

Increase Alpha: Performance and Risk of an AI-Driven Trading Framework

Sid Ghatak¹, Arman Khaledian², Navid Parvini², and Nariman Khaledian²

¹ Increase Alpha, LLC

s.ghatak@increasealpha.com

² Zanista AI Ltd.

arman.khaledian@zanista.ai, nariman.khaledian@zanista.ai,
navid.parvini@zanista.ai

Abstract. There are inefficiencies in the financial market, leaving unexploited patterns embedded in price, volume, and cross-sectional relationships. While recent advances increasingly employ large-scale transformer architectures, we take a domain-focused route: classical feed-forward and recurrent networks paired with expertly curated features to mine subtle regularities in noisy financial data. This smaller-footprint design is computationally lean and reliable under low signal-to-noise conditions—qualities that matter for daily production at scale. At *Increase Alpha*, we developed a deep-learning framework that maps a universe of over 800 U.S. equities into daily directional signals with minimal computational overhead.

The purpose of this paper is twofold. First, we outline the general overview of the predictive model without disclosing its core underlying concepts. Second, we evaluate its real-time performance through transparent, industry standard metrics. Forecast accuracy is benchmarked against both naive baselines and macro indicators. The performance outcomes are summarized via cumulative returns, annualized Sharpe ratio, and maximum drawdown. The best portfolio combination using our signals provides a low-risk, continuous stream of returns with a Sharpe ratio of more than 2.5, maximum drawdown of around 3%, and a near-zero correlation with the S&P 500 market benchmark. We also compare the model’s performance through different market regimes, such as the recent volatile movements of the US equity market in the beginning of 2025. Our analysis showcases the robustness of the model and significantly stable performance during these volatile periods.

Collectively, these findings show that market inefficiencies can be systematically harvested with modest computational overhead if the right variables are considered. This report will emphasize the potential of traditional deep learning frameworks for generating an AI-driven edge in the financial market.

Keywords: Finance · Algorithmic Trading · Risk Management · Deep Learning · Portfolio Construction

1 Introduction

Financial markets are often portrayed as an example of informational efficiency, a view that traces its intellectual lineage to the early-twentieth-century exchange between Louis Bachelier and Henri Poincaré. While reviewing Bachelier’s 1900 doctoral thesis, Poincaré remarked that if security prices truly followed an unrestricted random walk, the expectation of profit “could not rationally differ from zero.” Many decades later, Eugene Fama stated that intuition into the Efficient Market Hypothesis (EMH), arguing that all available information is instantaneously and correctly realized into prices. Yet the empirical literature now documents persistent pockets of predictability based on momentum, reversal, seasonal effects, and cross-sectional valuation that careful observers can still exploit.

In recent years, artificial intelligence has emerged as a transformative force in quantitative finance, lowering barriers to entry and enabling independent researchers and small trading teams to harness advanced methodologies once exclusive to major institutions. By leveraging open-source libraries and cloud computing platforms, practitioners can now prototype, validate, and deploy AI-powered strategies across diverse market regimes with unprecedented ease. This democratization of technology has given rise to a rich ecosystem of algorithms that span reinforcement learning, deep learning, ensemble models, and evolutionary techniques, each offering unique avenues for extracting predictive signals from complex financial data. For example, high-frequency trading now represents over 70 % of all executed trades [2], and modern HFT engines rely critically on machine-learning and deep-learning methods to scan multiple data streams, generate sub-second buy/sell signals, and execute vast numbers of orders [3,6]. Some implementations focus on drawdown control—e.g. a grid-trading DL system that adaptively reduces downside risk [8]—while others apply neural architectures to volatile assets such as Bitcoin, leveraging parallel processing and Bayesian regularization to forecast returns in noisy, high-frequency environments [4].

Reinforcement learning agents frame trading as a sequential decision-making problem, iteratively learning to select buy, hold, or sell actions that maximize cumulative reward under evolving market conditions. While these methods can uncover highly nonlinear policies that adapt to regime shifts, they require vast historical datasets and significant computational resources and carry a pronounced risk of overfitting unless subjected to rigorous backtesting and paper-trading validation [1]. Complementing this, Long Short-Term Memory (LSTM) networks excel at modeling temporal dependencies in price series, capturing both seasonality and momentum trends; however, they remain vulnerable to unanticipated “black swan” events and mandate careful feature engineering—integrating technical indicators alongside raw OHLCV data—and walk-forward validation to ensure robustness [11].

Supervised timing models—most notably tree-based ensembles such as Random Forests and XGBoost—classify market regimes by leveraging a curated set of features that may include momentum scores, volatility measures, and sentiment indices. Their interpretability via feature-importance metrics facilitates

systematic refinement, yet they also demand prudent cross-validation protocols to avoid capitalizing on spurious historical patterns [10]. In parallel, Generative Adversarial Networks (GANs) offer a novel data-augmentation paradigm, synthesizing realistic time-series scenarios to fill gaps in rare or stressed-market conditions; despite their promise in mitigating overfitting, GANs introduce training instability and necessitate thorough statistical validation of the synthetic output against empirical distributions [12].

Natural-language processing pipelines extend the AI toolkit by quantifying sentiment from unstructured text sources—news articles, social-media posts, and corporate filings—thus capturing behavioral drivers that often precede price movements. Beginning with lexicon-based scoring and progressing to fine-tuned transformer architectures, these systems must contend with evolving language trends, sarcasm, and manipulation, calling for continuous retraining and robust noise-filtering mechanisms [7]. Similarly, multi-factor AI models synthesize fundamental, technical, and alternative datasets to rank and select assets; machine-learning algorithms optimize factor combinations and, through systematic rebalancing and transaction-cost modeling, seek to sustain persistent outperformance while managing factor decay.

Dynamic portfolio allocators harness Bayesian inference or reinforcement learning to adaptively balance risk and return, continuously updating asset weights in response to shifting volatility and correlation structures. By incorporating frictional cost simulations and liquidity constraints into backtests, these systems aim to enhance risk-adjusted performance while controlling turnover. Across all methodologies, the pathway to robust deployment consistently emphasizes thorough backtesting, staged paper trading, and ongoing out-of-sample monitoring, thereby laying the groundwork for sustainable, data-driven trading paradigms in the era of AI. AI-driven trading frameworks typically fuse three pillars of information:

- **Technical analysis:** historical price, volume, and order-book dynamics
- **Fundamental analysis:** the impact of corporate and macroeconomic news on asset valuations
- **Investor sentiment:** emotion and opinion signals mined from social media and news feeds

At *Increase Alpha*, we believe that market inefficiencies can be harvested systematically by using the information mentioned above mixed with classical feed-forward and recurrent neural networks curated by domain experts. The use of transformers and large language models in Finance has gained popularity among the experts in the field in recent years [5,9,7]. However, in *Increase Alpha*, rather than pursuing massive transformer models on unstructured text, our minimalist design extracts only the variables that a seasoned fundamental analyst would consider economically meaningful, and then uses compact networks to uncover non-linear interactions at scale. The result is a daily, security-level directional signal for 814 U.S. equities—generated with negligible computational overhead and latency compatible with both discretionary and algorithmic execution.

Because the model and feature definitions are proprietary, we treat the architecture as a black box. In Section 2 we describe our data collection, signal generation, and execution infrastructure. In Section 3 we benchmark performance—using cumulative return, annualized Sharpe ratio, and maximum drawdown—against naive baselines and macroeconomic indicators, and examine behavior across different market regimes, including the turbulent U.S. equity markets of early 2025.

The remainder of the paper is organized as follows:

- Chapter 2 — Methodology
 - Describes the signal-generation pipeline, including multi-horizon forecasting, timestamp-tracked storage, and implementation mechanism.
 - Explains how execution logic is calibrated using hyperparameter grid search for Profit-Taker, Stop-Loss, and Maximum Holding Period.
 - Outlines the cloud-based deployment on Azure AKS and the associated scalability and cost-efficiency considerations.
- Chapter 3 — Accuracy and Statistical Significance
 - Measures signal performance across 814 tickers using multiple holding horizons and directional classifications (long, short).
 - Applies statistical rigor, including z-tests and confidence intervals, to validate signal effectiveness against random baselines.
 - Summarizes signal reliability by evaluating coverage, profitability, and sample-size-adjusted significance.
- Chapter 4 — Risk and Return
 - Translates signal outputs into cumulative return/PnL, drawdown, and Sharpe ratio metrics using the optimized execution configuration.
 - Compares the strategy’s performance against a buy-and-hold baseline and macro indices such as the S&P 500.
 - Highlights robustness under stress through a regime-based analysis—contrasting pre- and post-January 2025 performance—and visualizes results.
- Chapter 5 — Portfolio Construction and Dynamic Rebalancing
 - Converts signals into a real-world trading strategy using dynamic stock selection and quarterly rebalancing.
 - Portfolio configurations and evaluates performance via P&L, Sharpe ratio, and drawdown.
 - Showcasing the adaptability of signals across regimes, confirming the real-world viability of the approach.
- Chapter 6 — Conclusion
 - Summarizes the predictive power, statistical rigor, and market resilience of the *Increase Alpha* system.
 - Emphasizes that classical deep learning when paired with expert feature design and scalable infrastructure, offers a viable and interpretable alternative to more opaque architectures in financial signal generation.

2 Methodology

Our objective in this chapter is to discuss—without revealing proprietary intellectual property—the full life-cycle by which the *Increase Alpha* framework generates, stores, filters, and evaluates daily equity-level trading signals. We also describe the metrics we used to evaluate the performance of the generated signals. Every step is designed to eliminate *look-ahead bias* and *information leakage*, which are the main pitfalls in any trading strategies. This section is organized as follows:

1. Data-generation and signal specification (directional forecasts, storage, and tradable conventions)
2. Scenario analysis and cloud infrastructure (trading parameters search for profit-taker, stop-loss, and maximum holding period)
3. Evaluation framework (accuracy tests, economic metrics, and robustness checks)

2.1 Data-generation and signal specification

Since **28 June 2021**, the production system has executed a daily inference cycle for a universe of 814 U.S. equities. Each trading day, shortly after the market close, the latest data—including official corporate actions, fundamental metrics, and price/volume features—are ingested and processed by our deep learning model.

The model generates ten distinct directional predictions per stock for the upcoming ten trading sessions (i.e., $t + 1$ to $t + 10$). Each signal forecasts the expected percentage price change for a specific future session, forming a sequence of forecasts from a single prediction date. For example, the prediction on 28 June 2021 includes 10 separate directional estimates for each of the trading sessions from 29 June to 13 July 2021.

Each forecast is stored along with two immutable timestamps, capturing the creation and final update moments of each signal. Every prediction is finalized only after the market close, using solely available data, and is time-stamped and committed before the opening of the target session. This rigorous tracking ensures a true ex-ante prediction record, immune to post-hoc adjustment or backtest contamination, and serves as a strong guardrail against *look-ahead bias* and *information leakage*. Each model run generates a rolling set of multi-horizon signals—ten predictions that span the next ten trading sessions—forming a comprehensive forward-looking view. This automated pipeline has been running uninterrupted since June 2021, producing a longitudinal dataset with consistent daily operations.

We source price data from www.eodhd.com, a comprehensive market data provider. Their paid subscription gives access to a wide range of live and historical market information. We collected daily OHLC prices for all 814 U.S. equities from 28 June 2021 to 30 June 2025.

The automated system has incurred a consistent operational cost since inception. Between November 2024 and April 2025, daily AWS compute expenses averaged approximately \$95–\$100 USD, with occasional spikes on high-load processing days. When including all essential infrastructure services—such as EC2 instances, storage (S3), RDS databases, SageMaker for model hosting, and supporting utilities—the total infrastructure expenditure for the period amounted to approximately \$17,000 USD. Notably, this excludes optional analytics services (e.g., QuickSight, LIT for Traders), which were not part of the inference or training pipeline. This steady and predictable cost profile underscores the system’s viability for long-term deployment and financial forecasting operations.

Table 1 presents an example prediction table for AAPL on 28 June 2021:

Table 1. Example multi-horizon-ahead predictions with timestamp tracking

Current Date-time	Ticker	Target Date	Forecasted return	Horizon
2021/06/28 - 21:30	AAPL	29/06/2021	+0.5835	1
2021/06/28 - 21:30	AAPL	30/06/2021	-3.5856	2
2021/06/28 - 21:30	AAPL	01/07/2021	+1.1635	3
2021/06/28 - 21:30	AAPL	02/07/2021	-1.2820	4
2021/06/28 - 21:30	AAPL	06/07/2021	-0.5109	5
2021/06/28 - 21:30	AAPL	07/07/2021	-0.5405	6
2021/06/28 - 21:30	AAPL	08/07/2021	-0.2841	7
2021/06/28 - 21:30	AAPL	09/07/2021	-0.3977	8
2021/06/28 - 21:30	AAPL	12/07/2021	-0.4024	9
2021/06/28 - 21:30	AAPL	13/07/2021	-0.2335	10

For each ticker the engine outputs a predicted return. We use the sign of these predicted returns as *ternary direction code* $s \in \{+1, 0, -1\}$:

- +1 (Long) — expected positive open-to-close return next session,
- -1 (Short) — expected negative return,
- 0 (Flat) — neutral or low-confidence regime.

2.2 Scenario Analysis and Trading Parameter Optimisation

To convert directional signals into economic value, we extract three parameters necessary for trade execution.

- **Profit-Taker (PT)** — positive return threshold for exit;
- **Stop-Loss (SL)** — adverse move forcing liquidation;
- **Maximum-Holding Period (MHP)** — maximum duration (days) a position may remain open.

To obtain these values, we design a backtesting algorithm that simulates real-world trading on historical data. We then conduct a grid search as a scenario analysis. For each ticker, we store the performance of each item in the signal universe under different PT, SL, and MHP settings. For each ticker, we run the scenario analysis over:

-
- **Maximum Holding Period (MHP)**: ranges from 1 to 10;
 - **Profit-Taker (PT)**: ranges from 0.001 to 0.02 in increments of 0.0005;
 - **Stop-Loss (SL)**: ranges from -0.04 to -0.01 in increments of 0.005.

In each scenario, we use the signal directions and the prices below to compute the trade return r , Sharpe ratio (SR), and maximum drawdown (MDD). For day t and ticker i , let

$$r_{i,t} = \begin{cases} \text{PT}, & \text{if } H_{i,t:t+MHP} - O_{i,t} \geq \text{PT}, \\ \text{SL}, & \text{if } L_{i,t:t+MHP} - O_{i,t} \leq \text{SL}, \\ C_{i,t+MHP} - O_{i,t}, & \text{otherwise,} \end{cases}$$

where

$$H_{i,t:t+MHP} = \max\{H_{i,t}, H_{i,t+1}, \dots, H_{i,t+MHP}\},$$

$$L_{i,t:t+MHP} = \min\{L_{i,t}, L_{i,t+1}, \dots, L_{i,t+MHP}\},$$

and O_t , H_t , L_t , and C_t are the open, highest, lowest, and close price at day t .

In total, for each item in the trading universe we measure performance across 2,280 scenarios (MHP: 10; PT: 38; SL: 6). Given the large universe and the 10 different trading signals (horizon 1 through horizon 10 prediction signals), the computational load to extract the optimal execution values is substantial. Therefore, we use Microsoft Azure to containerize and run the computation in parallel.

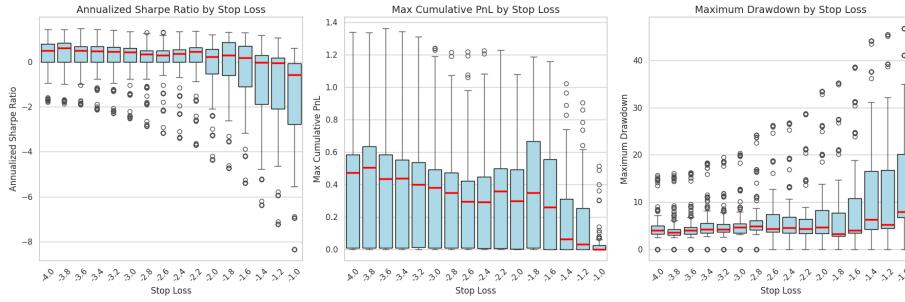
To efficiently execute this large-scale scenario analysis, we leverage Microsoft Azure Kubernetes Service (AKS) for parallel computation. To manage the grid search for optimal execution parameters (MHP, PT, and SL), which requires substantial compute, we containerize the jobs and distribute the workload across 20 pods. Each pod is provisioned with 1.6 CPU cores and 6 Gi of memory, and the average runtime per pod is approximately 8 days (192 hours). The underlying infrastructure uses Standard D4s v3 virtual machines, each offering 4 vCPUs and 16 Gi RAM, providing a balanced compute environment. The AKS cluster is autoscaled to 20 nodes to accommodate the distributed jobs efficiently.

From a cost perspective, we estimate the cloud compute expenditure as follows. The Standard D4s v3 VM is priced at approximately \$0.376 per hour. Given the configuration of 10 VMs supporting 20 pods over 192 hours, total VM usage amounts to 1,920 hours, leading to a total compute cost of \$722 (Table 2 summarizes the AKS configuration and cost). As obtaining the optimal execution values is a one-time task that does not need to be repeated during trading, using cloud infrastructure makes the scenario analysis both scalable and cost-effective.

For each ticker-signal pair, the system produced a comprehensive set of outputs. For every scenario evaluated, three visualizations summarize performance under variations in MHP, PT, and SL thresholds. Figure 1 shows an example for Commvault Systems, Inc. (CVLT). The figure illustrates scenario performance across different stop-loss levels. As shown, most scenarios perform better when the Profit-Taker (PT) is set to 3.8% of the entry price.

Table 2. AKS configuration and cost.

Pods	20 (approximately 20 tickers per pod)
Pod resources	1.6 vCPU, 6 GiB RAM
Node type	Standard D4s v3 (4 vCPU, 16 GiB RAM)
Autoscale window	1 – 20 nodes
Wall-clock runtime	~8 days
Total VM-hours	1 920
Total VMs	10
Azure list price	\$0.376 per hour
Total compute cost	\$722

**Fig. 1.** Boxplot showing the distribution of Sharp Ratio, cumulative return/PnL (%), and Drawdown for different cases in the scenario analysis.

Moreover, the results of each case are stored in a dataframe detailing the performance metrics. These results are saved to Azure Blob Storage and made available for further analysis. This scalable, cloud-based setup enabled a comprehensive scenario analysis, allowing us to test parameterized signal-execution strategies at scale while maintaining full reproducibility and minimizing operational overhead.

3 Accuracy and statistical significance

We evaluate the directional accuracy and performance of trading signals across our entire trading universe. In this section, we present the results of our analysis, including detailed accuracy scores and statistical significance of the measured accuracies.

3.1 Accuracy

When a signal is positive, a **long** position is opened on the next available trading day. When a signal is negative, a **short** position is opened on the next available trading day. For each asset, the directional accuracy over a specific *holding period* h is computed with respect to the sign of its return over that period. Consistent with the MHP search space in the previous section, we examine trade performance for 11 different holding periods (business days), labeled from 0 to 10. The

multi-horizon-ahead return for each ticker is calculated by comparing the close at holding period h , i.e., day $t + h$, with the open on day $t + 1$. Since positions are opened one business day after the signal is generated, we assume the signal pertains to day $t + 1$.³ For example:

- A holding period of **0** indicates that the return is calculated as the percentage change between the open on the next available day after the signal is generated and the close of that same day.
- A holding period of **1** indicates that the return is the percentage change between the open price on the next available day and the close price one day later.
- Similarly, a holding period of **k** indicates that the return is the percentage change between the open on the next available day after the signal is generated and the close **k** days after the position is opened.

For each ticker, we track how often the direction of the generated signals aligns with the direction of the return for the same signal horizon and holding period. Accordingly, **accuracy** is computed as the ratio of correctly predicted directions to total predictions:

$$\text{Accuracy} = \frac{\text{Number of correctly predicted directions}}{\text{Total number of predictions}} \times 100\%.$$

The final aggregated results of the accuracy analysis for 814 stocks consist of the items defined as follows:

- **Ticker**: The stock ticker symbol (e.g., AAPL).
- **long_0, long_1, …, long_10**: The accuracy (in percentage) of all long signals held for 0, 1, …, 10 days, respectively.
- **avg_long**: The average accuracy (in percentage) of all long signals across all holding periods (from 0 to 10).
- **max_long**: The maximum accuracy among long_0, long_1, …, long_10.
- **min_long**: The minimum accuracy among long_0, long_1, …, long_10.
- **pct_long**: The percentage of signals that were long out of all signals for the ticker.
- **best_day_long**: The specific holding period (0–10) at which the long strategy achieved max_long accuracy.
- **short_0, short_1, …, short_10**: The accuracy (in percentage) of all short signals held for 0, 1, …, 10 days, respectively.
- **avg_short**: The average accuracy of all short signals across the same holding periods (0 to 10).
- **max_short**: The maximum accuracy among short_0, short_1, …, short_10.
- **min_short**: The minimum accuracy among short_0, short_1, …, short_10.
- **pct_short**: The percentage of signals that were short out of all signals for the ticker.
- **best_day_short**: The specific holding period (0–10) at which the short strategy achieved max_short accuracy.

³ As previously discussed, the signal is generated at the end of the trading session on day t to forecast days $t + 1, t + 2, \dots, t + 10$. This ensures any potential look-ahead bias is addressed in all stages of analysis and simulation.

3.2 Statistical Significance and Reliability

Since the long/short signal distribution is inherently class-imbalanced, we apply robust statistical methods to ensure a fair comparison of signal accuracies. To determine whether the observed accuracies are statistically significant, we conduct formal tests and report p-values and confidence intervals. To evaluate reliability and significance, we consider:

- Using a **binomial test** or a **z-test** for proportions to assess whether measured accuracies deviate significantly from random chance (e.g., 50%).
- Constructing **confidence intervals** for each accuracy measure (e.g., a 95% confidence interval around the estimated accuracy).

Interpretation of Statistical Metrics:

- **Accuracy (%)**: This is the percentage of signals whose directions are aligned with the direction of the observed return for a particular signal type (long or short) and a particular holding period. For instance, if `long_0` is 57.0, it indicates that out of all same-day long signals generated for that ticker, 57% were pointing to the same direction as the true return observations.
- **Sample Size** (e.g., `n_long`, `n_short`): This is the total number of long or short signals generated for each ticker. If a particular ticker had 100 long signals, then the accuracy of `long_0` is based on those 100 signals. The larger the sample size, the more confidence one can have in the overall observed accuracy.
- **p_value_vs_50%** (**p-value from the proportions z-test**): We compare each observed accuracy to a baseline (commonly 50%) under the null hypothesis H_0 :

$$H_0 : \text{True accuracy is } 50\%,$$

$$H_1 : \text{True accuracy differs from } 50\%.$$

The p-value of this test will let us know how likely it is to observe the measured accuracy assuming the true accuracy is 50%. This is done using a z-test for proportions, where the standard error (SE_0) is defined as:

$$SE_0 = \sqrt{\frac{p_0(1 - p_0)}{n}},$$

with p_0 the baseline proportion (e.g., 0.5) and n the number of observations. The observed accuracy \hat{p} is then compared to the baseline p_0 using the z-score:

$$z = \frac{\hat{p} - p_0}{SE_0},$$

which quantifies how many standard errors the observed proportion is from the baseline. A larger absolute z-score indicates a greater deviation between observed and expected accuracy.

A smaller standard error—occurring when n is large—makes the test more sensitive to differences. Consequently, the p-value, computed from the z-score, tends to decrease with larger sample sizes if the observed accuracy remains consistently above (or below) 50%.

A low p-value (e.g., < 0.05) suggests that the observed accuracy is statistically different from 50% and unlikely to arise by chance. Conversely, a high p-value indicates that the observed accuracy could plausibly occur even if the signal were generated by a random binary process.

Example: Suppose the observed accuracy is 60% on a sample of $n = 200$ trades, and we test against a baseline of 50%:

$$\hat{p} = 0.6, \quad p_0 = 0.5, \quad n = 200.$$

The standard error is:

$$SE_0 = \sqrt{\frac{0.5 \times (1 - 0.5)}{200}} = \sqrt{\frac{0.25}{200}} \approx 0.0354.$$

The corresponding z-score is:

$$z = \frac{0.6 - 0.5}{0.0354} \approx 2.82.$$

A z-score of 2.82 corresponds to a p-value well below 0.05, indicating that the observed accuracy is significantly better than chance.

- **ci_lower, ci_upper (confidence interval for observed accuracy):** In addition to computing a p-value, we estimate a confidence interval (CI) around the observed accuracy \hat{p} , which provides a plausible range for the true accuracy based on the sample data.
For a 95% confidence level, the interval is computed using the normal approximation (Wald interval):

$$CI = \hat{p} \pm z_{\alpha/2} \times SE_{\hat{p}},$$

where $\hat{p} = \frac{\text{successes}}{n}$ denotes the observed accuracy, and $z_{\alpha/2}$ is the critical value from the standard normal distribution (e.g., 1.96 for a 95% CI). The standard error $SE_{\hat{p}}$ is given by:

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}.$$

The lower and upper bounds of the confidence interval are:

$$\text{Lower} = \hat{p} - z_{\alpha/2} \times SE_{\hat{p}}, \quad \text{Upper} = \hat{p} + z_{\alpha/2} \times SE_{\hat{p}}.$$

Effect of sample size: As the sample size n increases, the standard error decreases, yielding a narrower (more precise) confidence interval. Conversely, with a small sample size, the standard error is larger, making the confidence interval wider and less certain.

Example: Suppose the observed accuracy is 60% ($\hat{p} = 0.6$), and we compute the 95% confidence interval using the normal method:

- **Case A (small sample):** $n = 20$

$$SE_{\hat{p}} = \sqrt{\frac{0.6 \times 0.4}{20}} \approx 0.1095, \quad CI = 0.6 \pm 1.96 \times 0.1095 = [0.385, 0.815]$$

- **Case B (large sample):** $n = 200$

$$SE_{\hat{p}} = \sqrt{\frac{0.6 \times 0.4}{200}} \approx 0.0346, \quad CI = 0.6 \pm 1.96 \times 0.0346 = [0.532, 0.668]$$

As shown above, a larger sample size leads to a tighter confidence interval around the observed accuracy.

In summary, analyzing all four elements (accuracy, sample size, p-value, and confidence intervals) offers deeper insight into the reliability of the trading signals:

- **Accuracy** indicates the observed performance.
- **Sample size** shows how many trades underlie that observed performance.
- **p-value** reveals whether the result is *statistically* different from 50%.
- **Confidence intervals** give a range of plausible values for the *true* accuracy.

The results of the signal accuracy analysis is stored for each ticker where we have the observed accuracy (in percentage), the number of signals generated (sample size), the p-value from a z-test against a 50% success rate, and the 95% confidence interval for the true accuracy. As an example, Table 3 shows the accuracy results for "CommVault Systems, Inc." (CVLT), one of the items in the covered universe. In this table, apart from above mentioned information, we have the strategy that represents the direction at which the ticker should be traded and period signal that shows the best signal period for the ticker. Both of these parameters are found by analysing the optimization process in section 2.2.

Table 3. CVLT signal accuracy Summary.

Metric	Value
Ticker	CVLT
Name	CommVault Systems, Inc.
Strategy	Long Only
Period Signal	3
Pct_long	64.2 %
Long Accuracy	67.76 %
Short Accuracy	59.62 %
Sample_Size	577
p_value_vs_50%	6.86E-20
CI_95_lower_%	63.95 %
CI_95_upper_%	71.57 %

As we can see, the occurrence of the long trading signals for CVLT is relatively high at 64.2% of the generated signals, and its confidence interval and

p-values show that the measured accuracy for this stock is statistically significant. Beyond summary statistics such as long accuracy and short accuracy, this extended evaluation allows us to statistically interpret how reliable each signal truly is. Two tickers may have identical accuracies, but one may be statistically significant (based on sample size and confidence intervals), while the other is not.

Table 4. P-value Summary Statistics.

Signal Type	Mean	Med.	Std. Dev.	Min	Max	% $p < 1\%$	% $p < 5\%$	% $p < 10\%$
Long	0.0259	0.0002	0.0670	0.0000	0.5979	74.948	86.349	90.900
Short	0.0367	0.0005	0.0823	0.0000	0.5742	68.985	81.642	86.402
Both	0.0313	0.0003	0.0752	0.0000	0.5979	71.967	83.996	88.651

Taking a broader picture of the statistical test results, the summary presented in Table 4 reports the distributional characteristics of the p-values obtained across all tickers, separately for long, short, and both signals. Several important features stand out. First, the mean p-values are low (around 0.026 for long signals and 0.037 for short signals), which already suggests that, on average, the null hypothesis of random 50% accuracy is strongly rejected. The medians are even smaller (0.0002 for longs and 0.0005 for shorts), highlighting that the majority of p-values are concentrated near zero rather than around conventional significance cutoffs.

Most importantly, the proportion columns show that the vast majority of signals meet conventional statistical significance levels. For example, more than 74% of long-signal p-values fall below 1%, and over 86% fall below 5%. Even for short signals, 69% are below 1% and over 81% below 5%. When both directions are aggregated, nearly 84% of all signals achieve 5% significance. Such proportions are high for financial prediction tasks, where genuine edge is typically weak, multiple-testing is pervasive, and microstructure noise and near-efficiency usually limit the share of signals clearing 5%—let alone 1%—especially after accounting for serial dependence and out-of-sample validation.

Figure 2 displays the cross-universe accuracy and 95% confidence intervals for the long and short signals over the period from the first observation to the end of 2024, with tickers ordered by increasing observed accuracy on the horizontal axis. The solid green line shows the observed accuracy for each ticker, computed using the best-performing signal period and holding period identified for that stock; the shaded band and dashed edges show the empirical 95% confidence interval obtained from the z-test (binomial approximation) around the observed accuracy.

Two salient features emerge. First, the distribution of observed accuracies is skewed above the 50% benchmark: for long signals, more than 90% of tickers achieve an accuracy above 50%, while for short signals this figure is around 75%. Moreover, many of these also have 95% lower bounds above 50%, indicating

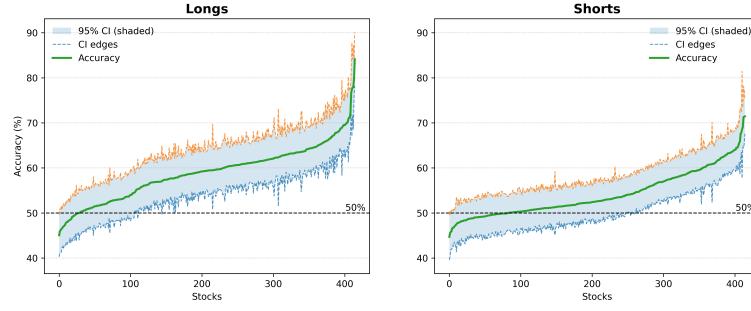


Fig. 2. Confidence interval plots for the period beginning from the first observation to the end of 2024. The stocks selected from the optimization process are ordered from lowest to highest accuracy on the horizontal axis. The solid green line shows the average accuracy for each ticker across the sample and the shaded areas shows the 95% confidence interval. The left plot shows long signals, while the right plot depicts short signals.

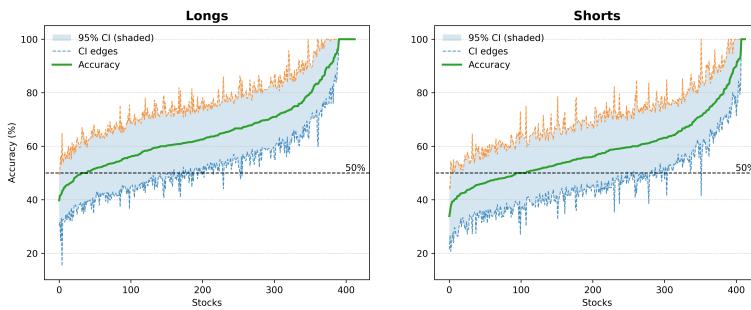


Fig. 3. Confidence interval plots for the period of the first two quarters of 2025. The stocks selected from the optimization process are ordered from lowest to highest accuracy on the horizontal axis. The solid green line shows the average accuracy for each ticker across the sample and the shaded areas shows the 95% confidence interval. The left plot shows long signals, while the right plot depicts short signals.

statistical evidence of predictive power beyond chance. Second, the confidence intervals maintain a relatively stable distance from the mean accuracy line across the entire ranking of tickers. This indicates that the level of statistical uncertainty is consistent regardless of whether the observed accuracy is moderate or high. In other words, the model does not become less reliable as accuracy improves; rather, the uncertainty remains controlled throughout the distribution. If the confidence intervals had widened disproportionately with higher accuracies, this would have suggested instability in the signal reliability. Instead, the uniform separation between the accuracy curve and its confidence bounds underscores the robustness of the results across the full range of signals.

Comparing panels, long signals are generally more accurate than short signals. This is consistent with the well-known upward drift in equity prices (positive equity risk premium), which creates an asymmetric environment where long-only signals are, on average, easier to exploit. However, the figure also shows many high-accuracy short signals, demonstrating that the signal generation process produces useful information on both sides of the market.

Figure 3 shows a similar plot for the first and second quarters of 2025. The analysis follows the same procedure as above (accuracies computed using each stock’s best-performing signal period and holding period), but the 95% confidence intervals are generally wider and their boundaries more volatile. This widening largely reflects averaging over a smaller sample observations per ticker and elevated market uncertainty during the period, both of which increase estimation variance. Despite the larger uncertainty, the model preserves its edge: a large majority of tickers remain above the 50% benchmark, and several tickers register perfect (100%) realized accuracy over the sample considered.

4 Risk and Return

While accuracy provides an initial indication of a trading signal’s potential effectiveness, it is not the sole determinant of success in financial markets. Ultimately, the most critical factors are the strategy’s returns and associated risks. To assess this balance, we rely on key performance metrics such as return, Profit and Loss (PnL)⁴, Sharpe ratio (SR), and drawdown. Having evaluated the accuracy and statistical significance of the trading signals in the previous section, we now analyze the strategy’s return profile and risk characteristics using these financial metrics.

4.1 Extracting Return and Risk Metrics

To evaluate the strategy’s risk-return characteristics, we use outputs from the scenario analysis and optimization step in Section 2.2, which identified the optimal Maximum Holding Period (MHP), Profit-Taker (PT), and Stop-Loss (SL) settings. These parameters, together with the OHLC price data, are used to compute both return and risk metrics for each ticker.

⁴ The concepts of PnL and returns are both used to refer to trade/portfolio returns in percentages throughout this study and used interchangeably.

Cumulative return/PnL: Cumulative return/PnL reflects the aggregated return of the strategy over time. It measures the percentage profit or loss generated by sequentially applying the trading signals over the historical price series, assuming no reinvestment of gains⁵. This metric captures both the frequency and magnitude of successful signals and provides an intuitive view of the strategy's performance trajectory.

Drawdown: Drawdown measures the percentage decline from a historical peak in cumulative returns to a subsequent trough. It captures the impact of sequential negative returns. **Maximum drawdown (MDD)** represents the largest observed drawdown over the evaluation period and serves as a key risk metric, highlighting the strategy's worst-case peak-to-trough loss and informing capital-preservation considerations under adverse conditions.

Annualized Sharpe Ratio: The Sharpe ratio measures risk-adjusted performance by comparing the portfolio's excess return over the risk-free rate to the volatility of its returns. To enable comparability across assets and strategies, it is typically expressed in annualized terms. A higher Sharpe ratio reflects more efficient compensation for risk undertaken.

4.2 Visualization of Risk-Return Characteristics

To better understand these metrics, we provide a three-part plot (Figure 4) visualizing the return and risk profiles of both the trading strategy and a benchmark buy-and-hold approach:

- **Top Subplot:** Shows the cumulative return/PnL in percentages (left y-axis) and drawdown (right y-axis) over time, with the high watermark displayed as a dotted line.
- **Middle Subplot:** Displays the price of CVLT (buy-and-hold) along with its drawdown, allowing comparison with the strategy's behavior. The legend includes correlation values between:
 - Daily return/PnL and CVLT price
 - CVLT and S&P 500
 - return/PnL and S&P 500
- **Bottom Subplot:** Shows the 60-day rolling Sharpe ratios of the trading strategy, CVLT buy-and-hold, and S&P 500, with annualized Sharpe ratios shown as horizontal dashed lines.

Key Takeaways from the plot is:

- The cumulative return/PnL of the strategy is significantly higher than the CVLT buy-and-hold return.

⁵ All cumulative returns in this study are calculated using a simple, non-compounded aggregation of returns, where each period's return is computed against a fixed notional capital base.

- The strategy's drawdown is approximately 10 times smaller than that of the CVLT buy-and-hold.
- The annualized Sharpe ratio of the strategy is greater than those of both CVLT and the S&P 500, demonstrating superior risk-adjusted returns.

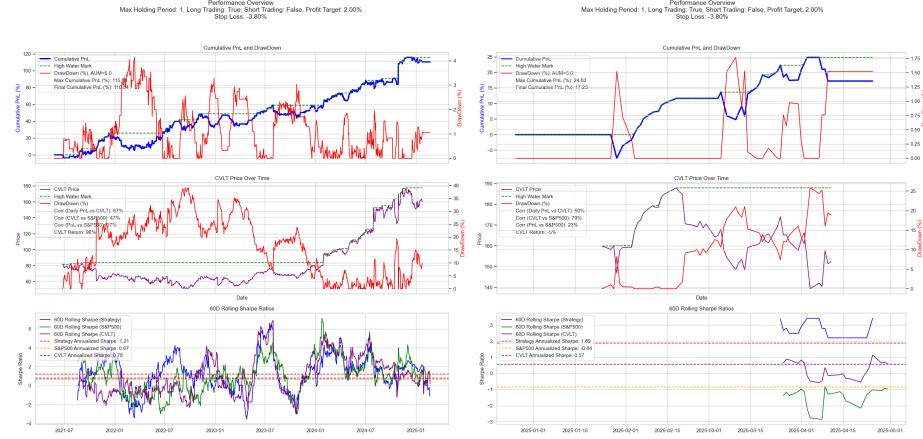


Fig. 4. Visualization of the strategy's returns and risk, compared with the macro benchmark (S&P 500) and the stock's buy-and-hold.

4.3 Stress Testing the Strategy

While the prior analysis focused on the period from June 2021 to January 2025, market conditions changed dramatically after January 2025.

Market Turbulence After January 2025: Following accelerated interest rate hikes, geopolitical tensions, and tech sector corrections, global markets saw increased volatility. Major indices experienced drawdowns exceeding 20%, correlations across asset classes broke down, and liquidity became strained. This period presented a substantial challenge for most trading systems.

Stress Test Methodology: We segmented the data into “Before January 2025” and “After January 2025” periods. Our goal was to evaluate whether the trading signals remained robust in these drastically different regimes.

From Figure 4, we observe that even though the returns of the S&P 500 and buy-and-hold scenario were not positive, our strategy signals managed to produce a positive return. This indicates that the signal performance was stable and unchanged during the harsh market regime.

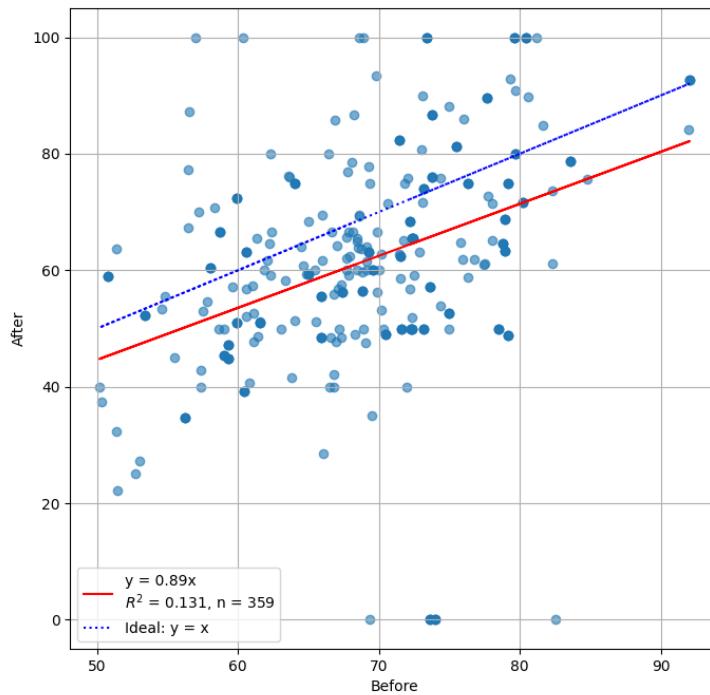


Fig. 5. Scatter plot showing the long accuracy of the trading signals, before and after January 2025.

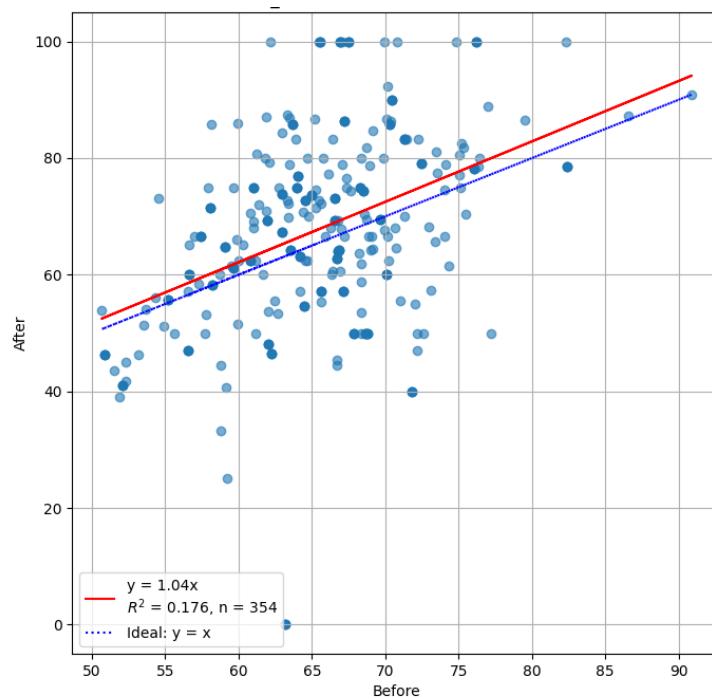


Fig. 6. Scatter plot showing the short accuracy of the trading signals, before and after January 2025.

Results: The scatter plots (see Figure 5 and Figure 6) display “Before” vs “After” accuracy of our sample.

Key findings include:

- The regression lines are noticeably close to the ideal $y = x$ line, indicating stability in signal performance across regimes.
- The slope and R^2 values demonstrate consistency and predictive strength.

Conclusion: These findings strongly support that:

- The trading strategy is robust and largely uncorrelated with broader market movements.
- It continues to function effectively during historically adverse periods.
- The low correlation to CVLT and the S&P 500 implies diversification benefits.

5 Portfolio Construction and Dynamic Rebalancing

In the previous section, we evaluated the generated signals’ ability to predict the price movement of individual stocks. In what follows, we translate these insights into a practical investment strategy with an example of systematic portfolio construction and dynamic rebalancing. We detail the methodology employed to build and adapt a trading portfolio using our generated signals and evaluate the portfolio’s performance over time. We adopt a *walk-forward* design that strictly separates model evaluation from live decision-making.

The portfolio construction follows a six-quarter rolling window with one-quarter steps. More specifically, we start by using model outputs between mid-2021 and the end of 2022, corresponding to six quarters, as a calibration window. During this period, we aggregate daily directional signals across 814 U.S. equities and compute performance metrics—including directional accuracy, annualized Sharpe ratio, maximum drawdown, etc.—for each stock. For the stock-selection process, we sort the stocks based on their performance in each rolling selection window and select the top-performing names.

The constructed portfolio is then traded over the subsequent quarter, specifically from the beginning of 2023 through the end of Q1 2023, using the latest daily predictions from the model. At the end of each quarter, the portfolio is rebalanced: the prior 18 months of trading signals are analyzed, and new top-performing stocks are selected to form the next quarter’s portfolio. This rolling mechanism ensures that the portfolio adapts to evolving market conditions while maintaining a systematic, data-driven foundation. Note that the selection of a stock does not guarantee its inclusion in the portfolio; this is a watch list, and stocks are included only if the signals generated for that stock during the trading quarter—e.g., the first quarter of 2023—are in line with the desired trading direction assigned to the stock as a result of the optimization process. However, based on the analysis in the results section, the majority of the selected stocks are traded during the trading quarter.

We continue this rebalancing process through the end of 2024 and into the first and second quarters of 2025, resulting in a comprehensive evaluation of the dynamic portfolio strategy across multiple market regimes. Therefore, this example demonstrates the practical viability and robustness of signals generated by our model in a real-world trading context.

5.1 Results

This section reports the out-of-sample trading results, examining how the strategy built from the model’s signals compares with a passive S&P 500 buy-and-hold benchmark. Because the principal aim is to demonstrate the economic value of the signals, our focal strategy is the maximum drawdown-based, equally weighted portfolio. Prior to selection, we remove stocks with too few observations over the selection window and exclude stocks with market beta greater than one. We then rank the remaining universe according to the chosen metric: for risk-oriented measures such as maximum drawdown (MDD), stocks are ordered from lowest to highest and the lowest-MDD names are selected; for return-oriented measures such as the Sharpe ratio (SR), stocks are ordered from highest to lowest. For the MDD-based portfolio, at each rebalancing quarter we take the 20 lowest-MDD names from the stocks selected for long trading during the optimization process and the 20 lowest-MDD names from the stocks selected for short selling. We compare cumulative returns, risk-adjusted performance, drawdown behavior, and time-varying performance of this portfolio to the passive benchmark. The effects of alternative selection criteria are examined later in the section. Linearly weighted variants deliver qualitatively similar behavior with slightly weaker outcomes; where relevant, we summarize those results but do not plot them. All performance statistics are computed on the rolling, walk-forward trading sample described in Section 5.

Portfolio performance

Aggregate returns and risk-adjusted performance: Figure 7 depicts the out-of-sample performance of the MDD-based, equally weighted (MDDEW) portfolio, benchmarked against a passive S&P 500 buy-and-hold. Over the test interval, the strategy’s final cumulative return is $\approx +26.4\%$ (peak cumulative return $\approx 26.95\%$) with a maximum drawdown of $\approx 3.0\%$ (values read from the annotated figure). Although the passive S&P 500 over the same interval delivers larger nominal appreciation (final cumulative return $\approx +51.4\%$), the MDDEW portfolio exhibits higher risk-adjusted performance. The strategy’s annualized Sharpe ratio (computed from daily returns and displayed in the plot legend) is ≈ 2.54 , compared with ≈ 1.19 for the S&P 500. Thus, the MDDEW portfolio achieves higher return per unit of daily volatility despite producing a lower absolute cumulative return. This difference—smaller nominal gains but higher Sharpe—summarizes the main empirical result: the model’s signals can be converted into a tradable book that is more efficient on a risk-adjusted basis than passive market exposure.

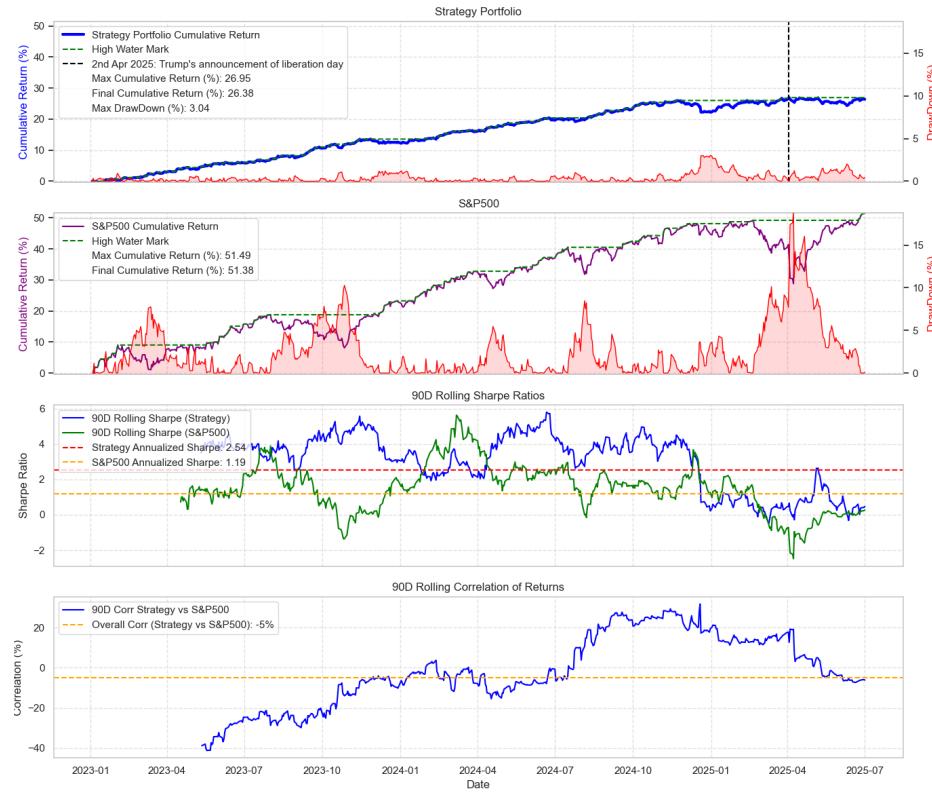


Fig. 7. Out-of-sample performance of the **MDD-based, equally-weighted (MD-DEW)** portfolio, benchmarked against the passive S&P 500 buy-and-hold.

Correlation with market returns: The relation between the strategy and broad market returns is small on average and time varying. The annotated overall correlation between the strategy’s daily returns and S&P500 returns is $\approx -5\%$, indicating that MDDEW portfolio is largely orthogonal to the benchmark over the full sample. This low average correlation implies that, from a portfolio construction standpoint, MDDEW portfolio behaves like an idiosyncratic (alpha) sleeve that can complement passive equity exposure; it does not appear to be a levered or variant expression of market beta.

The figure’s 90-day rolling correlation series, however, reveals that this correlation is not constant: correlation drifts upwards through 2024—reaching positive territory in mid-2024—before falling back toward slightly negative values in 2025. That time dependence indicates that MDDEW portfolio’s diversification benefit is regime dependent: in some subperiods the strategy’s returns becomes more aligned with market moves (reducing diversification), while in others it provides a defensive, low-correlation source of returns. Any practical allocation that pairs MDDEW portfolio with passive holdings should therefore treat correlation as a variable quantity and consider dynamic sizing or stress testing for periods of elevated co-movement.

Drawdowns and resilience through market events: A comparison of the drawdown traces highlights the different risk profiles of MDDEW portfolio and the S&P 500. The S&P line shows larger and deeper drawdown episodes (notably the drawdown spike visible in the middle panel), whereas MDDEW portfolio’s drawdown envelope is shallow and contained (maximum drawdown $\approx 3\%$). Importantly, MDDEW portfolio’s limited drawdowns imply that its positive cumulative return is achieved with relatively little depth of capital erosion between highs—a property attractive to risk-conscious investors even where absolute return is lower than the benchmark.

The annotated vertical event on the plot coincides with a period of elevated market stress. The S&P exhibits a sharp drawdown and rapid recovery during that interval, while MDDEW portfolio experiences a modest and brief deterioration in value before re-attaining prior highs. This behaviour suggests that the signals underlying MDDEW portfolio capture cross-sectional opportunities that are not tightly coupled to the market shock that produced the benchmark’s large drawdown, reinforcing the earlier observation that MDDEW portfolio is not simply capturing market beta.

Time-varying behaviour – rolling Sharpe and stability: The 90-day rolling Sharpe panel demonstrates that MDDEW portfolio’s performance pattern is both persistent and episodic. MDDEW portfolio’s rolling Sharpe spends the majority of the window above the S&P’s rolling Sharpe and repeatedly attains values in the region of 2–5 during favourable subperiods, consistent with the high aggregate Sharpe (≈ 2.54). This episodic concentration of high short-window Sharpe indicates that the model does not produce uniform small gains every day; instead it identifies pockets of cross-sectional opportunity that, when harvested repeatedly, build a high-quality return stream.

Toward late 2024 and into 2025 the strategy’s short-window Sharpe declines toward more moderate levels, mirroring the reduction in rolling correlation described above. The joint movement of rolling Sharpe and rolling correlation suggests a plausible narrative: when market conditions produce clear, coherent cross-sectional opportunities (for example, mid-2023 through mid-2024 in the figure), MDDEW portfolio tends to align with those trends, leading to higher gains during sustained market rallies. By contrast, in periods when the broad market weakens or fails to produce strong returns, MDDEW portfolio’s returns decouple from the benchmark—the strategy exhibits lower correlation and does not systematically amplify market losses—so that drawdowns remain contained. This dynamic — participation in market up-moves together with reduced co-movement during market stress — explains why MDDEW portfolio achieves higher gains in strong market advances while helping to mitigate losses during market declines.

Decomposed performance: Table 5 distills the out-of-sample performance by sleeve (long, short) and in aggregate. The combined long-short book produces a return/PnL stream: cumulative return $\approx 26.38\%$, annualized Sharpe ≈ 2.54 , and an MDD of $\approx 3.04\%$. Both sides contribute—longs at $\approx 15.75\%$ (Sharpe ≈ 1.87 ; MDD $\approx 3.99\%$) and shorts at $\approx 10.63\%$ (Sharpe ≈ 1.00 ; MDD $\approx 3.33\%$)—highlighting persistent edge on each sleeve. Execution is nimble: mean holding periods cluster around one trading day (0.98-1.08) under a six-day realized MHP, and the program executed 8,859 trades (4,336 long; 4,523 short). Hit rates are consistently above 55% (long 57.9%, short 55.4%, aggregate 56.6%), in line with the signal-accuracy results and supportive of the steady upward drift in cumulative return/PnL. In aggregate, the table shows a high-Sharpe portfolio with limited drawdowns and balanced contributions from long and short exposures.

Table 5. Portfolio performance metrics decomposed into long and short legs.

	Cum. Ret./PnL (%)	SR	MDD	Mean realized holding period (days)	Max. realized holding period (days)	Win rate (%)	Num. transac- tions
Long	15.75	1.872	3.988	1.080	6	57.9	4336
Short	10.63	1.001	3.333	0.885	6	55.4	4523
Combined	26.38	2.538	3.039	0.980	6	56.6	8859

Portfolio weighting: Throughout, all portfolios and model settings above are implemented as equally weighted baskets. As a robustness check, we also consider a simple linearly decaying scheme in which the top-ranked names receive higher weights that decrease in a linear fashion down the ranking list. Comparing with the best model identified above (MDDEW), this tilt modestly reduces overall

performance relative to equal weights. The linearly weighted variant delivers a strategy Sharpe of ≈ 2.25 , a maximum drawdown of $\approx 3.37\%$, a slightly negative market correlation ($\approx -4\%$), and a final cumulative return of $\approx 26.8\%$.

Role of selection metric: To further assess the robustness of the signals, Table 6 reports the out-of-sample performance of a broader set of portfolio selection rules and simple filters rather than focusing on a single construction. Every row in the table uses the same 6-quarter calibration window and the same quarterly rebalancing; what changes are the criterion used to rank names (Sharpe, drawdown, final cumulative return, Sortino, beta, downside risk, accuracy, etc.), the application of a simple beta filter, and the number of stocks selected. The purpose of this exercise is straightforward: we want to show how alternative, plausible selection choices affect realized returns, risk, and correlation to the market, and to assess whether the model’s signals produce performance across several settings.

The results are informative and largely consistent. Risk-aware ranking rules — in particular realized Sharpe and low-drawdown ranking — produce higher risk-adjusted outcomes. For example, ranking by Sharpe (row 1) yields a strategy Sharpe ≈ 1.98 , a final cumulative return $\approx 30.7\%$, and a maximum drawdown of $\approx 6.2\%$; ranking by drawdown (row 2) gives a slightly higher Sharpe (≈ 2.20) with an especially small maximum drawdown ($\approx 3.1\%$) and a final return $\approx 27.3\%$. Ranking by final cumulative return (row 3) and Sortino (row 4) also deliver positive cumulative returns ($\approx 24\text{--}31\%$) and Sharpe ratios above 1.5. By contrast, naive or purely direction-based rules perform poorly: the accuracy-ranked portfolio (row 11) posts a negative final return ($\approx -7.8\%$), a negative Sharpe (≈ -0.54) and a large drawdown, underscoring that directional accuracy alone does not reliably generate tradable, risk-adjusted alpha.

Simple filters and settings alter outcomes in expected ways. Excluding high-beta names (or, conversely, requiring $\beta > 1$) changes drawdown and return characteristics but does not erase the overall pattern: for example, applying a $\beta > 1$ filter and ranking by drawdown (row 8) produces a strong risk profile in the table (Sharpe ≈ 2.38 , MDD $\approx 2.5\%$, final return $\approx 29.3\%$), while retaining a top-10 selection under the same filter (row 7) yields comparable returns with a modestly higher drawdown. Increasing the number of selected names (row 9, top-20) slightly reduces final cumulative return but preserves a high Sharpe, illustrating that the signals remain useful when the portfolio is broadened. Two alternative ranking choices — ranking by beta alone (row 5) and ranking by downside risk (row 6) — produce weaker outcomes (low or negative Sharpe and small or negative cumulative return), which highlights that not all ex-post metrics are equally informative.

Stock turnover

Figures 8 and 9 display the color-coded quarter-by-quarter composition of the 20-stock trading book per side in a compact, rank-ordered grid: Figure 8 shows

Table 6. Summary of the other portfolio selection criteria and their performance.

	Window	Filter out	Ranked by	Top n	SR	MDD	Corr	Final Cum. Ret.
1	6Q	-	SR	10	1.98	6.23	9	30.72
2	6Q	-	DD	10	2.20	3.14	-1	27.31
3	6Q	-	Final Cum. Ret.	10	1.75	6.81	11	30.74
4	6Q	-	Sortino	10	1.60	7.36	11	24.34
5	6Q	-	Beta	10	0.70	6.70	6	7.92
6	6Q	-	Downside Risk	10	-0.07	11.30	7	-1.18
7	6Q	Beta >1	SR	10	1.85	4.74	-3	27.66
8	6Q	Beta >1	DD	10	2.38	2.50	2	29.33
9	6Q	Beta >1	SR	20	2.06	3.64	-1	24.28
10	6Q	Beta >1 & SR <1	DD	10	1.77	4.53	-1	24.58
11	6Q	-	Accuracy	10	-0.54	78.6	8	-7.77

the 20 names selected for long exposure in each quarter (columns = quarters, rows = rank 1..20) and Figure 9 shows the 20 names selected for short exposure. The visualisation is simple and categorical (one colour block per ticker) so that two structural features of the selection process are apparent: (i) how persistent particular tickers are across consecutive rebalances; and (ii) how much turnover and rank reordering occur from quarter to quarter. The following analysis describes those features and discusses their implications for implementation, risk, and interpretation of the signals.

First, the plots reveal a mixed pattern of persistence and turnover. Several tickers appear in the same side of the book in multiple consecutive quarters (for example, a handful of names in the short-panel — such as PDM and HCSG in the early quarters of the sample — recur repeatedly), which indicates that the selection rule sometimes identifies stocks with lived, stable favourable calibration metrics across the 18-month look-back. At the same time, many slots are filled by names that appear only for a single quarter before being replaced. This alternation between repeat selections and rapid replacement is consistent with a selection signal that is both stable enough to capture persistent cross-sectional patterns and responsive to regime or firm-specific changes. For a practitioner, this mix is useful: persistent names provide the basis for multi-quarter exposures (lower turnover, clearer implementation plan), while rotating names capture freshly emergent opportunities identified by the model.

Second, rank mobility within a quarter-block is informative. The matrix rows reflect the rank ordering of chosen names; in many cases the same name migrates up or down the rank ladder across adjacent quarters rather than disappearing

entirely. This graded movement suggests the model is providing a continuous score (or at least a rankable quantity) rather than a binary “in/out” flag, and it implies that weighting schemes which respect rank (for example, size or linear weighting) will change realised exposures relative to simple equal-weighting. In practice, this behaviour argues for careful consideration of weighting: an aggressive weighting on rank will amplify the impact of a stock that remains ranked first across quarters, increasing concentration risk; conversely, equal weighting will reduce this effect and reduce idiosyncratic exposure from persistent top-ranked names.

Third, comparing the long and short matrices highlights asymmetries in selection dynamics. The long-side grid in Figure 8 shows a different pattern of reuse and rotation than the short-side grid in Figure 9. Concretely, one side may contain more recurring names while the other rotates more rapidly; this asymmetry has two implications. On the modeling side it suggests that the information content of signals for “longable” names differs from that for “shortable” names (for example, some firms consistently show the metric the model rewards while others flip sign more often). On the implementation side it implies that turnover—and therefore transaction costs—will differ across the two sides of the book. Any realistic P&L projection should therefore account for asymmetric execution friction between longs and shorts.

Finally, the plots implicitly inform concentration and diversification. Visual inspection shows periods in which the rank palette is relatively homogeneous (many different names, low persistence) and other periods in which a small set of names repeatedly occupies top ranks. Those concentrated periods raise the risk of idiosyncratic shocks: if the strategy’s gains depend heavily on a small subset of names that remain in the portfolio across rebalances, the strategy becomes exposed to single-name tail risk. Conversely, the more diversified quarters—where each column contains many different tickers—offer stronger mechanical diversification and should be less vulnerable to idiosyncratic failures. This observation reinforces the need for explicit concentration controls (position caps, volatility scaling) when moving from a simulated equally-weighted construction to a live implementation.

In sum, the colour-coded rank matrices convey two central characteristics of the selection mechanism: the model delivers both repeatable (multi-quarter) signals for a subset of names and an ability to adapt and rotate the book when prior winners decay. Those characteristics are desirable for a tradable signal set, but they also create trade-offs—between exploitation of persistent edges and protection against single-name concentration—that need to be explicitly managed in a real implementation.

5.2 Further Discussion

The Impact of Leverage

To explore the strategy’s potential when capital is not a limiting constraint, we analyze the performance of the MDDEW portfolio with a 2x leverage multi-

Ranked Tickers (Top 20) per quarter — long positions											
	Q1 2023	Q2 2023	Q3 2023	Q4 2023	Q1 2024	Q2 2024	Q3 2024	Q4 2024	Q1 2025	Q2 2025	
Rank	VRXT	PRI	PRI	PRI	CAH	PRI	PRI	PKG	FNF	FNF	
1	VRSK	VRSK	VRSK	FSS	PRI	R	NRG	WAB	PKG	NI	
2	PRI	MSI	ABM	MPC	R	MMC	R	TJX	WAB	BK	
3	AFL	AZO	FSS	R	MPC	ALSN	CLH	TRGP	PGR	CLH	
4	AZO	AFL	CAH	CAH	TJX	MPC	TRGP	FNF	PRI	IDCC	
5	FNF	EXEL	NXST	PLXS	IRM	DECK	VRSK	KEX	ETR	SPG	
6	MPC	BRO	MMSI	ABM	CLH	CAH	PGR	PGR	SPG	JWN	
7	CAH	CAH	FCN	MMSI	MMC	NSIT	PKG	MMC	MATX	PRI	
8	AEE	FNF	MLI	NRG	CW	VRSK	COST	PRI	WMB	ETR	
9	BRO	NRG	AFL	AFL	NSIT	PKG	MMC	TSCO	ATR	COST	
10	OKE	OKE	R	IRM	VRTX	TSCO	TDG	R	BK	EXEL	
11	CASY	FCN	OKE	PWR	AFL	EXEL	NSIT	CNX	COST	PGR	
12	CW	CW	CW	CW	WAB	WMB	EXEL	BWXT	CMI	PKG	
13	ETN	HUBB	ATO	AGO	OKE	KEX	TSCO	COST	KEX	ORLY	
14	WMB	ATO	AZO	ATO	ORI	COST	AJG	VRSK	FFIV	ORI	
15	ATO	CLH	WAB	CLH	TSCO	PG	WMB	AJG	MAR	AJG	
16	CLH	MMSI	MSI	OKE	EXEL	ORI	CAH	ORI	EXEL	WMB	
17	MMSI	IRM	CLH	AJG	WMB	ORLY	ORI	WMB	ORLY	VRSK	
18	MMC	MMC	IRM	WAB	AJG	AJG	PG	ORLY	VRSK	LNG	
19	GD	CASY	AJG	ORI	PG	EXC	ORLY	PG	AJG	OMC	
20											

Fig. 8. Color-coded quarter-by-quarter composition of the **long-leg** 20-stock trading book, displayed in a compact rank-ordered grid. Each column corresponds to a quarter (as indicated in the column header) and shows the stocks selected for long positions during that period. Rows indicate the ranking of stocks based on their historical MDD.

Ranked Tickers (Top 20) per quarter — short positions											
Rank	Q1 2023	Q2 2023	Q3 2023	Q4 2023	Q1 2024	Q2 2024	Q3 2024	Q4 2024	Q1 2025	Q2 2025	
1	HCSG	MTN	MDRX	MPW	MNRO	WERN	WERN	WERN	WERN	WERN	
2	PDM	PDM	PDM	HCSG	HCSG	BIIB	HRL	MED	LEG	HELE	
3	HELE	LUV	MPW	TFX	WERN	HRL	PFE	LEG	MED	VC	
4	TFSL	MPW	WERN	WERN	NUS	IRDM	NUS	NUS	PII	LEG	
5	NWL	WU	SRCL	LEG	BAX	NUS	MNRO	CENTA	VC	LEA	
6	WU	HCSG	WU	BAX	BIIB	MNRO	IRDM	SRCL	PFE	MED	
7	HLF	TFSL	KRC	QDEL	PFE	LEG	LEG	XRAY	OMCL	NEOG	
8	LEG	KRC	HCSG	BIIB	MDRX	PFE	BIIB	OMCL	HELE	IRDM	
9	TECH	TECH	TECH	ARE	SRCL	RNST	XRAY	PFE	NUS	OMCL	
10	WERN	NTCT	SASR	MDRX	BGS	CBU	QDEL	HRL	IRDM	HRL	
11	HAIN	MDRX	BAX	SRCL	SASR	AMED	MED	SHEN	XRAY	SAM	
12	MED	BAX	ARE	PDM	CBU	CENTA	HAIN	IRDM	BIIB	THRM	
13	VIAV	LEG	NTCT	SIRI	LEG	QDEL	CENTA	SAM	LEA	XRAY	
14	MDRX	HAIN	LEG	VZ	DG	BAX	TFSL	QDEL	HRL	AES	
15	KRC	VZ	QDEL	HR	WU	WU	CRI	BIIB	MAN	BIIB	
16	VZ	WERN	BIIB	WU	QDEL	FOLD	BBY	HAIN	CENTA	ARI	
17	BIIB	HR	HR	TECH	VZ	MAN	HR	CRI	TFSL	BGS	
18	HR	PFE	PFE	PFE	ADM	MTN	WU	WU	BBY	PFE	
19	MNRO	BIIB	VZ	NTCT	AMED	BGS	BAX	BAX	SAM	MAN	
20	BAX	MD	AMED	ADM	FOLD	ADM	ADM	ADM	ADM	TFSL	

Fig. 9. Color-coded quarter-by-quarter composition of the **short-leg** 20-stock trading book, displayed in a compact rank-ordered grid. Each column corresponds to a quarter (as indicated in the column header) and shows the stocks selected for long positions during that period. Rows indicate the ranking of stocks based on their historical MDD.

plier, assuming a 4% annual cost of capital. This exercise sheds light on how the strategy's core properties—its Sharpe ratio, low market correlation, and drawdown behavior—scale under higher nominal exposure.

With a 2x leverage factor, the MDDEW portfolio's cumulative return at the end of the sample is approximately +47%, with a peak gain near +49%. This represents an increase in absolute nominal returns compared to the non-leveraged MDDEW portfolio's cumulative return of +26.38%. While a larger capital base and increased exposure naturally lead to higher nominal returns, the leveraged portfolio maintains an annualized Sharpe ratio of 2.26. This value, while slightly lower than the non-leveraged portfolio's Sharpe of 2.54, still exceeds the S&P 500 (1.19). This result reinforces the conclusion from our primary analysis: the signals generate a tradable return stream whose volatility profile is attractive enough to sustain favorable risk-adjusted outcomes even under leverage.

The leveraged portfolio's daily return remains weakly correlated with the broad market, with a correlation of -5% to the S&P 500. This low correlation is consistent with the non-leveraged results and demonstrates that the portfolio's gains are still largely idiosyncratic and not simply an amplified bet on systematic market drivers. The time-varying performance, as shown by the rolling Sharpe series, similarly indicates that the leveraged strategy continues to produce robust, positive performance episodes throughout the test period, further validating the stability of the model's signals.

The analysis of drawdowns reveals a key insight into the strategy's risk profile under leverage. The maximum drawdown for the leveraged portfolio increases to 5.33% from the non-leveraged version's 3.04%. This increase in absolute drawdown is an expected consequence of amplifying both positive and negative returns and including borrowing cost. However, the portfolio recovered and resumed its prior trajectory after these spikes, highlighting its ability to withstand adverse market events and resume profitable trading.

Beyond US equity market

The current framework has been validated on a universe of large and mid-cap U.S. equities, but its architecture and feature set are naturally extendable to other markets and instruments. Within equities, one immediate extension is to small-cap stocks in the U.S., where microstructural inefficiencies may amplify the predictive power of the signals. Beyond the U.S., the same methodology can be applied to mature exchanges such as the London Stock Exchange or the Tokyo Stock Exchange, as well as selected European and emerging markets that share structural similarities with U.S. equity markets. Because the feature definitions are rooted in economically interpretable variables rather than market-specific artifacts, the signals are expected to retain predictive power across geographies. A further line of extension is to fixed-income markets. The same predictive framework that currently forecasts directional equity returns can be repurposed to model corporate bond yields, credit spreads, and rate dynamics, thereby opening an avenue into substantially larger markets. In addition, the multi-horizon

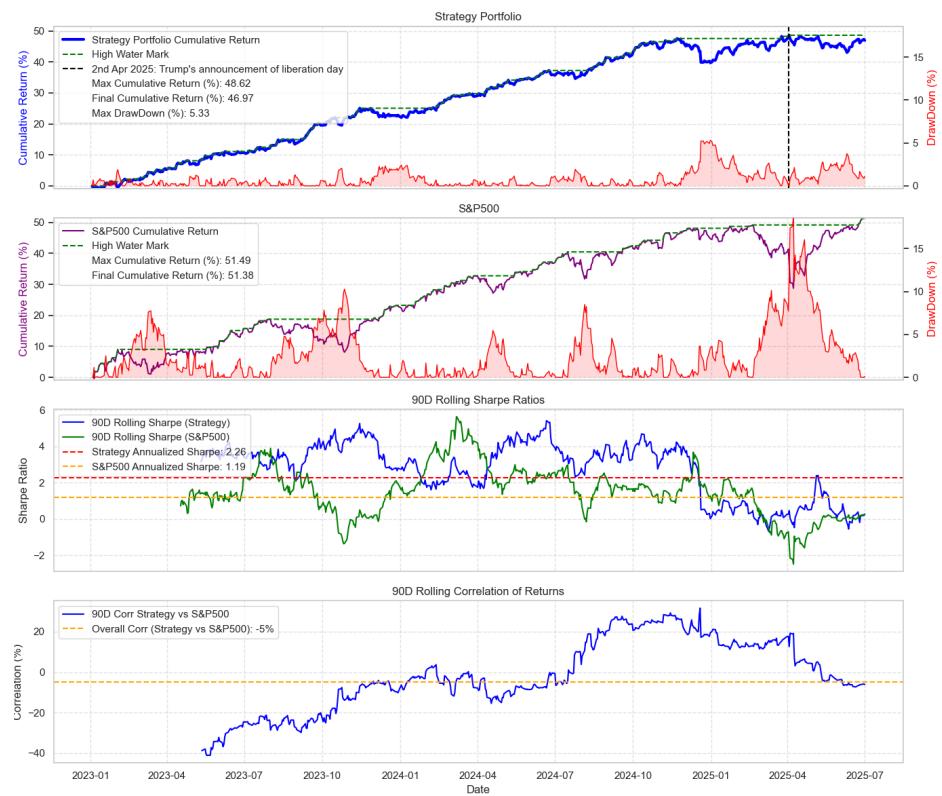


Fig. 10. Out-of-sample performance of the MDD-based, equally-weighted (MDDEW) portfolio with **2x leverage and 4% cost of capital**, benchmarked against a passive S&P 500 buy-and-hold.

design of the signals—where predictions are generated across several future time steps—lends itself to derivative markets. In particular, the short-dated horizons align with the maturities of futures contracts and short-lived options, including zero-day-to-expiry (0DTE) options that have recently gained significant liquidity. This adaptability suggests that the proposed framework is not confined to equity returns alone but can evolve into a more general platform for multi-asset signal generation and risk forecasting across global capital markets.

6 Conclusion

In this paper, we presented an end-to-end evaluation of an AI-driven trading framework developed at *Increase Alpha*. The model utilizes deep learning architectures—primarily feed-forward and recurrent networks—combined with expert-selected financial features to generate directional trading signals for 814 U.S. equities. The system’s architecture emphasizes operational efficiency, avoiding large-scale transformer models in favor of a leaner, interpretable design that delivers consistent performance with minimal computational overhead.

We demonstrated the predictive strength and economic viability of the model across multiple axes. Our accuracy analysis revealed that both long and short signals significantly outperform random baselines, as evidenced by statistical tests including p-values and confidence intervals. Furthermore, the signal’s effectiveness persisted across various holding periods and market regimes, including the volatile period following January 2025.

To translate these predictions into actionable trades, we conducted an extensive grid search across profit-taking, stop-loss, and holding-period parameters using Azure’s scalable cloud infrastructure. The resulting configuration enabled a robust simulation framework that quantified the strategy’s profitability and risk metrics. Notably, the framework achieved superior cumulative returns and Sharpe ratios compared to both buy-and-hold strategies and macro benchmarks like the S&P 500, while maintaining significantly lower drawdowns.

Stress testing confirmed that the model remains stable and adapts under adverse market conditions. By modulating activity in low-confidence regimes, the framework avoids overtrading and limits exposure to high-risk periods without explicit correlation inputs. This emergent behavior highlights the system’s ability to internalize uncertainty and act accordingly.

Ultimately, our findings suggest that local inefficiencies in financial markets can be systematically harvested using targeted deep learning tools grounded in domain expertise. The *Increase Alpha* platform exemplifies how rigorous engineering, paired with scalable infrastructure and disciplined evaluation, can produce a reliable and uncorrelated source of alpha. We believe this work opens the door to further applications of interpretable and computationally efficient AI in modern asset management.

References

1. Bai, Y., Gao, Y., Wan, R., Zhang, S., Song, R.: A review of reinforcement learning in financial applications. *Annual Review of Statistics and Its Application* **12**(1), 209–232 (2025)
2. Cohen, G.: Algorithmic trading and financial forecasting using advanced artificial intelligence methodologies. *Mathematics* **10**(18), 3302 (2022)
3. Huang, B., Huan, Y., Xu, L.D., Zheng, L., Zou, Z.: Automated trading systems statistical and machine learning methods and hardware implementation: a survey. *Enterprise Information Systems* **13**(1), 132–144 (2019)
4. Lahmiri, S., Bekiros, S.: Intelligent forecasting with machine learning trading systems in chaotic intraday bitcoin market. *Chaos, Solitons & Fractals* **133**, 109641 (2020)
5. Lopez-Lira, A., Tang, Y.: Can chatgpt forecast stock price movements? return predictability and large language models. arXiv preprint arXiv:2304.07619 (2023)
6. McGroarty, F., Booth, A., Gerding, E., Chinthalapati, V.R.: High frequency trading strategies, market fragility and price spikes: an agent based model perspective. *Annals of Operations Research* **282**(1), 217–244 (2019)
7. Rahimikia, E., Drinkall, F.: Re (visiting) large language models in finance. Available at SSRN (2024)
8. Rundo, F., Trenta, F., di Stallo, A.L., Battiatto, S.: Grid trading system robot (gtsbot): A novel mathematical algorithm for trading fx market. *Applied Sciences* **9**(9), 1796 (2019)
9. Sarkar, S.K., Vafa, K.: Lookahead bias in pretrained language models. Available at SSRN (2024)
10. Simsek, A.I.: Improving the performance of stock price prediction: A comparative study of random forest, xgboost, and stacked generalization approaches. In: *Revolutionizing the Global Stock Market: Harnessing Blockchain for Enhanced Adaptability*, pp. 83–99. IGI Global (2024)
11. Song, J., Cheng, Q., Bai, X., Jiang, W., Su, G.: Lstm-based deep learning model for financial market stock price prediction. *Journal of Economic Theory and Business Management* **1**(2), 43–50 (2024)
12. Vuletić, M., Prenzel, F., Cucuringu, M.: Fin-gan: Forecasting and classifying financial time series via generative adversarial networks. *Quantitative Finance* **24**(2), 175–199 (2024)