss is computed from the distribution of captured returns in each mini-batch. Let $\mathcal{B}$ denote a mini-batch of instrument-date observations. For each observation $(i, t) \in \mathcal{B}$, the model computes

$$c_{i,t} = w_{i,t} y_{i,t}.$$

The batch mean captured return is

$$\bar{c}_{\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{(i,t) \in \mathcal{B}} c_{i,t},$$

and the corresponding batch second moment is

$$\overline{c^2}_{\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{(i,t) \in \mathcal{B}} c_{i,t}^2.$$

The batch standard deviation is therefore computed as

$$\sigma_{\mathcal{B}}(c) = \sqrt{\overline{c^2}_{\mathcal{B}} - \bar{c}_{\mathcal{B}}^2 + \varepsilon}, \quad \varepsilon = 10^{-9}.$$

The loss is the negative annualized Sharpe ratio of captured returns:

$$\mathcal{L}_{\text{Sharpe}} = -\frac{\bar{c}_{\mathcal{B}}}{\sigma_{\mathcal{B}}(c)} \sqrt{252}.$$

Minimizing this loss is equivalent to maximizing the batch Sharpe ratio of the positions generated by the network. The annualization factor $\sqrt{252}$ is constant and therefore does not affect the optimizer's ranking of parameter

updates, but it keeps the objective on the same scale as the reported annualized Sharpe ratios.

This objective aligns training with the trading problem. The model is not penalized directly for forecasting error; instead, it is penalized when its positions produce low average captured returns relative to their variability. In practical terms, the network is encouraged to output positive exposure before positive target returns, negative exposure before negative target returns, and smaller exposure when the signal is weak or uncertain.

Hyperparameters are selected by random search over hidden layer size, dropout rate, learning rate, gradient clipping norm and mini-batch size. The optimizer is Adam with gradient clipping, which stabilizes training under the Sharpe-based objective. Model selection is based on validation portfolio Sharpe: at the end of each epoch, validation positions are multiplied by validation returns, aggregated by date across instruments, and then converted into a diversified validation Sharpe ratio. The selected checkpoint is the model with the highest validation Sharpe after an initial burn-in period.

C Backtesting Pipeline and Hyperparameter Optimisation

This appendix summarises the implementation pipeline used to train, select and evaluate the LSTM models across eras.

C.1 Data Universe and Feature Engineering

The experiment uses a diversified universe of 26 futures contracts sourced from Refinitiv, covering equity indices, fixed income, energy, metals, grains, soft commodities and livestock. Raw price series are converted into daily percentage returns and then transformed into volatility-scaled target returns, as described in the main text.

For each instrument and date, the LSTM receives a rolling 63-day sequence of eight normalized momentum features. These are:

$$[r^{1d}, r^{1m}, r^{3m}, r^{6m}, r^{12m}, MACD_{8,24}, MACD_{16,48}, MACD_{32,96}].$$

The first five variables capture return momentum at increasing horizons, while the three MACD variables capture trend-following signals at progressively slower timescales. The annual return feature r^{12m} corresponds to the classical TSMOM signal used in the benchmark strategy.

C.2 Era Structure and Episode Bootstrap

The evaluation horizon is partitioned into four non-overlapping eras:

$$2005\text{--}2010, \quad 2010\text{--}2015, \quad 2015\text{--}2020, \quad 2020\text{--}2025.$$

Each era represents a different market regime. This split is important because the objective is not only to estimate average LSTM performance, but also to examine whether the model behaves differently across market environments.

Within each era, one-year deployment episodes are generated by randomly sampling start dates from the corresponding out-of-sample window. Each episode contains $T = 252$ trading days. For each era and random seed, 200 starts are sampled, with duplicate episode keys removed before analysis. This produces approximately 930 unique one-year episodes per era and 3,722 episodes in total across the full experiment.

C.3 Training, Validation and Hyperparameter Selection

For each era, the LSTM is trained only on data available before the beginning of that era. A validation split inside the pre-test historical period is used for hyperparameter selection, early stopping and checkpoint selection. The test era itself is kept strictly out-of-sample: no observations from the test era are used to update model weights, select hyperparameters or choose checkpoints.

Hyperparameters are selected independently for each era through a structured search over five dimensions:

$$h, \quad lr, \quad do, \quad bs, \quad gnorm,$$

where h is the LSTM hidden size, lr the learning rate, do the dropout rate, bs the mini-batch size, and $gnorm$ the gradient clipping threshold.

The selected configurations are:

Table 6: Selected LSTM hyperparameters by era.

Era	Hidden size h	Learning rate lr	Dropout do	Batch size bs	Gradient clip $gnorm$
2005–2010	160	0.001	0.50	128	1.00
2010–2015	200	0.010	0.40	128	1.00
2015–2020	200	0.001	0.50	256	0.10
2020–2025	160	0.003	0.20	128	0.01

The variation in selected hyperparameters across eras is itself informative. It suggests that the optimisation

problem and the useful degree of regularisation differ across market regimes, consistent with the non-stationary nature of financial time series.

C.4 Training Objective and Checkpoint Selection

The LSTM is trained to output a trading position directly rather than a return forecast. Given the next-day volatility-scaled target return $\tilde{r}_{i,t+1}$, the captured return is

$$c_{i,t+1} = w_{i,t}^{LSTM} \tilde{r}_{i,t+1}.$$

The training loss is the negative annualized Sharpe ratio of captured returns computed over each mini-batch:

$$\mathcal{L}_{Sharpe} = -\frac{\bar{c}}{\sqrt{c^2 - \bar{c}^2 + \varepsilon}} \sqrt{252}.$$

This objective aligns training with the trading problem: the model is rewarded for producing positions that generate positive captured returns relative to their variability.

Because the mini-batch Sharpe loss is noisy, a burn-in period of 15 epochs is imposed. No checkpoint is eligible before the end of this burn-in period. After burn-in, the training loop monitors the smoothed validation Sharpe ratio. The best checkpoint is defined as the epoch with the highest smoothed validation Sharpe after burn-in. This checkpoint is then saved and used for all subsequent out-of-sample inference in the corresponding era.

C.5 Learning Curves

Figure 13 reports representative learning curves for seed 1 across the four eras. The left axis shows the mini-batch training loss, while the right axis shows the annualized validation Sharpe ratio. The shaded grey region marks the burn-in period, and the vertical dashed line indicates the selected checkpoint.

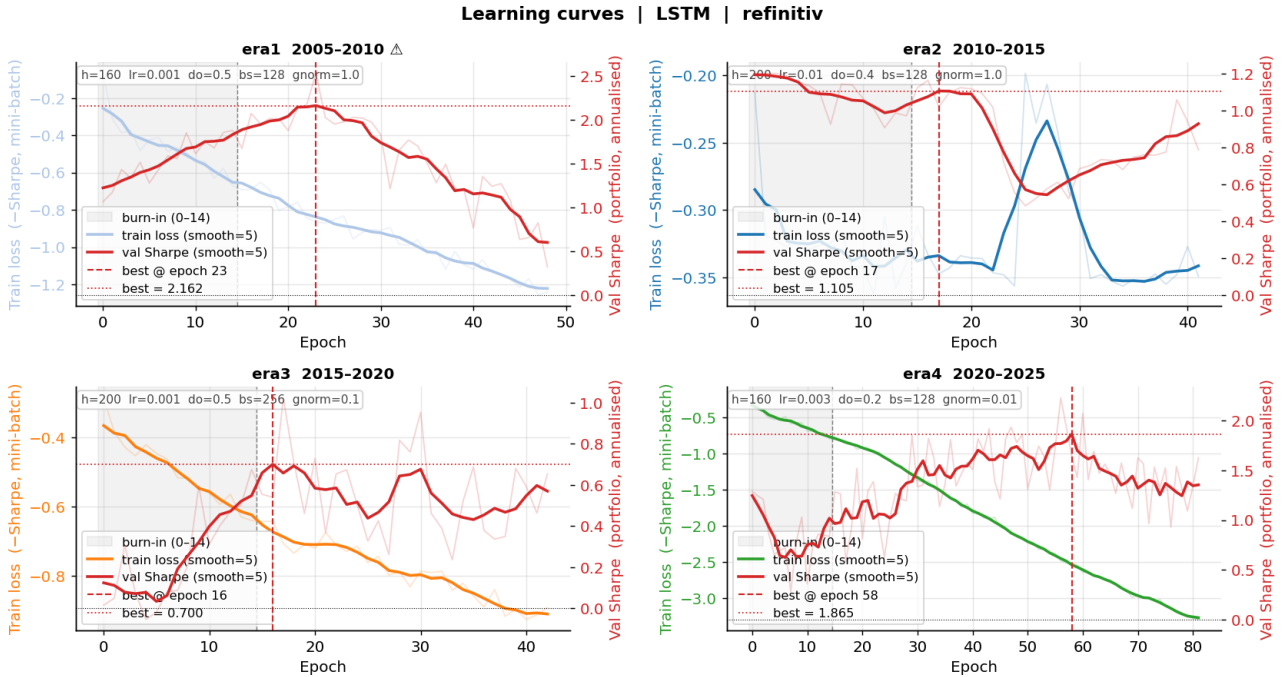


Figure 13: LSTM learning curves by era for seed 1. The left axis reports mini-batch training loss, while the right axis reports annualized validation Sharpe. The grey region marks the burn-in period and the vertical dashed line marks the selected checkpoint.

The learning curves show that optimisation behaviour differs substantially across eras. Era 1 and era 4 have longer training runs, while eras 2 and 3 reach their best validation checkpoints earlier. Era 2 converges quickly but achieves a lower peak validation Sharpe, consistent with the more difficult 2010–2015 regime observed in the out-of-sample performance analysis. Era 4 reaches its best checkpoint later, suggesting that the model continues to refine its position function over a longer training horizon in the post-2020 regime.

These curves also illustrate why checkpoint selection is necessary. The training loss can continue to improve even after validation Sharpe stops improving, so the final epoch is not necessarily the best model for out-of-sample inference. The selected checkpoint therefore reflects validation performance after burn-in, not simply the last epoch of training.

D Benchmark Strategies

This appendix defines the benchmark strategies used in the empirical comparison. All strategies produce an instrument-level signed exposure $w_{i,t}^{(k)}$, which is later applied to the volatility-scaled target return $\tilde{r}_{i,t+1}$.

D.1 Long Only

The Long Only benchmark keeps a constant long exposure in every instrument:

$$w_{i,t}^{LO} = 1.$$

It is not a timing strategy, but it provides a useful reference for periods where persistent long exposure performs well.

D.2 Classical Time-Series Momentum

The main classical timing benchmark is annual time-series momentum. More generally, an intermediate trend strategy can be written as a convex combination of monthly and annual momentum signals:

$$w_{i,t}^{TSMOM} = \omega \operatorname{sign}\left(r_{i,t}^{(21)}\right) + (1 - \omega) \operatorname{sign}\left(r_{i,t}^{(252)}\right).$$

The case $\omega = 0$ corresponds to the annual time-series momentum rule used in the main empirical analysis:

$$w_{i,t}^{TSMOM} = \operatorname{sign}\left(r_{i,t}^{(252)}\right).$$

Thus, the benchmark is long when the past annual return is positive and short when it is negative.

D.3 MACD

For MACD, the raw signal for short and long exponential moving-average timescales (S, L) is

$$\text{MACD}_{i,t}(S, L) = \text{EWMA}_S(P_{i,t}) - \text{EWMA}_L(P_{i,t}).$$

The implementation first normalizes this signal by a rolling price standard deviation:

$$q_{i,t}(S, L) = \frac{\text{MACD}_{i,t}(S, L)}{\text{RollStd}_{63}(P_{i,t})},$$

and then by a rolling signal standard deviation:

$$M_{i,t}(S, L) = \frac{q_{i,t}(S, L)}{\text{RollStd}_{252}(q_{i,t}(S, L))}.$$

The three normalized MACD signals are averaged:

$$m_{i,t} = \frac{1}{3} [M_{i,t}(8, 24) + M_{i,t}(16, 48) + M_{i,t}(32, 96)].$$

The final MACD position is

$$w_{i,t}^{MACD} = \frac{m_{i,t} \exp(-m_{i,t}^2/4)}{0.89}.$$

This produces a smooth signed position rather than a hard long/short rule. The nonlinear transformation dampens very large MACD signals, preventing the position from growing mechanically with signal magnitude.

E Portfolio Construction and Performance Metrics

This appendix gives the return-construction, allocation and performance-metric formulas used in the empirical analysis. The key point is that all strategies are evaluated through equal-weighted, volatility-scaled instrument sleeves. The LSTM controls only the signed exposure inside each sleeve; it does not allocate capital across instruments.

E.1 Volatility-Scaled Target Returns

Let $r_{i,t+1}$ denote the next-day percentage return of instrument i , and let $\sigma_{i,t}^{ann}$ denote its estimated annualized volatility at decision date t . The target return is

$$\tilde{r}_{i,t+1} = r_{i,t+1} \frac{\sigma_{\text{target}}}{\sigma_{i,t}^{ann}}, \quad \sigma_{\text{target}} = 15\%.$$

This expresses each instrument return in risk-normalized units. Price levels do not enter the strategy directly; only percentage returns and volatility-scaled momentum features are used.

E.2 Instrument-Level Captured Returns

For a strategy k , the instrument-level captured return is

$$c_{i,t+1}^{(k)} = w_{i,t}^{(k)} \tilde{r}_{i,t+1},$$

where $w_{i,t}^{(k)}$ is the signed exposure generated by the strategy. For the LSTM, this is the bounded output $w_{i,t}^{LSTM} \in [-1, 1]$. For the benchmarks, it is the corresponding Long Only, TSMOM or MACD exposure.

E.3 Portfolio Aggregation

Instrument-level captured returns are aggregated into an equal-weight cross-sectional portfolio:

$$R_{t+1}^{(k)} = \frac{1}{N_t} \sum_{i=1}^{N_t} c_{i,t+1}^{(k)} = \frac{1}{N_t} \sum_{i=1}^{N_t} w_{i,t}^{(k)} \tilde{r}_{i,t+1}.$$

In this project $N_t = 26$ throughout the test sample, but the notation allows for the general case where the number of available instruments varies over time.

E.4 Instrument Allocation Interpretation

Starting from portfolio value V_t , each instrument receives a base sleeve:

$$\frac{V_t}{N_t}.$$

The strategy position $w_{i,t}^{(k)}$ scales only the signed exposure inside that sleeve:

$$\text{effective exposure}_{i,t}^{(k)} = \frac{V_t}{N_t} w_{i,t}^{(k)}.$$

After the daily return is realized, the portfolio value evolves as

$$V_{t+1} = V_t \left(1 + R_{t+1}^{(k)} \right).$$

At the next decision date, base sleeves are recomputed from the updated portfolio value. Thus, the strategy behaves like a daily rebalanced equal-sleeve portfolio with learned or rule-based signed exposure inside each sleeve.

This distinction is important for interpreting the LSTM heatmaps. A position of $w_{i,t}^{LSTM} = 1$ means full long exposure to that instrument's equal risk-normalized sleeve. It does not mean that the model allocates 100% of the portfolio to that instrument. Similarly, $w_{i,t}^{LSTM} = -1$ means full short exposure to that sleeve, and $w_{i,t}^{LSTM} = 0$ means no exposure inside that sleeve.

E.5 Episode-Level Performance Metrics

For each one-year episode e and strategy k , we compute total return, annualized return, annualized volatility, Sharpe ratio, Sortino ratio, maximum drawdown, Calmar ratio, VaR, CVaR and hit rate.

The main risk-adjusted metric is the annualized Sharpe ratio:

$$\text{Sharpe}_e^{(k)} = \frac{\bar{R}_e^{(k)}}{\sigma_e^{(k)}} \sqrt{252},$$

where $\bar{R}_e^{(k)}$ is the mean daily portfolio return over the episode and $\sigma_e^{(k)}$ is the corresponding daily return standard deviation:

$$\bar{R}_e^{(k)} = \frac{1}{T} \sum_{t=1}^T R_{e,t}^{(k)}, \quad \sigma_e^{(k)} = \sqrt{\frac{1}{T-1} \sum_{t=1}^T \left(R_{e,t}^{(k)} - \bar{R}_e^{(k)} \right)^2}.$$

The factor $\sqrt{252}$ annualizes the daily Sharpe ratio under the standard approximation of 252 trading days per year.

E.6 Ex-Post Volatility Scaling

For the main risk-adjusted comparisons, each episode’s realised portfolio return series is also rescaled to a common 15% annualized volatility target:

$$\hat{R}_{e,t}^{(k)} = R_{e,t}^{(k)} \frac{\sigma_{\text{target}}}{\hat{\sigma}_{e,\text{ann}}^{(k)}}, \quad \hat{\sigma}_{e,\text{ann}}^{(k)} = \sigma_e^{(k)} \sqrt{252}, \quad \sigma_{\text{target}} = 0.15.$$

Here, $R_{e,t}^{(k)}$ is the raw daily portfolio return and $\hat{R}_{e,t}^{(k)}$ is the volatility-scaled return.

Scaled metrics are used for Sharpe-based comparisons and for the main timing-edge analysis. Raw returns are retained as economic diagnostics because they show the realised profit and loss of each strategy before ex-post volatility rescaling. Thus, raw returns answer whether the strategy made money in economic terms, while scaled returns answer whether it generated superior performance per unit of risk.

This episode-level rescaling is used only for ex-post comparability of realised risk-adjusted performance. It is not intended to represent an implementable ex-ante volatility-targeting rule.

F Tail Episode Selection and Diagnostic Framework

This appendix reports the exact definitions used to select the diagnostic tail episodes and to interpret the tail-analysis results.

F.1 Risk-Adjusted Timing Edge

To evaluate LSTM timing skill in risk-adjusted terms, we compare the LSTM against the stronger classical timing benchmark in the same episode:

$$\Delta \text{Sharpe}_e^{\text{timing}} = \text{Sharpe}_e(\text{LSTM}) - \max\{\text{Sharpe}_e(\text{TSMOM}), \text{Sharpe}_e(\text{MACD})\}.$$

A positive value means that the LSTM achieved a higher Sharpe ratio than both classical timing rules in the same one-year episode. A negative value means that at least one classical timing rule performed better.

This measure is used as the main risk-adjusted comparison throughout the empirical analysis.

F.2 Raw Benchmark-Relative Tail Selection

For the diagnostic tail analysis, episodes are selected using a raw benchmark-relative return measure:

$$\Delta_{\text{bestBM},e}^{\text{raw}} = \text{Ret}_e^{\text{raw}}(\text{LSTM}) - \max\{\text{Ret}_e^{\text{raw}}(\text{LongOnly}), \text{Ret}_e^{\text{raw}}(\text{TSMOM}), \text{Ret}_e^{\text{raw}}(\text{MACD})\}.$$

This measures the LSTM’s raw total-return advantage or disadvantage relative to the best available benchmark in the same calendar window.

The use of raw returns for selection is deliberate. It identifies economically material episodes where the LSTM most clearly adds or destroys value relative to feasible alternatives. The subsequent diagnostic analysis still reports risk-adjusted quantities, especially $\Delta \text{Sharpe}_e^{\text{timing}}$, so tail selection and performance interpretation are separated.

F.3 Diagnostic Sample Construction

The full performance dataset contains 3,722 unique one-year episodes across four eras. From these, we select 160 diagnostic episodes in total: 40 per era, consisting of 20 breakdown episodes and 20 outperformance episodes.

Breakdown episodes are the lowest-ranked episodes by Δ_{bestBM} within each era, while outperformance episodes are the highest-ranked episodes by the same metric.

The selection is performed separately within each era. This is important because the underlying market regime changes across the sample, and pooling all tails could allow one regime to dominate the diagnostic analysis. By selecting the same number of breakdown and outperformance episodes in each era, the diagnostic sample preserves balanced regime coverage and allows us to compare whether the LSTM fails or succeeds through similar mechanisms across different market environments.

The choice of 20 episodes per group and era reflects a practical compromise. The goal is to study the most informative regions of the distribution: the cases where the LSTM breaks most clearly and the cases where it adds the most value. At the same time, the Integrated Gradients pipeline is computationally expensive, since each selected episode requires attributions over all instruments, trading days, lookback steps and input features.

The selected sample is therefore large enough to compare diagnostic patterns across regimes, but it should still be interpreted cautiously. The resulting propositions are exploratory mechanisms, not universal statistical laws.

F.4 Diagnostic Dimensions

The selected episodes are analysed along five dimensions:

1. LSTM position dynamics: turnover, sign-flip rate and output saturation;
2. MACD–TSMOM signal conflict;

3. LSTM wrong-call frequency;
4. feature-level Integrated Gradients attribution shares;
5. instrument-level P&L contribution.

The formulas for these diagnostic measures are reported in Appendix G.

G Diagnostic Measures Used in the Tail Analysis

This appendix defines the episode-level diagnostics used in the tail analysis. For each selected episode e , let $\mathcal{T}_e$ denote its set of trading days, with length $T = 252$, and let $N = 26$ denote the number of instruments. The LSTM position for instrument i on day t is denoted by $w_{i,t}$.

G.1 Position Turnover

Position turnover measures the average absolute change in the LSTM signal:

$$TO_e = \frac{1}{N(T-1)} \sum_{i=1}^N \sum_{t=2}^T |w_{i,t} - w_{i,t-1}|.$$

This diagnostic is used to test whether frequent position adjustment is associated with LSTM underperformance or outperformance.

G.2 Sign-Flip Rate

The sign-flip rate measures how frequently the model switches from long to short, or from short to long:

$$SF_e = \frac{1}{N(T-1)} \sum_{i=1}^N \sum_{t=2}^T \mathbf{1}[\text{sign}(w_{i,t}) \neq \text{sign}(w_{i,t-1})].$$

This is a directional switching diagnostic. It is stricter than turnover because small changes in position size do not count unless they change the sign of the exposure.

G.3 Output Saturation

Output saturation measures the fraction of LSTM positions close to the tanh boundaries:

$$SAT_e = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \mathbf{1}(|w_{i,t}| > 0.95).$$

High saturation indicates that the LSTM frequently collapses toward extreme positions, which may reduce its ability to discriminate across instruments and time.

G.4 Cross-Sectional Dispersion

Cross-sectional dispersion measures how differentiated the LSTM positions are across instruments on each day:

$$DISP_e = \frac{1}{T} \sum_{t=1}^T \text{std}_i(w_{i,t}).$$

This was used as a control diagnostic. In the empirical results it did not show a meaningful relationship with LSTM timing edge.

G.5 MACD–TSMOM Signal Conflict

Classical signal conflict measures the fraction of instrument-days where MACD and TSMOM have opposite directional signs:

$$Conflict_e = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \mathbf{1}[\text{sign}(w_{i,t}^{MACD}) \neq \text{sign}(w_{i,t}^{TSMOM})].$$

This diagnostic is used to test whether disagreement between classical timing rules is associated with LSTM breakdown or outperformance.

G.6 Wrong-Call Frequency

For the LSTM, the instrument-level captured return is

$$c_{i,t+1}^{LSTM} = w_{i,t}^{LSTM} \tilde{r}_{i,t+1}.$$

An instrument-day is classified as a wrong call when the captured return is negative:

$$\mathbf{1} [c_{i,t+1}^{LSTM} < 0].$$

The episode-level wrong-call frequency is

$$WC_e = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \mathbf{1} [w_{i,t}^{LSTM} \tilde{r}_{i,t+1} < 0].$$

This diagnostic measures whether breakdowns are driven by repeated directional errors rather than isolated large shocks. It is an absolute LSTM diagnostic, not a benchmark-relative measure.

G.7 Feature Attribution Shares

Integrated Gradients is used to analyse which input features push the LSTM toward a given position. Let

$$IG_{i,t,\tau,j}$$

denote the attribution for instrument i , decision date t , lookback step τ , and feature j . Aggregating over the lookback dimension gives

$$IG_{i,t,j} = \sum_{\tau=1}^L IG_{i,t,\tau,j}.$$

Thus, $IG_{i,t,j}$ measures the total contribution of feature j , across the full input window, to the LSTM position.

For each episode e , feature-level attribution strength is summarized using absolute Integrated Gradients:

$$S_{e,j} = \sum_{t \in \mathcal{T}_e} \sum_{i=1}^N |IG_{i,t,j}|.$$

The attribution share of feature j is

$$Share_{e,j} = \frac{S_{e,j}}{\sum_{k=1}^F S_{e,k}}.$$

These shares are used to examine whether wrong-call frequency is associated with greater relative dependence on particular input horizons or MACD features.

For each feature j , the diagnostic null hypothesis is:

$$H_0 : \rho_s(WC_e, Share_{e,j}) = 0,$$

where ρ_s denotes the Spearman rank correlation. Since Integrated Gradients explains the model output rather than realised returns, this evidence is interpreted as suggestive rather than causal.

G.8 Instrument-Level P&L Contribution

For each strategy k , instrument i , and episode e , cumulative instrument-level P&L contribution is computed as

$$C_{e,i}^{(k)} = \sum_{t \in \mathcal{T}_e} \frac{1}{N} w_{i,t}^{(k)} \tilde{r}_{i,t+1}.$$

The total episode return can therefore be written as

$$R_e^{(k)} = \sum_{i=1}^N C_{e,i}^{(k)}.$$

This decomposition is used to identify which instruments contributed most to LSTM losses and to compare those same instruments against TSMOM and MACD.

H Return-Divergence and P&L-Weighted Attribution

This appendix gives an auxiliary decomposition used to interpret selected breakdown and outperformance episodes. It is not used as a formal performance metric. Instead, it helps identify whether LSTM underperformance comes from missed upside, wrong short exposure, bad long exposure, or concentrated instrument-level losses.

For a benchmark B , the daily return divergence between the LSTM and the benchmark is

$$\Delta R_{t+1}^{LSTM,B} = R_{t+1}^{LSTM} - R_{t+1}^B.$$

Substituting the equal-weighted portfolio construction gives

$$\Delta R_{t+1}^{\text{LSTM},B} = \frac{1}{N} \sum_{i=1}^N (w_{i,t}^{\text{LSTM}} - w_{i,t}^B) \tilde{r}_{i,t+1}.$$

Define the position gap

$$\Delta w_{i,t}^{\text{LSTM},B} = w_{i,t}^{\text{LSTM}} - w_{i,t}^B.$$

Then

$$\Delta R_{t+1}^{\text{LSTM},B} = \frac{1}{N} \sum_{i=1}^N \Delta w_{i,t}^{\text{LSTM},B} \tilde{r}_{i,t+1}.$$

This decomposition separates three intuitive sources of relative underperformance:

$$\text{missed upside: } w_{i,t}^B > w_{i,t}^{\text{LSTM}}, \quad \tilde{r}_{i,t+1} > 0,$$

$$\text{wrong short: } w_{i,t}^{\text{LSTM}} < 0, \quad \tilde{r}_{i,t+1} > 0,$$

$$\text{bad long: } w_{i,t}^{\text{LSTM}} > 0, \quad \tilde{r}_{i,t+1} < 0.$$

Integrated Gradients explains the model output, not realised P&L. To connect model sensitivity with realised economic outcomes, we also define an approximate P&L-weighted attribution:

$$A_{i,t+1,j}^{\text{P\&L}} = IG_{i,t,j} \tilde{r}_{i,t+1}.$$

Aggregating across instruments and over an episode $\mathcal{T}_e$ gives

$$A_{e,j}^{\text{P\&L}} = \sum_{t \in \mathcal{T}_e} \frac{1}{N_t} \sum_{i=1}^{N_t} IG_{i,t,j} \tilde{r}_{i,t+1}.$$

This connects model sensitivity to realised economic outcome, but it should be interpreted as a diagnostic approximation rather than a causal return decomposition.

References

- [1] Bryan Lim, Stefan Zohren, and Stephen Roberts. Enhancing time-series momentum strategies using deep neural networks. *The Journal of Financial Data Science*, 2019.
- [2] Marcos López de Prado. *Advances in Financial Machine Learning*. Wiley, 2018.
- [3] Daniel P. Palomar. *Portfolio Optimization: Theory and Application*. Cambridge University Press, 2025.
- [4] Kieran Wood, Sven Giegerich, Stephen Roberts, and Stefan Zohren. Trading with the momentum transformer: An intelligent and interpretable architecture. 2022. arXiv:2112.08534.
- [5] Zihao Zhang and Stefan Zohren. *Deep Learning in Quantitative Trading*. Elements in Quantitative Finance. Cambridge University Press, 2025.