How Reliable Are Reward-Objective Conclusions in Financial Reinforcement Learning?

Anonymous authors
Paper under double-blind review

Abstract

Financial reinforcement learning (RL) studies use materially different reward objectives, while separate work shows that repeated RL training can yield substantially different outcomes across random seeds. Neither observation is treated as new here. Instead, this study asks how reliable conclusions attributed to reward choice are when evaluated across independent stochastic retraining and different subsets of observed runs. Six reward families—portfolio-value change, log wealth growth, differential Sharpe ratio, rolling Sortino ratio, negative maximum drawdown, and Sharpe minus conditional-value-at-risk penalty—are evaluated with Proximal Policy Optimization (PPO) on a controlled single-asset task using 10 matched seeds per reward. Omnibus tests do not detect a significant reward effect on annualized return, Sharpe ratio, Sortino ratio, or maximum drawdown; reward-associated differences are detected for turnover and long exposure, although most policies remain predominantly long. Counterfactual single-seed comparisons produce different apparent reward leaders. Robust summaries reinforce that caution: replacing the mean with the median or interquartile mean (IQM) substantially compresses the apparent performance separation among rewards, and matched-seed bootstrap resampling does not yield a stable IQM leader. Exhaustive enumeration of matched seed subsets further shows that the non-significant return and Sharpe conclusions are stable, whereas the long-exposure conclusion becomes increasingly stable as more runs are included: 99.2% of seven-seed subsets and all subsets of eight or more seeds yield a significant Friedman reward effect. Turnover stabilizes more slowly. Reward rankings also become progressively more similar to the 10-seed reference as additional runs are included, while winner claims remain fragile when leading estimates are close. No significant performance reward effect is detected across three temporal windows or with a second actor–critic algorithm. The contribution is therefore a reward-specific conclusion-reliability analysis: the results support reporting full run distributions and directly examining whether conclusions attributed to reward choice remain stable under stochastic retraining.

1 Introduction

Reward design is a central modeling choice in reinforcement learning (RL) because it determines the objective optimized by the learning process. In financial RL, reward functions span raw or net profit, log wealth growth, Sharpe- and Sortino-based criteria, drawdown penalties, tail-risk penalties, utilities, and combinations of these quantities (Moody and Saffell, 2001; Leem and Kim, 2020; Jung and Oh, 2025; Makni, 2026; Sahu and Bhandari, 2026). Prior studies already show that changing the reward can alter learned trading behavior or reported performance. Reward sensitivity itself is therefore not the novelty claim of this paper.

A separate empirical-RL literature shows that learned outcomes can vary substantially across independent stochastic training runs (Henderson et al., 2018; Agarwal et al., 2021; Patterson et al., 2024). Recent financial evidence makes the same issue explicit: repeated training under a fixed configuration can produce economically different outcomes, and favorable-run selection can distort reported performance (Grądzki, 2026). Random-seed sensitivity is therefore also not a new observation here.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Taken together, these two established findings create an attribution problem that is less directly characterized in reward-comparison studies. If reward A and reward B have different sample means, how stable is that apparent separation relative to the variation obtained by retraining each reward? If only one or a few runs are observed, would the same reward ordering or "winner" be inferred? These questions concern the reliability of the *conclusion about reward choice*, rather than the existence of reward sensitivity or seed sensitivity by themselves.

This study uses *reward-conclusion reliability* as a descriptive notion: a reward-level conclusion is more reliable when it remains similar across repeated stochastic training realizations and across reasonable subsets of those runs. It is not introduced as a new statistical estimator or universal criterion. The primary research question is:

> *When researchers attribute an observed difference to reward-objective choice, how reliably does that conclusion persist across independent stochastic training runs?*

Three research questions (RQs) organize the analysis. **Research question 1 (RQ1)** asks what performance and behavioral differences are observed among six reward objectives under matched Proximal Policy Optimization (PPO) training. **Research question 2 (RQ2)** asks whether those reward-level differences are stable relative to within-reward seed variability, including what a single-seed or small-seed comparison would have concluded. **Research question 3 (RQ3)** asks whether the resulting performance-level pattern persists across temporal windows, state-information conditions, and a second actor–critic algorithm.

The contributions are deliberately narrow. First, the study combines reward comparison and repeated-run uncertainty in one matched design and evaluates reward attribution through omnibus matched tests, mean, median, and interquartile mean (IQM) summaries, single-seed counterfactuals, exhaustive seed-subset ranking stability, and an *inference-stability* diagnostic that asks how often the same significance conclusion is recovered from every possible subset of the available seeds. Second, it shows empirically that apparent mean performance gaps shrink substantially under robust summaries and that performance-level reward conclusions remain non-significant across nearly all seed subsets, whereas the long-exposure conclusion becomes progressively more stable across run subsets as additional runs are included; turnover is intermediate and requires more runs before significance becomes common. Third, it separates performance conclusions from policy behavior and tests the performance-level pattern across time and a second actor–critic algorithm. The paper does not propose a new reward taxonomy or RL algorithm, does not claim a universal minimum seed count, and does not infer trading superiority from the observed policies.

2 Related Work

This section separates prior work on financial reward design from work on stochastic evaluation, then identifies the reward-attribution question that connects them.

2.1 Reward objectives in financial RL

Direct reinforcement methods have long optimized financial criteria beyond raw profit. Moody and Saffell (2001) introduced direct trading optimization and the differential Sharpe ratio (DSR), while Du et al. (2009) compared Q-learning and recurrent reinforcement learning under interval profit, Sharpe, and differential-Sharpe value functions. Modern financial-RL frameworks extend the setting to end-to-end portfolio allocation and benchmark environments (Jiang et al., 2017; Liu et al., 2022).

Reward comparison itself is not new. Leem and Kim (2020) explicitly evaluated profit, Sharpe, and Sortino rewards; each approach was trained 10 times, after which the top five models were selected for summary reporting. Jung and Oh (2025) compared five reward designs across several asset classes and three deep reinforcement learning (DRL) algorithms: PPO, Soft Actor-Critic (SAC), and Twin Delayed Deep Deterministic Policy Gradient (TD3); the documented main training setup used a fixed seed. Makni (2026) evaluated a grid of risk-aware PPO reward configurations with five seeds per configuration. Sahu and Bhandari (2026) examined several reward specifications together with multiple DRL algorithms and market-efficiency regimes. These studies establish both the diversity and practical importance of financial reward design.

2.2 Stochastic evaluation in RL

General RL evaluation research has repeatedly shown that point estimates from a small number of runs can be unreliable (Henderson et al., 2018; Agarwal et al., 2021; Patterson et al., 2024). The issue includes random initialization, action sampling, optimization order, environment randomness, and implementation details. In this paper, the broader concept of training stochasticity is operationalized concretely as *random-seed variability across independent training runs* under an otherwise identical configuration.

In financial DRL, Grądzki (2026) directly studies stochastic-training and model-selection uncertainty across repeated seeds. Its central experiments hold the reward definition fixed within a diagnostic configuration and characterize the distribution generated by retraining. The closest reward-comparison papers and the stochastic-evaluation literature therefore establish the two ingredients of the present study separately: rewards can matter, and stochastic retraining can matter. The distinction here is the reward-specific attribution question—whether a conclusion assigned to changing the reward remains stable once stochastic retraining is treated as part of the comparison. This distinction is operationalized through paired reward–seed measurements, counterfactual single-seed conclusions, exhaustive seed-subset ranking stability, and seed-subset inference stability. Appendix B gives a study-by-study comparison.

3 Problem Formulation and Reward Objectives

This section defines the common portfolio-control environment and the six reward-objective conditions compared within that fixed setting.

3.1 Portfolio-control environment

The environment uses daily returns of the Standard & Poor's Depositary Receipt (SPDR) Dow Jones Industrial Average exchange-traded fund (ETF) Trust, traded under ticker symbol DIA. At time t, the compact state is

$$s_t = [R_{t-1}, \ldots, R_{t-5}, p_{t-1}, W_t, D_t], \tag{1}$$

where the first five terms are lagged market returns, $p_{t-1} \in \{-1, 0, 1\}$ is the previous position, W_t is current wealth, and D_t is current drawdown. The action $a_t \in \{-1, 0, 1\}$ denotes the desired short, cash, or long position. A transaction cost of $c = 0.001$ is charged per unit change in position:

$$C_t = c|p_t - p_{t-1}|, \qquad R_t^{\text{net}} = p_t R_t^{mkt} - C_t. \tag{2}$$

Wealth and drawdown update as

$$W_{t+1} = W_t(1 + R_{t+1}^{\text{net}}), \quad D_t = 1 - \frac{W_t}{\max_{0 \le j \le t} W_j}, \tag{3}$$

with $W_0 = 1$, $p_0 = 0$, and $D_0 = 0$. The market trajectory is fixed by historical data; actions, positions, wealth, drawdown, and reward histories are endogenous to each learned policy.

3.2 Reward-objective conditions

Table 1 summarizes the six reward families. Differential Sharpe ratio (DSR) follows the direct-trading formulation of Moody and Saffell (2001); profit and Sortino-based rewards are compared by Leem and Kim (2020); log-return and Sharpe–conditional-value-at-risk (CVaR) objectives are represented in Sahu and Bhandari (2026); and drawdown-sensitive reward shaping is examined by Makni (2026). The negative maximum-drawdown (MDD) condition used here is a representative drawdown objective. These conditions are literature-grounded operationalizations under a common simulator, not a new taxonomy or verbatim reproductions of every cited formulation. Path-dependent reward histories are reset independently at train, validation, and test boundaries.

For DSR, A_t and B_t are exponentially updated first and second moments of net return, with ΔA_t and ΔB_t denoting their one-step deviations. In the Sortino condition, R_f denotes the risk-free rate and is set to zero.

Table 1: Reward-objective conditions. $\eta = 0.01$ for DSR, Sortino uses a 20-day window, and Sharpe–CVaR uses a 252-day window with $\alpha = 0.05$ and $\lambda = 0.5$.

Objective	Training reward
Profit	$r_t = W_{t+1} - W_t$
Log return	$r_t = \log(1 + \text{clip}(R_t^{\text{net}}, -0.5, 5))$
DSR	$r_t = \frac{B_{t-1}\Delta A_t - \frac{1}{2}A_{t-1}\Delta B_t}{(B_{t-1} - A_{t-1}^2)^{3/2}}$
Sortino	Rolling $(\bar{R} - R_f)/\text{downside-deviation}(R)$, $R_f = 0$
Neg. MDD	$r_t = -\,\text{MDD}_t$
Sharpe–CVaR	$r_t = \text{Sharpe}_t - \lambda \lvert\, \text{CVaR}_{\alpha,t} \,\rvert$

4 Experimental Design

The experimental design holds the learning setup fixed across reward conditions and evaluates reward conclusions using matched repeated training runs.

4.1 Data, training, and controls

Daily open, high, low, close, and volume (OHLCV) data were retrieved from Yahoo Finance using `yfinance` (Aroussi, 2024); adjusted close is used to compute returns. The primary split is 2010–2020 for training, 2021–2022 for validation, and 2023–2025 for testing. Feature standardization is fit on training data only and then applied unchanged to validation and test data.

PPO (Schulman et al., 2017) is implemented directly in PyTorch using a shared 64×64 tanh actor–critic network, $\gamma = 0.99$, generalized advantage estimation (GAE) with $\lambda = 0.95$, clipping $\epsilon = 0.2$, entropy coefficient 0.01, value coefficient 0.5, Adam learning rate 3×10^{-4}, gradient clipping at 0.5, and 10 PPO update epochs. Training runs for 250 epochs. Actions are sampled from a categorical policy during training and selected by deterministic argmax during validation/test. Advantages are normalized within each update. This reduces sensitivity to raw reward scale in the policy update, although differences in raw reward and critic targets remain a limitation.

Ten seed values (0–9) initialize Python, NumPy, and PyTorch pseudo-random number generators. The same seed identifiers are matched across all reward conditions. Checkpoints are evaluated every 10 epochs and selected using the same external validation criterion—annualized net portfolio return—rather than the internal training reward. Test data are never used for checkpoint selection. Across reward conditions, the market data, split, state, action space, architecture, optimizer, PPO hyperparameters, training budget, cost model, seed set, checkpoint rule, and evaluation metrics are fixed. Appendix A records the full experimental configuration.

4.2 Evaluation and statistical analysis

External test metrics are annualized return, Sharpe ratio, Sortino ratio, maximum drawdown (MDD), mean turnover, and short/cash/long action shares. Buy-and-hold and cash are simple economic references. Reward conditions are compared with a Friedman test using seed as the matched block; Kendall's W is reported as an omnibus effect size. Holm-corrected paired Wilcoxon tests are used for follow-up behavioral comparisons.

Reward-conclusion reliability is evaluated directly from the repeated-run design rather than through a ratio of reward-mean variance to run variance. Such a raw ratio is sample-size dependent because averaging over n seeds mechanically reduces the variance of reward means; it is therefore not used as a threshold for attribution. Instead, three diagnostics are used. First, a single-seed counterfactual asks which reward would appear highest if only one matched seed were observed. Second, for every subset size $k \in \{1, \ldots, 10\}$, all matched seed subsets are enumerated and the resulting reward ordering is compared with the 10-seed sample ordering using Spearman correlation. Third, for $k \in \{3, \ldots, 10\}$, the Friedman test is recomputed for every one of the $\binom{10}{k}$ matched seed subsets. The fraction of subsets yielding $p < 0.05$ is reported as *inference stability*: how often the same reward-effect conclusion would be reached from a subset of the available stochastic runs. These subset analyses are diagnostics of conclusion stability, not additional hypothesis tests used to manufacture significance.

Table 2: Primary PPO test results over 10 seeds. Return is annualized; values are mean ± sample standard deviation where shown. The buy-and-hold row is deterministic for this test window.

Reward	Return	Sharpe	Turnover	Long
Profit	$15.24 \pm 1.91\%$	1.134	0.0149	0.993
Log return	$14.25 \pm 5.26\%$	1.094	0.0129	0.894
DSR	$15.10 \pm 2.68\%$	1.120	0.0125	0.976
Sortino	$12.38 \pm 5.42\%$	0.976	0.0285	0.805
Neg. MDD	$15.05 \pm 1.13\%$	1.115	0.0059	0.999
Sharpe–CVaR	$14.04 \pm 2.43\%$	1.064	0.0181	0.929
Buy-and-hold	15.26%	1.129	0.0013	1.000

The 10-seed ordering and full-sample p-values are reference estimates, not ground truth or a universal seed requirement.

Following robust RL evaluation practice (Agarwal et al., 2021), annual return and Sharpe are also summarized by the median and the interquartile mean (IQM), defined here as the exact empirical mean over the central 50% of the run distribution (25% trimmed from each tail, with fractional boundary weights for 10 observations). To characterize finite-run uncertainty without breaking the matched design, 50,000 bootstrap resamples jointly resample seed identifiers across all reward conditions. IQM confidence intervals and the fraction of bootstrap resamples in which each reward has the highest IQM are reported descriptively. These leader frequencies are not posterior probabilities that a reward is truly best and are not used as hypothesis tests.

The repeated-run outcome can be viewed conceptually as $y_{r,s} = \mu + \alpha_r + \epsilon_{r,s}$, where α_r denotes the reward-associated component and $\epsilon_{r,s}$ the run-specific stochastic deviation. With one aggregate test outcome per reward–seed cell, a richer reward-by-seed interaction cannot be separated cleanly from run-level residual variation without stronger modeling assumptions. The analysis therefore favors matched nonparametric tests, robust summaries, and exhaustive subset diagnostics over an over-parameterized mixed-effects model.

5 Results

The results are organized around the primary reward comparison, stability of reward conclusions across stochastic runs, robustness checks, and policy-behavior context.

5.1 Primary PPO experiment

Table 2 summarizes the primary test results. Profit, DSR, and negative-MDD rewards have similar mean annualized returns near the buy-and-hold reference (15.26%). Log-return, Sharpe–CVaR, and especially Sortino conditions show lower means, but dispersion across seeds is substantial. Appendix B reports the complete omnibus statistics. The Friedman tests do not reject equal reward-condition distributions for annualized return ($p = 0.430$), Sharpe ($p = 0.364$), Sortino ($p = 0.468$), or MDD ($p = 0.822$). Omnibus reward-associated differences are detected for turnover ($p = 0.017$) and long exposure ($p = 0.0014$), but the post-hoc separation is sparse: after Holm correction, only the log-return versus negative-MDD long-exposure comparison remains significant ($p = 0.029$). Thus the primary evidence does not support a broad claim that all reward objectives learn distinct behaviors; it supports a narrower conclusion that reward-associated behavioral differences are detectable even though aggregate performance distributions overlap strongly.

Figure 1 shows why mean-only interpretation is fragile: seed-level annual-return distributions overlap heavily across reward conditions. A small number of reward–seed combinations produce large departures, including a log-return run that converges almost entirely to cash. Robust summaries quantify the effect of those tails. Across the six rewards, the max–min annual-return spread is 2.86 percentage points for the means, but only 0.56 points for medians and 1.00 point for IQMs; the Sharpe spread contracts from 0.157 by mean to 0.039 by median and 0.063 by IQM. Profit has the highest point estimate under all three summaries, but matched bootstrap resampling does not support a stable winner interpretation: for annual-return IQM, Profit is highest in 46.4% of resamples, log return in 41.7%, and DSR in 9.5%; Sharpe IQM is similarly split at 45.5%, 44.0%, and 8.1%. These

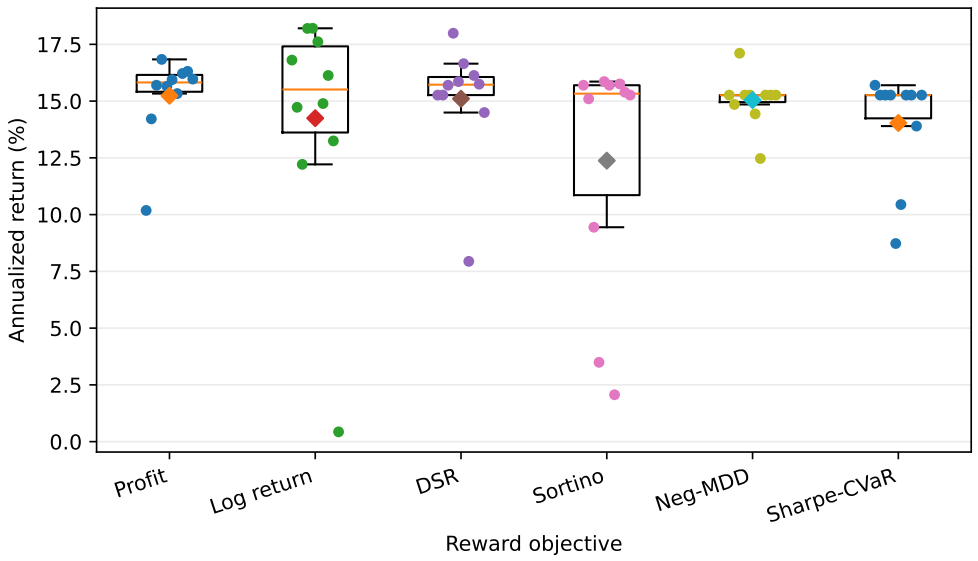

Figure 1: Out-of-sample annualized return for each of 10 independent PPO training seeds under the six reward objectives. Individual points are retained rather than selecting favorable runs.

are descriptive resampling frequencies, not probabilities of true superiority. Appendix E reports the full robust summaries.

5.2 Reliability of reward conclusions across stochastic training

The single-seed counterfactual first shows how unstable a winner-style claim can be. The reward with the highest annual return is log return for seeds 0 and 3–5, negative MDD for seed 2, profit for seeds 6–9, and a DSR/Sharpe–CVaR tie for seed 1. Only four of ten individual-seed comparisons identify Profit, the reward with the highest 10-seed sample mean. This does not establish Profit as a true superior reward; it shows that the apparent leader depends on which stochastic realization is observed.

Figure 2 asks a stronger question: whether the *statistical conclusion* about a reward effect survives changes in which runs are observed. For annual return and Sharpe, fewer than 4% of all subsets at any size $k = 3, \ldots, 8$ yield a significant Friedman test, and none of the nine-seed or full 10-seed analyses do. Thus the non-significant performance conclusion is highly stable. Long exposure behaves differently: the reward effect is significant in 47.6% of five-seed subsets, 80.5% of six-seed subsets, 99.2% of seven-seed subsets, and every subset of eight or more seeds. Turnover stabilizes more slowly: significance occurs in 27.8% of five-seed subsets, 55.6% of eight-seed subsets, 70% of nine-seed subsets, and in the full 10-seed analysis. This distinguishes a behavioral conclusion that is stable across most larger subsets (long exposure), a more subset-dependent behavioral conclusion (turnover), and consistently unsupported performance conclusions.

Increasing the number of seeds also stabilizes the overall reward ordering (Figure 4 in Appendix B). Mean Spearman agreement with the 10-seed annual-return ordering increases from 0.23 at $k = 1$ to 0.59 at $k = 5$, 0.81 at $k = 8$, and 0.90 at $k = 9$; Sharpe is similar. Yet the identity of the single highest sample-mean reward remains fragile because the leading means are close. Leave-one-seed-out analysis reinforces this point: Profit remains the annual-return leader in eight of ten nine-seed subsets, while omitting seed 6 makes DSR highest and omitting seed 9 makes log return highest. The experiment therefore supports distributional and conclusion-stability reporting rather than a universal minimum seed count.

5.3 Robustness across time, state, and algorithm

The temporal experiment repeats the six-reward, 10-seed PPO design in three expanding training windows with disjoint test periods: 2017–2019, 2020–2022, and 2023–2025 (Appendix F). Annual-

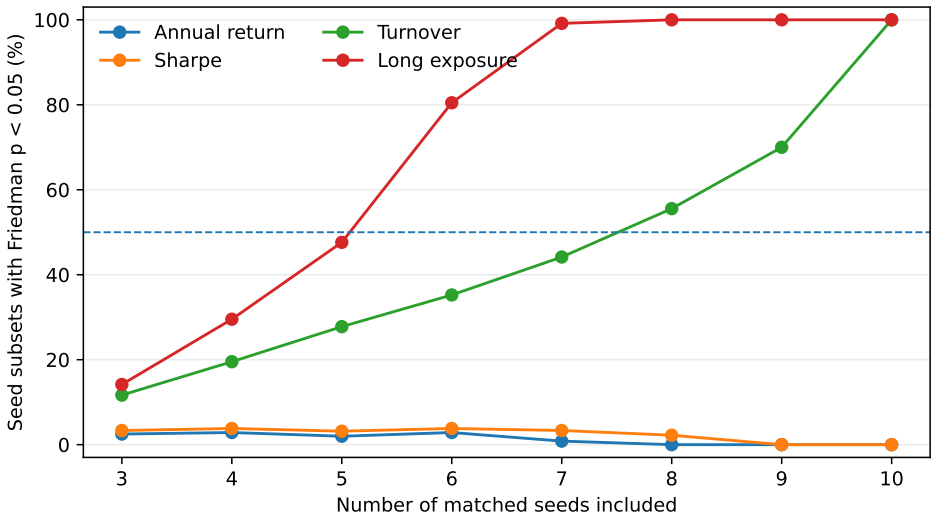

Figure 2: Inference stability across all matched seed subsets. For each subset size k, the figure reports the percentage of $\binom{10}{k}$ subsets for which the Friedman test across six rewards yields $p < 0.05$. The analysis summarizes conclusion stability; subset tests are not treated as independent discoveries.

return Friedman p-values are 0.315, 0.0736, and 0.430; Sharpe p-values are 0.380, 0.0931, and 0.364. Thus no performance reward effect reaches $\alpha = 0.05$ in any window. Omnibus turnover and long-exposure differences are detected in all three windows. The magnitude of outcomes changes materially with the market period, but the performance-level conclusion remains the same across the three tests: the experiment does not establish a reward-dependent performance ordering.

State ablation evaluates four nested states: five return lags alone; plus previous position; plus wealth; and plus drawdown (Appendix G). Across reward–seed blocks, no overall state main effect is detected for annual return ($p = 0.574$), Sharpe ($p = 0.616$), turnover ($p = 0.138$), or long exposure ($p = 0.579$). This does not establish state irrelevance; it indicates that the tested state additions did not produce a detectable global shift under the fixed protocol.

A second experiment replaces PPO with a direct synchronous Advantage Actor-Critic (A2C) implementation related to the actor–critic family of Mnih et al. (2016), while holding the market data, state, rewards, seed set, costs, checkpoint rule, and external metrics fixed (Appendix H). The interaction/training-trajectory budget is matched, although the optimizer-step budget is not; this is therefore an algorithm robustness check rather than a tuned algorithm contest. Under A2C, none of the six reward comparisons reaches $p < 0.05$: annual return $p = 0.0648$, Sharpe $p = 0.0790$, turnover $p = 0.459$, and long exposure $p = 0.926$. The absence of a statistically established performance reward effect therefore persists under this second actor–critic implementation, whereas the PPO behavioral significance does not.

5.4 Policy behavior and benchmark context

Most PPO policies converge toward a predominantly long strategy. Across the 60 primary reward–seed policies, 53 (88.3%) are long on at least 95% of test observations and 48 (80%) are long on at least 99%. Median long exposure exceeds 99% for every reward condition. Figure 5 in Appendix I shows the full exposure distribution. The policies below 95% long exposure have much lower average annual return than the near-long group, but this association is descriptive and does not establish causality.

This behavior is an important limitation and diagnostic. The DIA buy-and-hold baseline earns 15.26% annually with Sharpe 1.129 over the primary test window, and no reward condition has a higher mean return. Similar aggregate performance among many successful runs therefore partly reflects convergence toward the same high-exposure solution rather than six distinct active trading strategies. Policy behavior should consequently be reported alongside financial metrics. A separate

XGBoost (Chen and Guestrin, 2016) reference using only the five exogenous return lags produces weak out-of-sample next-day prediction and a mostly-cash policy; because it lacks the endogenous portfolio state supplied to RL, it is contextual rather than a direct algorithm comparison (Appendix J).

6 Discussion

Reward sensitivity and seed sensitivity are both established in prior financial-RL work. The contribution here is narrower: when reward choice is the controlled factor, the analysis asks whether the conclusion attributed to that factor remains stable across stochastic retraining.

In the primary experiment, reward means differ, but omnibus tests do not establish performance separation. Median and IQM summaries compress the apparent gaps, the matched bootstrap does not identify a stable performance leader, and single-seed comparisons produce different apparent leaders. Exhaustive subset analysis sharpens that result: non-significance for annual return and Sharpe is stable across essentially all subsets, long-exposure significance becomes stable across nearly all seven-seed subsets and all larger subsets, and turnover remains more subset-dependent. These findings do not imply a universal seed requirement; they show that the reliability of a reward conclusion should be demonstrated rather than inferred from the mere use of multiple seeds.

Behavioral interpretation also requires care. PPO shows omnibus differences in turnover and long exposure, but most policies are predominantly long and post-hoc separation is limited. Similar financial performance can therefore coincide with a common exposure pattern, while a few reward–seed combinations produce large departures. Reporting exposure, turnover, and benchmark distance alongside return and risk metrics makes that distinction visible.

This study follows broader empirical-RL guidance on repeated-run evaluation (Henderson et al., 2018; Agarwal et al., 2021; Patterson et al., 2024) without claiming a new general principle. For reward studies, the practical implication is to retain all runs, report within-condition distributions, and test whether substantive conclusions remain stable across run subsets and relevant experimental contexts. The numerical findings remain specific to the evaluated financial setting.

6.1 Scope of claims and threats to validity

The study isolates reward definitions under shared hyperparameters rather than separately tuning each reward; advantage normalization reduces, but does not remove, reward-scale differences in critic targets. DSR, Sortino, MDD, and Sharpe–CVaR are history-dependent, and the tested state does not establish full Markov sufficiency. The simulator uses a fixed 10-basis-point transaction cost but omits slippage, liquidity, market impact, latency, and endogenous price response. The primary task is one liquid U.S. ETF with a discrete short/cash/long action space; temporal windows add regime variation but not multi-asset or multi-market validation. Most successful policies are close to buy-and-hold, limiting behavioral diversity. Finally, 10 seeds enable exhaustive subset diagnostics but do not establish a universal seed requirement; subset significance fractions are stability diagnostics, not independent tests or power guarantees.

7 Conclusion

Prior work separately establishes reward sensitivity and stochastic-training variability in financial RL. This study evaluates their intersection as a reward-attribution problem. In the controlled six-reward PPO experiment, performance reward effects are not statistically established, robust summaries reduce apparent mean separation, and no stable performance winner emerges from single-seed or bootstrap analyses. Seed-subset analysis distinguishes stable performance non-significance from increasingly stable long-exposure differences and more subset-dependent turnover differences. Performance non-significance also persists across temporal windows and a second actor–critic implementation. Thus, under this controlled setting, performance-based reward conclusions can be sensitive to stochastic training, whereas some behavioral differences become increasingly stable as additional runs are considered. The resulting recommendation is reward-specific: report full within-reward run distributions and test whether conclusions attributed to reward choice remain stable under stochastic retraining.

AI USE STATEMENT

Generative artificial intelligence (AI) tools were used for language editing, readability, literature-search assistance, figure preparation, and document formatting. AI-assisted suggestions were checked against the cited literature, retained experiment outputs, and implementation metadata. Research questions, experimental design, implementation choices, statistical procedures, interpretation of results, and final claims remain the responsibility of the authors. Generative AI was not used as an autonomous substitute for running the reported experiments or for accepting unverified empirical results.

ETHICS STATEMENT

The study uses historical public market data and does not involve human participants or personal data. Results are presented as empirical findings about a controlled research environment, not as investment advice or claims of deployable trading superiority.

REPRODUCIBILITY STATEMENT

The manuscript specifies the data source, preprocessing and calendar splits, state and action definitions, reward equations, model architecture, hyperparameters, seed set, checkpoint-selection rule, external metrics, and statistical procedures. Per-seed outputs and experiment metadata were retained for every reported experiment. Executable code, retained per-seed results, experiment metadata, and analysis scripts are provided in the anonymous supplementary package. The package also documents data acquisition and the commands needed to regenerate the reported analyses from retained outputs.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Ran Aroussi. *yfinance: Download market data from Yahoo! Finance's API*, 2024. Software package, `https://github.com/ranaroussi/yfinance`.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi:10.1145/2939672.2939785.

Xin Du, Jinjian Zhai, and Koupin Lv. Algorithm trading using Q-learning and recurrent reinforcement learning. Stanford University CS229 Machine Learning final project report, 2009.

Przemysław Grądzki. Unstable gains: Multiplicity-aware evaluation of financial deep reinforcement learning. *The Journal of Finance and Data Science*, 12:100205, 2026. doi:10.1016/j.jfds.2026.100205.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi:10.1609/aaai.v32i1.11694.

Zhengyao Jiang, Dixing Xu, and Jinjun Liang. A deep reinforcement learning framework for the financial portfolio management problem. arXiv:1706.10059, 2017.

Nak Hyun Jung and Taeyeon Oh. Factor-based deep reinforcement learning for asset allocation: Comparative analysis of static and dynamic beta reward designs. *PLOS ONE*, 20(12):e0332779, 2025. doi:10.1371/journal.pone.0332779.

JoonBum Leem and Ha Young Kim. Action-specialized expert ensemble trading system with extended discrete action space using deep reinforcement learning. *PLOS ONE*, 15(7):e0236178, 2020. doi:10.1371/journal.pone.0236178.

Xiao-Yang Liu, Ziyi Xia, Jingyang Rui, Jiechao Gao, Hongyang Yang, Ming Zhu, Christina Dan Wang, Zhaoran Wang, and Jian Guo. FinRL-Meta: Market environments and benchmarks for data-driven financial reinforcement learning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 35, 2022. doi:10.52202/068431-0134.

Salwa Makni. Risk-aware reward shaping for portfolio optimization using proximal policy optimization. Research Square preprint, 2026. doi:10.21203/rs.3.rs-9847041/v1.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1928–1937, 2016.

John E. Moody and Matthew Saffell. Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks*, 12(4):875–889, 2001. doi:10.1109/72.935097.

Andrew Patterson, Samuel Neumann, Martha White, and Adam White. Empirical design in reinforcement learning. *Journal of Machine Learning Research*, 25(318):1–63, 2024.

Hitesh Kumar Sahu and Avishek Bhandari. Benchmarking deep reinforcement learning and classical models for portfolio optimization across market efficiency regimes. *Discover Artificial Intelligence*, 6:866, 2026. doi:10.1007/s44163-026-01869-x.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv:1707.06347, 2017.

APPENDIX

The following supplementary sections provide implementation details, additional statistical results, robustness analyses, and benchmark context referenced from the main paper.

A DETAILED EXPERIMENTAL CONFIGURATION

The primary configuration is collected here so that the main text can emphasize the controlled comparison rather than implementation detail. Table 3 records the settings shared across the six PPO reward conditions.

Table 3: Primary PPO configuration.

Component	**Setting**
Asset	DIA, daily adjusted-close returns
Train / validation / test	2010–2020 / 2021–2022 / 2023–2025
State	5 lagged returns, previous position, wealth, drawdown
Actions	$\{-1, 0, +1\}$ short/cash/long
Cost	$0.001\|p_t - p_{t-1}\|$
Network	shared multilayer perceptron (MLP), 64–64, tanh; actor + critic heads
γ / GAE λ	0.99 / 0.95
PPO clip / update epochs	0.2 / 10
Entropy / value coefficient	0.01 / 0.5
Learning rate	3×10^{-4}, Adam
Training	250 epochs; evaluation every 10 epochs
Seeds	0–9, matched across rewards
Checkpoint	maximum validation annualized net return
Test policy	deterministic argmax

B CLOSEST RELATED WORK

Table 4 makes the positioning explicit. "Not reported" means that the examined article text did not establish a repeated-seed distribution analysis; it should not be read as proof that no such run existed.

Table 4: Comparison with closely related reward-comparison and stochastic-evaluation studies.

Study	Reward/objective variation	Repeated runs	Run variability retained/analyzed	Direct reward-vs-seed comparison	Primary purpose
Moody & Saffell (2001)	Profit/risk-adjusted criteria incl. DSR	Multiple trials in reported experiments	Partial distributional reporting	No	Direct RL for trading and financial objectives
Du et al. (2009)	Profit, Sharpe, DSR	Not systematic	Not reported	No	Q-learning vs. recurrent reinforcement learning under value functions
Leem & Kim (2020)	Profit, Sharpe, Sortino	10 trainings	Top five selected for mean/standard error	No	Ensemble trading and reward sensitivity
Jung & Oh (2025)	Sharpe, Sortino, three factor rewards	Main setup seed 42	No seed distribution	No	Factor-aware reward design across assets/algorithms
Makni (2026)	16 risk-aware reward configurations	5 seeds/config.	Mean/standard deviation and bootstrap reporting	No	Risk-aware PPO reward shaping
Sahu & Bhandari (2026)	Log, Sharpe, Sharpe–CVaR	Not reported in examined DRL setup	Not reported	No	Market-efficiency regimes and algorithm/reward performance
Gradzki (2026)	Reward fixed within each diagnostic experiment	20 seeds	Yes; stochastic/multiplicity analysis	No	Stochastic optimization and selection uncertainty
This study	**Six reward families**	**10 matched seeds/reward**	**All runs retained**	**Yes**	**Reliability of reward-attribution under repeated training**

The comparison therefore motivates the narrower question examined here: whether conclusions attributed specifically to reward choice remain stable under matched stochastic retraining.

C Primary Statistical Detail

Table 5 reports the complete primary omnibus tests. Figure 3 complements the annual-return distribution in Figure 1 by showing the corresponding seed-level Sharpe distribution.

Table 5: Primary PPO Friedman tests across six reward conditions using 10 matched seeds.

Metric	χ^2_F	p	**Kendall** W
Annual return	4.882	0.4305	0.0976
Sharpe	5.444	0.3641	0.1089
Sortino	4.586	0.4685	0.0917
Maximum drawdown	2.190	0.8223	0.0438
Turnover	13.742	0.0173	0.2748
Long exposure	19.682	0.0014	0.3936

Table 6: Inference stability across exhaustive matched seed subsets. Entries are the percentage of subsets for which the Friedman test across six rewards yields $p < 0.05$.

Metric	$k = 3$	$k = 5$	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$
Annual return	2.5	2.0	2.9	0.8	0.0	0.0	0.0
Sharpe	3.3	3.2	3.8	3.3	2.2	0.0	0.0
Turnover	11.7	27.8	35.2	44.2	55.6	70.0	100.0
Long exposure	14.2	47.6	80.5	99.2	100.0	100.0	100.0

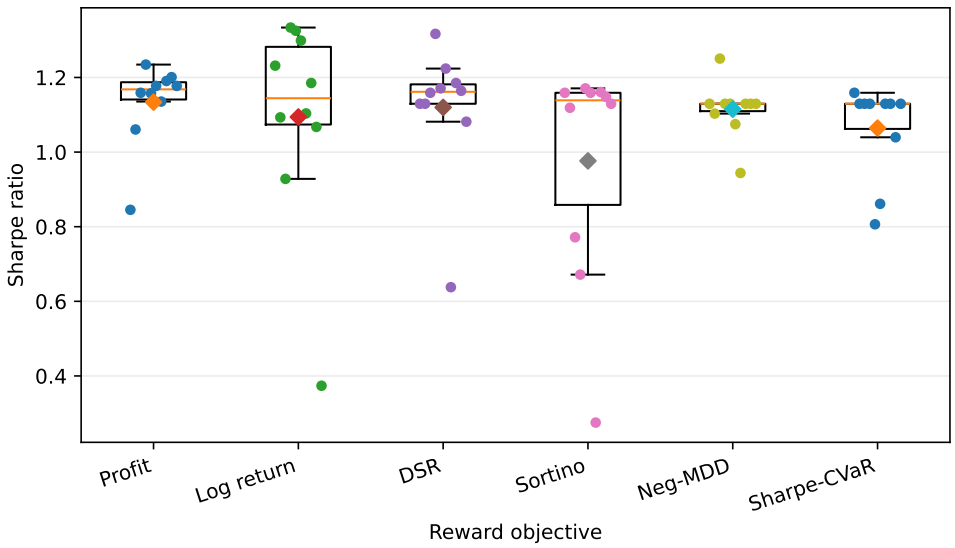

Figure 3: Sharpe ratio across the 10 primary PPO seeds for each reward objective. The overlap parallels the annual-return result in Figure 1.

Together, Tables 5 and 6 and Figure 3 show that the primary performance distributions overlap substantially even though some behavioral conclusions are detectable. The next appendix examines how those conclusions change as the observed seed set changes.

D Single-Seed and Seed-Count Diagnostics

Figure 4 summarizes how reward-ranking agreement with the 10-seed reference changes as additional matched seeds are included. The accompanying table then shows the apparent annual-return leader when each seed is considered alone.

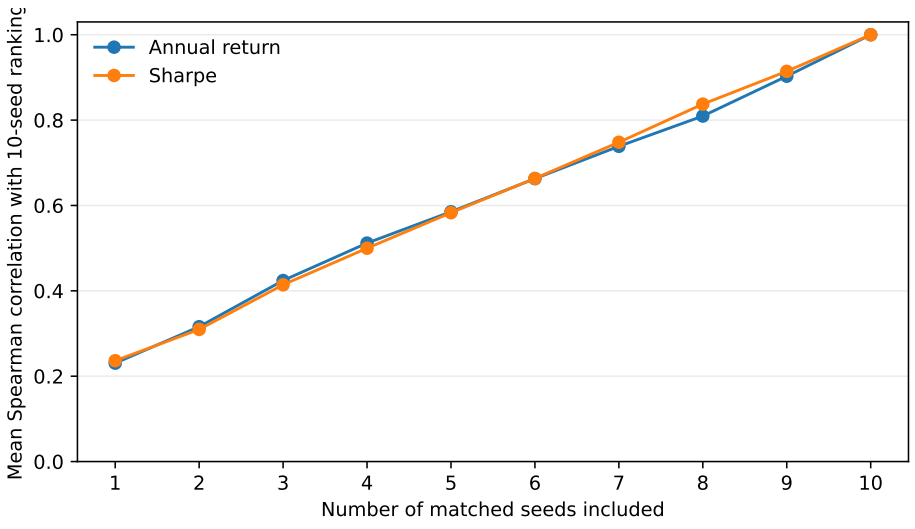

Figure 4: Mean Spearman agreement between reward rankings computed from all matched subsets of k seeds and the full 10-seed sample ordering. The full-sample ordering is a reference estimate rather than ground truth.

Table 7: Reward with the highest annualized return when each seed is considered alone. "Tie" denotes numerically equal test outcomes.

Seed	Apparent leader	Annual return
0	Log return	16.81%
1	DSR / Sharpe–CVaR (tie)	15.26%
2	Neg. MDD	17.11%
3	Log return	18.20%
4	Log return	18.21%
5	Log return	17.61%
6	Profit	15.33%
7	Profit	16.21%
8	Profit	16.31%
9	Profit	15.97%

For annual return, leave-one-seed-out analysis retains Profit as the highest sample-mean reward in eight of ten nine-seed subsets; omitting seed 6 makes DSR highest, while omitting seed 9 makes log return highest. Spearman agreement with the full 10-seed annual-return ordering ranges from 0.60 to 1.00 across these leave-one-out subsets (0.657 to 1.00 for Sharpe). This identifies influential runs without deleting them from the reported analysis.

E Robust Performance Summaries

The interquartile mean (IQM) is the exact empirical 25% trimmed mean used in the main text. With 10 runs, fractional boundary weighting retains the central 50% of empirical probability mass. Table 8 shows that mean performance gaps are materially larger than median/IQM gaps for the reward conditions most affected by occasional poor runs.

Table 8: Mean, median, and IQM performance summaries in the primary PPO experiment. Annual return is shown in percent.

Reward	Annual return (%)			Sharpe		
	Mean	Median	IQM	Mean	Median	IQM
Profit	15.24	15.82	15.81	1.134	1.168	1.167
Log return	14.25	15.51	15.60	1.094	1.144	1.159
DSR	15.10	15.72	15.65	1.120	1.162	1.156
Sortino	12.38	15.33	14.80	0.976	1.139	1.104
Neg. MDD	15.05	15.26	15.22	1.115	1.129	1.127
Sharpe–CVaR	14.04	15.26	15.13	1.064	1.129	1.121

Table 9: Matched-seed bootstrap frequency with which each reward has the highest IQM across 50,000 resamples. Frequencies summarize resampling instability and are not probabilities of true superiority.

Reward	Annual-return IQM	Sharpe IQM
Profit	46.4%	45.5%
Log return	41.7%	44.0%
DSR	9.5%	8.1%
Sortino	0.8%	0.9%
Neg. MDD	0.9%	0.9%
Sharpe–CVaR	0.8%	0.6%

F Temporal Robustness

The temporal check repeats the six-reward comparison across three disjoint test windows. Table 10 reports the omnibus reward-effect p-values used in the main-text robustness discussion.

Table 10: Temporal-window robustness. Entries are Friedman p-values across the six reward conditions.

Window	Annual return	Sharpe	Turnover	Long exposure
Window 1 (2017–2019)	0.315	0.380	0.00083	0.00213
Window 2 (2020–2022)	0.0736	0.0931	0.0127	0.00101
Window 3 (2023–2025)	0.430	0.364	0.0173	0.00143

G State-Information Ablation

The state ablation adds portfolio information incrementally to the five return lags. Table 11 summarizes pooled outcomes for the four nested representations before the corresponding omnibus state tests are reported.

Table 11: PPO results pooled across rewards and seeds for four nested state representations.

State	Return	Sharpe	Turnover	Long
S1: 5 market lags	14.27%	1.047	0.0263	0.933
S2: + previous position	15.23%	1.131	0.0119	0.993
S3: + wealth	14.71%	1.103	0.0186	0.951
S4: + drawdown	14.34%	1.084	0.0155	0.933

The corresponding state-main-effect Friedman p-values are 0.574 (annual return), 0.616 (Sharpe), 0.761 (Sortino), 0.858 (MDD), 0.138 (turnover), and 0.579 (long exposure). S2 is descriptively less variable in several summaries but also converges especially strongly toward passive-long behavior; no causal stability claim is made.

H A2C Algorithm Robustness

The second actor–critic experiment holds the market data, S4 state, reward objectives, seed set, costs, checkpoint rule, and external metrics fixed while replacing PPO with A2C. Table 12 summarizes its per-reward results.

Table 12: A2C results over 10 seeds per reward. This is a controlled robustness check, not a tuned PPO–A2C leaderboard.

Reward	Return	Sharpe	Turnover	Long
Profit	$13.92 \pm 5.05\%$	1.121	0.0240	0.827
Log return	$16.27 \pm 0.37\%$	1.198	0.0088	0.996
DSR	$15.74 \pm 2.62\%$	1.182	0.0177	0.925
Sortino	$13.44 \pm 5.73\%$	0.938	0.0278	0.874
Neg. MDD	$9.03 \pm 10.69\%$	0.693	0.0802	0.775
Sharpe–CVaR	$12.44 \pm 9.52\%$	0.919	0.0125	0.896

A2C Friedman p-values are 0.0648 (annual return), 0.0790 (Sharpe), 0.3088 (Sortino), 0.0655 (MDD), 0.4589 (turnover), and 0.9264 (long exposure). Thus the second actor–critic implementation does not reproduce PPO's behavioral significance and also does not establish a performance reward effect under the shared protocol.

I Behavior Relative to Buy-and-Hold

Figure 5 and Table 13 show how strongly the primary PPO policies concentrate near full long exposure and where the comparatively rare departures occur.

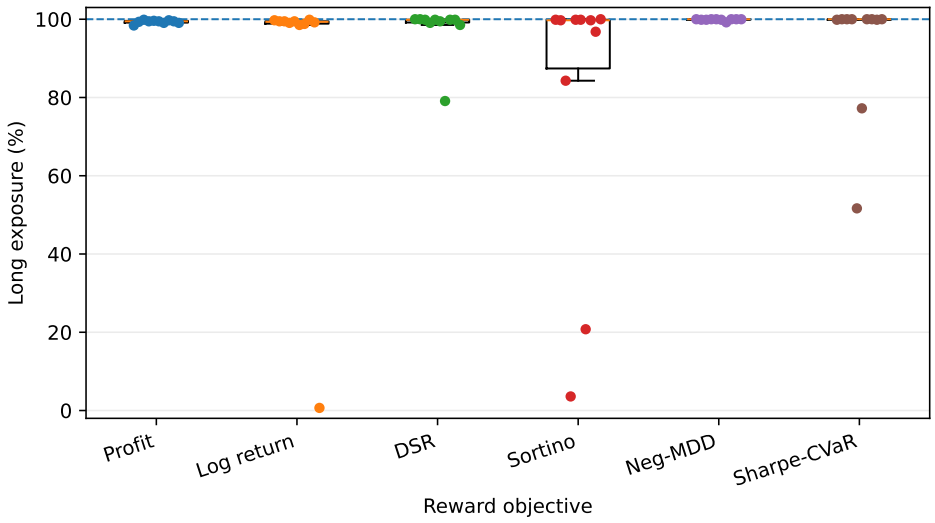

Figure 5: Long-exposure distribution across reward conditions in the primary PPO experiment. Most runs cluster near one, with a small number of large departures.

Table 13: Mean and median long exposure by reward in the primary PPO experiment.

Reward	**Mean long**	**Median long**	**Mean return**
Profit	0.9935	0.9947	15.24%
Log return	0.8943	0.9933	14.25%
DSR	0.9756	0.9987	15.10%
Sortino	0.8045	0.9973	12.38%
Neg. MDD	0.9988	1.0000	15.05%
Sharpe–CVaR	0.9286	1.0000	14.04%

J Supervised-Learning Context

A deterministic XGBoost reference (Chen and Guestrin, 2016) uses only the five lagged returns to predict next-day DIA return. A validation-selected threshold $\tau = 0.002$ maps predictions to short/cash/long actions under the same transaction-cost rule. On the primary test window, annualized return is 1.05%, Sharpe is 0.233, MDD is 5.46%, turnover is 0.125, and the policy is 92.0% cash. Prediction correlation is 0.063 and directional accuracy is 52.9%. Because the supervised model omits the endogenous position, wealth, and drawdown variables available to the RL policy, these values provide context for the limited predictive content of the exogenous five-lag signal rather than a direct algorithm ranking.