Regime-Switching Reinforcement Learning for Portfolio Allocation in Pairs Trading

Tsvetelina Ilieva[1]
Supervisors: Frans Oliehoek[1]**, Fenghui Yu**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Tsvetelina Ilieva
Final project course: CSE3000 Research Project
Thesis committee: Frans Oliehoek, Fenghui Yu, Neil Yorke-Smith

Abstract

Pairs trading is a well-studied strategy in statistical arbitrage. By using asset pairs with correlated changes in their historical prices, the strategy profits from exploiting the non-permanent divergence of their price relationship, assuming that this relationship will revert to its long-term equilibrium. However, the dynamics of this relationship may vary over time, as the spread, which measures the deviation between the prices of paired assets, can exhibit different levels of volatility and mean-reverting behavior under different market conditions. In this paper, we propose a regime-aware reinforcement learning framework for portfolio optimization in pairs trading. We model the spread between assets and characterize its behavior using statistical features capturing its relative position to historical equilibrium, its volatility, and the strength of its mean-reverting behavior. These features are used within a Hidden Markov Model to infer latent market regimes, which represent distinct states of spread dynamics over time. The inferred regimes are incorporated into the state representation of a reinforcement learning agent, which learns to dynamically allocate capital across pairs. We evaluate the proposed approach against a regime-agnostic reinforcement learning benchmark and a classical z-score threshold strategy. In a controlled simulation study, the regime-aware agent achieves a mean Sharpe ratio of 1.354 versus 0.738 for the baseline on V/MA (ΔSharpe = +0.616) and 1.183 versus 0.564 on V/JKHY (ΔSharpe = +0.619), consistent across 10 training seeds. On real out-of-sample data from 2023 to 2026, the regime agent achieves Sharpe ratios of 0.567 and 0.609 on V/MA and V/JKHY respectively, outperforming the baseline in both cases.

1 Introduction

Pairs trading is a statistical arbitrage strategy that exploits the difference in the price between two historically related stocks, or also called the spread. However, different market regimes may require distinct allocation and hedging strategies. Standard trading models often fail to capture these abrupt transitions, leading to suboptimal performance during chaotic or shifting market phases.

While existing literature has extensively explored pairs trading and reinforcement learning (RL) models for portfolio optimization, these approaches typically assume a stationary environment. For example, Gatev et al. [3] explore statistical arbitrage strategies based on the historical price spread between pairs of co-integrated assets. In the RL domain, Moody and Saffell [10] explore learning to trade via direct reinforcement, laying foundational RL methods for dynamic portfolio allocation. Building upon this, Jiang et al. [7] present a deep RL framework for financial portfolio management, dynamically reallocating portfolios while considering transaction costs. Furthermore, using a Hidden Markov Model (HMM) to identify hidden states representing the market regimes, Kim et al. showed that an adaptive portfolio optimizing for Sharpe ratio can yield superior portfolio performance across a variety of different asset classes [8]. More recently, Nixon Raj [12] proposes a regime-aware Proximal Policy Optimization (PPO) framework that conditions agents on HMM-derived regime probabilities for long-horizon portfolio optimization, though the approach operates on annual data across broad asset classes and does not exploit the cointegration structure specific to pairs trading. Despite these advancements, existing regime-aware RL frameworks focus on broad portfolio allocation across asset classes and do not exploit the cointegration structure that underpins pairs trading. Modeling the spread as a mean-reverting process with regime-specific dynamics and integrating that structure directly into the agent's observation signal remains unexplored.

To address this gap, our primary research question is: *Does a regime-aware reinforcement learning agent outperform regime-neutral alternatives in portfolio allocation for pairs trading?*

To answer this, we decompose the question into three sub-questions:

- Does a regime-aware PPO agent outperform a regime-neutral PPO baseline trained on the same data?
- Does a regime-aware PPO agent outperform a classical z-score threshold strategy?
- Does the performance advantage of the regime-aware agent observed in simulation persist when both agents are trained and tested on real historical data?

The main contribution of this paper is the introduction of a Regime-Switching Reinforcement Learning framework designed specifically for pairs trading. The framework comprises a complete simulation pipeline combining HMM regime detection and Ornstein-Uhlenbeck process calibration, a regime-adjusted z-score observation signal that makes regime information directly actionable without increasing model complexity, and an online calibration mechanism that adapts the per-regime equilibrium estimate over time. We evaluate the framework in both a controlled simulation study and on real out-of-sample data, demonstrating that the regime-aware agent consistently outperforms both the regime-neutral baseline and the classical z-score threshold strategy across both settings and both pairs.

2 Methodology

This section describes the complete framework for training and evaluating regime-aware reinforcement learning agents for pairs trading. The framework operates in two settings. In the primary setting, a regime-switching simulation pipeline generates synthetic spread episodes from a calibrated Hidden Markov Model and Ornstein-Uhlenbeck process, providing sufficient training data to overcome the scarcity of historical observations. In the secondary setting, both agents are trained directly on real historical spread windows to assess whether the regime advantage transfers beyond the simulation. In both settings the PPO agent, observation space, reward function, and hyperparameters remain identical; only the data source changes. The following pipeline is executed for each pair: pair selection and cointegration testing, OU calibration, HMM regime detection, per-regime parameter estimation, and either synthetic episode generation or historical sliding window construction.

2.1 Data and Pair Selection

Daily adjusted prices are obtained from Yahoo Finance [17]. Adjusted prices account for dividends, ensuring the price series reflect total economical returns rather than raw market quotations. Pairs are selected from within the same economic sector, ensuring that both assets share a common fundamental driver that motivates the existence of a long-run equilibrium relationship. For a specific sector, we take all stocks in that sector that are part of S&P 500 and for each two of them we run statistical tests to filter out unsuitable pairs. A fixed formation window is used for all pair selection filters. Only data within this window is used for screening; we hold out the test period entirely and never inspect it during pair selection or model calibration. All statistical tests are run in log-prices.

For each candidate pair (A,B), the Pearson correlation between the daily log returns of the two assets is computed over the formation window. Pairs with correlation below 0.70

are discarded. A correlation threshold of 0.70 ensures that surviving pairs share sufficient common variation.

The pairs surviving the correlation filter are tested for cointegration using the Engle-Granger two-step procedure [1]. First, the log-price of asset A is regressed on the log-price of asset B via the Ordinary Least Squares (OLS) method to obtain the hedge ratio $\hat{\beta}$:

$$p_t^A = \alpha + \hat{\beta} \cdot p_t^B + \varepsilon_t$$

The residual $z_t = p_t^A - \hat{\beta} \cdot p_t^B$ defines the spread, which removes the common price trend shared by the two assets. Second, an Augmented Dickey-Fuller (ADF) test is applied to z_t:

$$\Delta z_t = \gamma z_{t-1} + \sum_{j=1}^{k} \delta_j \Delta z_{t-j} + u_t$$

Under $H_0 : \gamma = 0$, the spread is a random walk with no tendency to revert to its mean, implying the pair is not cointegrated. Pairs for which H_0 cannot be rejected at the 5% significance level are discarded. A full description of the ADF test is provided in [11].

For each pair surviving the cointegration filter, we estimate the half-life of mean reversion. The half-life is a measure of how quickly the spread returns to its long-run equilibrium after a deviation. Formally, for an Ornstein-Uhlenbeck (OU) process with mean-reversion speed κ, the half-life is the expected time for the spread to cover half the distance back to its mean μ from any initial displacement:

$$\mathrm{HL} = \frac{\ln 2}{\kappa}$$

More information on the OU model is provided in 2.2. Intuitively, if the spread is currently one standard deviation above its mean, after one half-life it is expected to be only half a standard deviation above the mean; after two half-lives, one quarter; and so on. A short half-life indicates rapid mean reversion - the spread corrects quickly and trading opportunities close fast. A long half-life indicates slow reversion with the spread drifting for months before returning to equilibrium. Therefore, we keep only pair with a half-life between 1 day and 60 days. We put this filter as a practical tradability constraint rather than a statistical one. A half-life below one day implies that the spread reverts faster than the daily observation frequency, making the signal unactionable at daily resolution. A half-life above 60 days implies holding periods of several months, locking up capital for extended periods and reducing the number of complete mean-reversion cycles available within a fixed training episode.

2.2 Spread Model: Ornstein-Uhlenbeck Process

The spread $z_t = p_t^A - \hat{\beta} \cdot p_t^B$ is modelled as a continuous-time Ornstein-Unhlenbeck process:

$$dz_t = \kappa(\mu - z_t)dt + \sigma dW_t$$

where $\kappa > 0$ is the mean-reversion speed, μ is the long-run equilibrium level, $\sigma > 0$ is the diffusion coefficient, and W_t is a standard Brownian motion.
We use the discrete version of the process:

$$z_t = z_{t-1}\, e^{-\kappa \Delta t} + \mu\big(1 - e^{-\kappa \Delta t}\big) + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left(0,\, \frac{\sigma^2}{2\kappa}\left(1 - e^{-2\kappa \Delta t}\right)\right)$$ [1]

[1] Exact discretization; see [2], Eq. 71-76.

where Δt represents one trading day.

2.3 Regime Detection: Hidden Markov Model

Real equity spreads do not exhibit constant dynamics throughout time. Periods of low volatility and fast mean reversion alternate with periods of high volatility and slow reversion, often triggered by macroeconomic events or changes in sector fundamentals. Modeling time-varying market conditions as a Markov-switching process is well established in the time-series literature as a way to capture nonlinear dynamics and regime-dependent behavior [5]. To capture this behavior, we use a three-state Gaussian Hidden Markov Model (HMM) [14] to identify latent market regimes in the calibration spread. The three states are interpreted as *stable* ($r = 0$), *neutral* ($r = 1$), and *crisis* ($r = 2$).

Rather than feeding raw spread values to the HMM, we compute three features from the calibration spread that each capture a different aspect of the current market dynamics. The first feature is the rolling z-score of the spread, computed over a window of 20 days:

$$z_t = \frac{s_t - \bar{s}_{t,20}}{\hat{\sigma}_{t,20}}$$

where $\bar{s}_{t,20}$ and $\hat{\sigma}_{t,20}$ are the sample mean and standard deviation of the spread over the past 20 days. This feature measures how far the current spread is from its recent history in units of its own volatility. For regime detection, the z-score is informative because we expect that different regimes exhibit characteristically different deviation patterns: stable regimes tend to produce small, rapidly mean-reverting z-scores, while crisis regimes produce larger and more persistent deviations as the cointegration relationship temporarily weakens. The second feature we use is the rolling log-annualized volatility, defined as $\log(\hat{\sigma}_t \cdot \sqrt{252})$ where $\hat{\sigma}_t$ is the rolling standard deviation of daily spread changes, again over 20 days. The third feature is the logit-transformed reversion rate, introduced in this work as an empirical measure of the strength of mean reversion in the current window. It measures the fraction of days on which the spread moves back towards its rolling mean after a deviation exceeding one standard deviation:

$$\text{logit}(\hat{r}_t) = \log \frac{\hat{r}_t}{1 - \hat{r}_t}, \qquad \hat{r}_t = \frac{\#\{j \le t : |z_j| > 1 \text{ and } (z_{j+1} - z_j)(\bar{z} - z_j) > 0\}}{\#\{j \le t : |z_j| > 1\}}$$

This feature distinguishes regimes that are merely volatile from regimes where mean reversion has genuinely broken down. All three features are standardized to zero mean and unit variance before being passed to the HMM.

A three-state Gaussian HMM with full covariance matrices is fitted to the standardized feature sequence using the Baum-Welch algorithm [14], as implemented in `hmmlearn` [6]. To avoid convergence to poor local optima, the model is re-fitted multiple times with different random initializations and the solution with the highest log-likelihood is kept. The three hidden states are then ordered canonically by their median log-volatility feature value, so that state 0 always corresponds to the lowest-volatility (stable) regime and state 2 to the highest-volatility (crisis) regime. This ordering ensures consistency across different pairs and calibration windows.

The fitted HMM provides three outputs used downstream: the 3×3 transition matrix $\mathbf{A}$, where $A_{rr'} = P(s_t = r' \mid s_{t-1} = r)$, the per-state emission parameters (means and covariances), and the hard state sequence $\hat{r}_1, \ldots, \hat{r}_T$ inferred via the Viterbi algorithm [14]. The hard sequence is used to label each calibration day with a regime, enabling the per-regime parameter estimation described in the next section.

2.4 Per-Regime Parameter Estimation and Spread Simulation

Once the HMM has labeled each calibration day with a regime, we estimate separate OU parameters for each regime from the labeled sub-samples of the spread. For each regime $r \in \{0, 1, 2\}$ we collect all consecutive pairs (z_{t-1}, z_t) where both days carry label r and estimate:

- μ_r: the conditional equilibrium mean, taken as the sample mean of the spread over all days labeled r.
- κ_r: the per-regime mean-reversion speed, estimated from an in-regime AR(1) regression $\Delta z_t = a + c \cdot z_{t-1}$, giving $\kappa_r = -c/\Delta t$, clipped to $[0.1, 50]$. If the fitted coefficient implies no mean reversion ($c \geq 0$), the global κ is used as a fallback.
- σ_r: the per-regime volatility, estimated from the mean squared residuals of the in-regime OU transitions after conditioning on κ_r and μ_r:

$$\sigma_r = \sqrt{\frac{\overline{(z_t - \hat{z}_t)^2}}{(1 - e^{-2\kappa_r \Delta t})/(2\kappa_r)}}$$

  where $\hat{z}_t = z_{t-1} e^{-\kappa_r \Delta t} + \mu_r (1 - e^{-\kappa_r \Delta t})$ is the conditional OU mean.

The parameters $(\mu_r, \kappa_r, \sigma_r)$ and the transition matrix $\mathbf{A}$ are saved and used both for simulation and for computing the regime-adjusted observation.

We use the calibrated parameters to generate a large number of synthetic spread episodes, each covering one trading year. Each episode begins from the last observed spread value in the calibration window. At each step t, the active regime r_t is drawn from the Markov chain:

$$P(r_t = r' \mid r_{t-1} = r) = A_{rr'}$$

and the next spread value is drawn from the per-regime OU transition, using the discrete version from Section 2.2.

2.5 Output and Integration with the PPO Agent

The generator produces, for each pair, a dataset of synthetic spread episodes stored as two arrays: `spreads` of shape (N, T) containing spread values and `regimes` of shape (N, T) containing integer regime labels $r_t \in \{0, 1, 2\}$ at each step. Additionally, the fitted HMM, the feature scaler, the hedge ratio $\hat{\beta}$, and all calibrated OU parameters $(\kappa_r, \mu_r, \sigma_r, \mathbf{A})$ are saved to disk. We need this information at test time to compute the regime-adjusted observation and to run the HMM forward filter on real data.

The dataset is split into training and evaluation sets using a random permutation, so that both splits contain a representative mix of regime compositions. The evaluation split is used only to monitor agent performance during training and save the best checkpoint; it is never used to select hyperparameters.

2.6 Observation Space

We train and compare two agents - a regime-aware and regime-agnostic one used as a baseline. Both receive a three-dimensional observation.

The **baseline agent** receives:

$$o_t^{\text{base}} = \left[z_t^{\text{roll}},\ w_{t-1},\ \hat{v}_t\right]$$

where z_t^{roll} is the rolling z-score of the spread over a window of W days, w_{t-1} is the current portfolio weight, and $\hat{v}_t$ is the annualized realized volatility computed over the same window. Both z_t^{roll} and $\hat{v}_t$ are clipped to prevent outliers from destabilizing training.

The **regime-aware agent** receives:

$$o_t^{\text{regime}} = \left[z_t^{\text{adj}},\ w_{t-1},\ \hat{v}_t\right]$$

where the rolling z-score is replaced by the *regime-adjusted z-score*:

$$z_t^{\text{adj}} = \frac{z_t - \tilde{\mu}_{r_t}}{\sigma_{r_t}/\sqrt{2\kappa_{r_t}}}$$

Here $\tilde{\mu}_{r_t}$ is the current online estimate of the equilibrium mean for the active regime r_t, and $\sigma_{r_t}/\sqrt{2\kappa_{r_t}}$ is the theoretical stationary standard deviation of the OU process in that regime, used as a normalization constant. This z-score centers on the regime-specific equilibrium rather than the rolling mean, providing a more informative trading signal. The baseline agent uses a rolling mean that may be misled around regime transitions while the regime-aware agent uses the deviations from the correct target based on the current regime.

The estimate $\tilde{\mu}_{r_t}$ is updated online via an exponential moving average each time regime r_t is visited:

$$\tilde{\mu}_{r_t} \ \leftarrow\ \alpha \cdot z_t + (1-\alpha) \cdot \tilde{\mu}_{r_t}$$

where α is the calibration smoothing factor. At the start of each training episode $\tilde{\mu}_r$ is initialized to the calibrated value μ_r for all regimes. This online update allows the agent to track gradual equilibrium drift without requiring full re-calibration, partially addressing a limitation that historically estimated parameters may not generalize to future market conditions.

2.7 Action Space and Reward Function

The action at each step is a continuous portfolio weight $w_t \in [-1, 1]$, representing the fraction of capital allocated to the spread position. A positive weight corresponds to a long position in the spread (long asset A, short asset B scaled by $\hat{\beta}$); a negative weight corresponds to a short position. The continuous action space allows the agent to express graded conviction rather than a binary enter/exit decision.

The reward at step t is the net profit-and-loss of the position held over that step, net of a transaction cost:

$$r_t = w_{t-1} \cdot \Delta z_t - c \cdot |w_t - w_{t-1}|$$

where $\Delta z_t = z_t - z_{t-1}$ is the one-day spread change and $c > 0$ is the transaction cost coefficient per unit of spread position change. During evaluation this reward is converted to dollar P&L by scaling with the current capital and the beta-adjusted notional:

$$\text{Dollar PnL}_t = \frac{C_t \cdot r_t}{1 + \hat{\beta}}$$

Capital is updated each step as $C_{t+1} = C_t + \text{Dollar PnL}_t$, starting from $C_0 = \$10{,}000$. The full derivation of this conversion is provided in Appendix A.

2.8 PPO Algorithm and Training

Both agents are trained using Proximal Policy Optimisation [16] as implemented in Stable-Baselines3 [15] within a custom Gymnasium environment. PPO is an on-policy actor-critic algorithm that optimizes a clipped surrogate objective to prevent excessively large policy updates. It has been widely adopted in financial RL applications [13, 9, 12] and is used here for its stability and compatibility with continuous action spaces.

Both agents use a multi-layer perceptron (MLP) policy with two hidden layers and Tanh activations. Since both agents receive a three-dimensional observation, the network architectures are identical. Multiple parallel environments are used to collect rollouts simultaneously, with each rollout covering one complete trading-year episode. An evaluation callback monitors agent performance on the held-out evaluation split throughout training and saves the checkpoint with the highest mean episode reward as the best model. All specific hyperparameter values are reported in the experimental setup chapter.

2.9 Real Historical Training

The simulation pipeline in Sections 2.2–2.5 and usage together with the PPO agent helps us asses whether regime awareness helps when there is a clean pair relationship between the assets. However, there is much more variability in real data. To assess whether the regime advantage transfers to real market conditions, we train both agents directly on sliding windows of historical spread data. The PPO agent, observation space, reward function, and all hyperparameters remain identical to the simulated setting; only the data source changes.

Rolling hedge ratio. For each training episode starting at day t, the hedge ratio is re-estimated from the 252 trading days immediately preceding the episode:

$$\hat{\beta}_t = \mathrm{OLS}(\log p^A_{t-252:t},\ \log p^B_{t-252:t})$$

The episode spread is then $s_\tau = \log p^A_\tau - \hat{\beta}_t \cdot \log p^B_\tau$ for $\tau \in [t, t+252]$. This ensures that the spread within each episode is computed with a locally appropriate hedge ratio, maintaining approximate stationarity throughout the window.

Causal regime labelling. In the simulation pipeline, we use Viterbi decoding to label the calibration spread for per-regime parameter estimation - an offline procedure that introduces no look-ahead bias since training episodes use fresh Markov chain regime sequences, not the Viterbi labels. For real historical training the situation is different: regime labels are computed from the actAt each day t, the label $\hat{r}_t$ must therefore be computed using only observations up to t. We replace Viterbi with the HMM forward filter [14], which propagates the posterior regime distribution one step at a time using only past observations, giving a causal estimate $\hat{r}_t = \arg\max_r \alpha_t(r)$ at each day.

Sliding window episodes. Historical training episodes are constructed by sliding a $T = 252$-day window over the full historical spread with stride S days, giving approximately $(T_{\text{hist}} - T)/S$ episodes. Each episode is a pair $(\mathbf{s}, \mathbf{r})$ where $\mathbf{s} \in \mathbb{R}^{252}$ is the rolling-beta spread slice and $r \in \{0, 1, 2\}^{252}$ contains the causal forward filter labels. The training and evaluation sets are split chronologically: the first 80% of episodes form the training set and the last 20% the evaluation set. We don't use a random permutation, as it would allow future episodes to appear in training. At test time, the beta is re-estimated from the twelve months immediately preceding the test period.

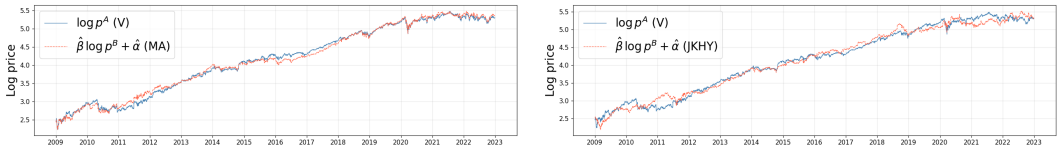
3 Experimental Setup and Results

This chapter applies the methodology described in Chapter 2 with specific parameters and evaluates the proposed framework under two settings. In the primary setting, both agents are trained on synthetic spread data generated by the calibrated regime-switching OU model and evaluated on a held-out simulated test set, providing a controlled comparison under known and reproducible market conditions. In the secondary setting, both agents are trained directly on historical spread episodes and evaluated on real out-of-sample market data, assessing how well the regime advantage transfers beyond the simulation. All hyperparameter values are reported in full to ensure reproducibility.

3.1 Pair and Data

Experiments are conducted on two pairs from the S&P 500 financial sector: Visa / Mastercard (V/MA) and Visa / Jack Henry & Associates (V/JKHY). Daily adjusted closing prices are downloaded from Yahoo Finance. For both pairs, the calibration window spans 2009-01-01 to 2022-12-31, providing approximately 3,500 trading days for HMM fitting and OU calibration. This window is used for pair selection, parameter estimation, and synthetic data generation in the simulated setting, and as the historical training period in the real data setting. The out-of-sample test period covers 2023-01-01 to 2026-05-29 and is held out entirely during calibration and training.

Figure 1 shows the log-price series of each asset alongside its cointegrated counterpart $\hat{\beta} \log p_t^B + \hat{\alpha}$ over the calibration window.



(a) V/MA pair. (b) V/JKHY pair.

Figure 1: Log-price series of each asset (solid) alongside its cointegrated counterpart $\hat{\beta} \log p_t^B + \hat{\alpha}$ (dashed) over the calibration window 2009-2022. The two series track each other closely, confirming a stable cointegrating relationship for both pairs.

3.2 Synthetic Data Generation

The generator is run with the parameters shown in Table 1.

Table 1: Data generation parameters.

Parameter	**Value**	**Parameter**	**Value**
Episodes N	3,000	HMM feature window W	20 days
Episode length T	252 steps	Reversion-rate window	60 days
Time step Δt	1/252	HMM random restarts	20
Regimes	3	Random seed	42

3.3 Agents and Baselines

Both agents share the same three-dimensional observation structure. The regime-aware agent uses the regime-adjusted z-score defined with online EMA calibration as defined in chapter 2.6; the baseline uses the standard rolling z-score. The observation window $W = 20$ days is used for both the z-score and volatility features; observations are clipped to $[-5, 5]$ and $[0, 3]$ respectively. The online calibration smoothing factor is $\alpha = 0.05$, corresponding to an effective memory of approximately 20 visits per regime.

As a classical benchmark, we additionally evaluate a z-score threshold strategy that requires no training. A long position is entered when $z_t < -1.5$, a short position when $z_t > 1.5$, and the position is closed when $|z_t| < 0.5$, where z_t is the rolling z-score defined in Section 2.6. The entry threshold of ± 1.5 follows the convention established in [3].

3.4 PPO Hyperparameters

Table 2 lists the PPO hyperparameters used for both agents.

Table 2: PPO training hyperparameters.

Parameter	**Value**	**Parameter**	**Value**
Network architecture	$[128, 128]$, Tanh	Discount factor γ	0.99
Entropy coefficient	0.005	GAE parameter λ	0.95
Learning rate	3×10^{-4}	PPO clip parameter	0.2
Rollout steps n	252	Value function coefficient	0.5
Mini-batch size	64	Max gradient norm	0.5
Epochs per update	10	Total timesteps	250,000
Parallel environments	4	Transaction cost c	0.001
Training seeds: 7, 8, 12, 15, 42, 43, 67, 123, 456, 789			

3.5 Simulated Test Set

The simulated test set consists of 500 fresh episodes generated with `TEST_SEED` = 999, ensuring no overlap with the training data. To model volatility estimation uncertainty, each test episode applies sigma-only domain randomization: the per-regime volatility σ_r is perturbed by a multiplicative factor drawn uniformly from $[0.8, 1.2]$ independently for each episode and each regime. The equilibrium means μ_r and reversion speed κ are kept at their calibrated values, so the regime-adjusted z-score signal remains directionally correct. The domain randomization reduces the win rate from its oracle value toward a more realistic range, providing a fairer assessment of the agent's robustness to model uncertainty.

3.6 Simulated Test Results

Table 3 reports the performance of all three strategies on the simulated held-out test set of 500 episodes over 10 random seeds.

Across both pairs, the regime-aware agent consistently outperforms the baseline. On V/MA it achieves a mean Sharpe of 1.354 ± 0.064 compared to 0.738 ± 0.040 for the baseline (ΔSharpe $= +0.616$), nearly doubling the mean annual return (6.58% versus 3.54%) while incurring a lower maximum drawdown (-2.89% versus -3.61%). On V/JKHY the advantage

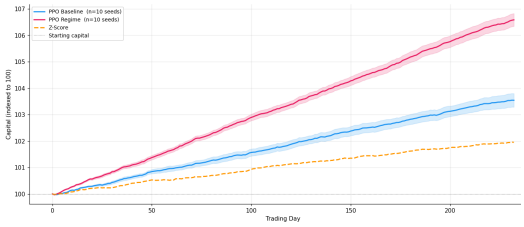
Table 3: Simulated test set results (500 held-out episodes, `TEST_SEED`= 999, σ noise $\pm 20\%$).

Pair	Strategy	Return (%)	Sharpe	MaxDD (%)	Trades
V/MA	PPO Regime	6.58 ± 0.26	1.354 ± 0.064	-2.89 ± 0.11	81.0 ± 40.9
	PPO Baseline	3.54 ± 0.27	0.738 ± 0.040	-3.61 ± 0.17	78.1 ± 54.6
	Z-Score	1.95	0.509	-3.34	32.0
V/JKHY	PPO Regime	8.32 ± 0.80	1.183 ± 0.084	-5.10 ± 0.12	59.3 ± 26.1
	PPO Baseline	3.41 ± 1.16	0.564 ± 0.085	-5.57 ± 1.50	130.9 ± 91.4
	Z-Score	-0.37	-0.035	-6.27	29.8

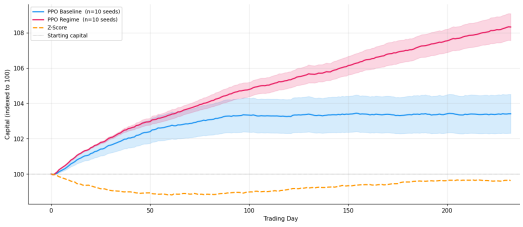
is similarly consistent, with the regime agent achieving a Sharpe of 1.183 ± 0.084 against 0.564 ± 0.085 for the baseline (ΔSharpe $= +0.619$). The narrow standard deviations across seeds confirm that the outperformance is robust rather than driven by a single lucky training run.

The behaviour of the z-score strategy differs markedly between the two pairs. On V/MA it achieves a Sharpe of 0.509, however on V/JKHY it loses money outright (Sharpe $= -0.035$), despite the data being generated from a stationary OU process. This divergence reflects the more pronounced regime switching in V/JKHY: when the Markov chain transitions between regimes with different equilibria, the 20-day rolling mean lags behind the new equilibrium, causing the z-score to generate false mean-reversion signals that the regime-aware agent avoids by centering on the per-regime equilibrium $\tilde{\mu}_{r_t}$ instead.

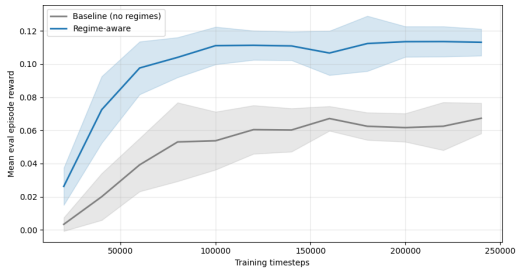
Figures 2a and 2b show the median capital curve for each strategy over the 252-step test episodes, averaged across 10 training seeds. Figures 2c and 2d show the training reward curves for both agents on each pair, confirming that both agents converge stably.



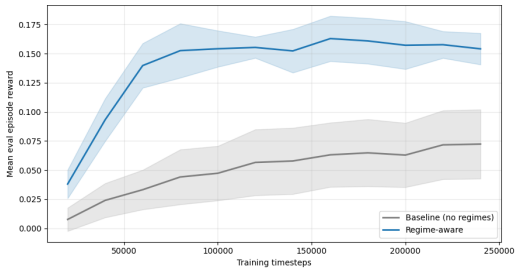
(a) V/MA capital curve.



(b) V/JKHY capital curve.



(c) V/MA training curve.



(d) V/JKHY training curve.

Figure 2: Top: median capital curve for all three strategies on the simulated test set. The shaded regions indicate ± 1 standard deviation over the 10 random training seeds. Bottom: smoothed training reward curves for the baseline and regime-aware PPO agents.

These results suggest that the regime-adjusted z-score provides a more informative trading signal than the rolling z-score under the assumed regime-switching OU model. However, both agents are evaluated on data generated from the same parametric model used for training, and the degree to which the advantage persists on real market data - with its fat tails, non-constant beta, and genuine microstructure noise - remains an open question examined in the following section.

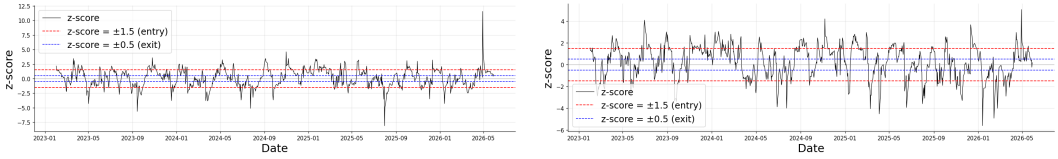
3.7 Real Data Results

Table 4 reports out-of-sample performance on real market data from 2023-01-03 to 2026-05-29 for both pairs, averaged across 10 training seeds.

Table 4: Out-of-sample real data results (2023-01-03 to 2026-05-29).

Pair	Strategy	Return (%)	Sharpe	MaxDD (%)	Trades
V/MA	PPO Regime	7.56 ± 1.34	0.567 ± 0.066	-4.46 ± 0.54	631.8 ± 76.4
	PPO Baseline	-3.40 ± 0.91	-0.172 ± 0.051	-8.22 ± 0.63	373.1 ± 83.7
	Z-Score	-3.48	-0.221	-9.48	106.0
V/JKHY	PPO Regime	26.33 ± 6.33	0.609 ± 0.071	-12.63 ± 2.03	683.7 ± 82.2
	PPO Baseline	15.72 ± 2.72	0.442 ± 0.066	-11.28 ± 1.13	797.1 ± 31.7
	Z-Score	12.37	0.318	-13.77	106.0

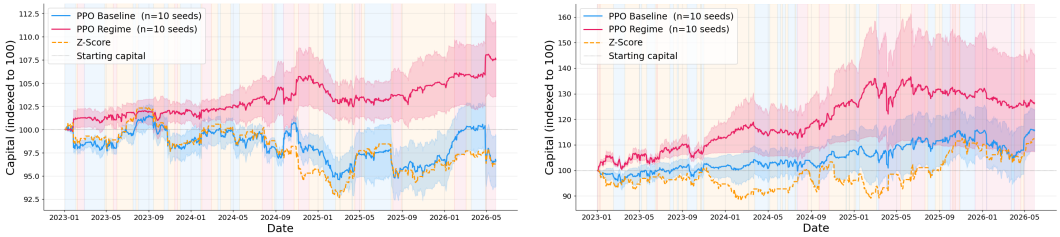
Figure 3 shows the rolling 20-day z-score for both pairs over the test period, confirming that both spreads exhibit regular crossings of the ± 1.5 entry threshold throughout 2023-2026 and thus present genuine mean-reversion opportunities.



(a) V/MA. (b) V/JKHY.

Figure 3: Rolling 20-day z-score over the test period with ± 1.5 entry thresholds (red dashed) and ± 0.5 exit thresholds (blue dashed).

Figure 4 shows the mean capital curve across 10 training seeds for each pair.



(a) V/MA pair.

(b) V/JKHY pair.

Figure 4: Mean capital curve (indexed to 100 at the start of the test period) for all three strategies on real out-of-sample data from 2023-01-03 to 2026-05-29. The shaded regions indicate ±1 standard deviation across 10 random training seeds. Background colors indicate the regime inferred by the causal HMM forward filter: blue = stable, orange = neutral, pink = crisis.

The regime-aware agent outperforms on both pairs, though the nature of the advantage differs. On V/MA, where the baseline and z-score both lose money, the regime agent achieves a Sharpe of 0.567 ± 0.066 (ΔSharpe = +0.739 over the baseline), demonstrating that the regime signal successfully identifies mean-reversion opportunities that the regime-agnostic strategies miss entirely.

On V/JKHY all three strategies are profitable, with the z-score achieving a return of 12.37% and Sharpe of 0.318, confirming that the spread was mean-reverting throughout the test period. The regime agent still outperforms with a Sharpe of 0.609(ΔSharpe = +0.167), though the advantage over the baseline is more modest and the standard deviation across seeds is larger (±6.33% in return), reflecting higher sensitivity to the specific training run on this pair.

Across both pairs the regime-aware agent consistently achieves the highest Sharpe ratio. On V/MA the maximum drawdown is substantially lower, supporting the conclusion that regime information provides a genuine edge in live market conditions when the cointegrating relationship is active.

The mean position held by each strategy over the test period is shown in Appendix C, illustrating how the regime-aware agent responds to the inferred market regimes.

It is important to note that the real data performance of both agents is sensitive to the choice of pair and, more specifically, to whether the sector was significantly disrupted by recent geopolitical or macroeconomic events. The pairs selected in this study - V/MA and V/JKHY - operate in the financial technology and payments sectors, which were largely insulated from the commodity price shocks, supply chain disruptions, and energy market volatility that characterized the 2022–2026 period. Pairs from sectors directly affected by such events, such as energy or consumer staples, are more likely to experience structural breaks in their cointegrating relationship during this period, rendering the regime-switching framework ineffective regardless of the agent's learning capacity. The results presented here should therefore be interpreted as evidence that the approach works when its core assumption - a stable cointegrating relationship in the test period - is satisfied, rather than as a general claim about the strategy's performance across all pairs and market conditions.

4 Responsible Research

4.1 Ethical Considerations

This research investigates statistical arbitrage through reinforcement learning applied to publicly traded equity pairs. We reflect on several ethical dimensions relevant to this work.

Market impact and fairness. Pairs trading is a form of statistical arbitrage that profits from temporary mispricings between cointegrated assets. At the scale of this research (a simulation study with no live capital deployment) there is no direct market impact. At larger institutional scales, statistical arbitrage is generally considered to contribute to market efficiency by accelerating the correction of price deviations, though this view is not without controversy; if sufficiently many agents pursue the same strategy simultaneously, crowding effects can amplify rather than dampen volatility. The strategy developed in this thesis is not deployed in live markets, and the scale at which it could affect market dynamics is well beyond the scope of this work.

Data and privacy. All data used in this research consists of publicly available daily adjusted closing prices obtained from Yahoo Finance. No personal, sensitive, or proprietary data is used at any stage. There are no human subjects and no privacy concerns.

Environmental cost. Training the PPO agents requires repeated forward and backward passes through a small neural network over 250,000 environment steps. The computational footprint of this research is modest: all experiments were run on a standard laptop CPU with training times in the range of minutes per agent. No specialized hardware or large-scale compute was required.

4.2 Reproducibility

We have taken several steps to ensure that the results reported in this thesis can be reproduced.

Fixed random seeds. All stochastic components of the pipeline are controlled by fixed seeds. The training data generator uses `RANDOM_SEED` = 42 for the Monte Carlo simulation. PPO training uses 10 random seeds for both weight initialization and environment episode sampling mentioned in the Experimental Setup and Results chapter. The simulated test set uses a separate `TEST_SEED` = 999, which differs from the training seed to ensure no overlap between training and test episodes.

Open-source tools. The entire pipeline is built on open-source libraries: Stable-Baselines3 [15] for PPO, `hmmlearn` [6] for the Gaussian HMM, `yfinance` [17] for price data, Gymnasium [4] for the trading environment, and standard scientific Python (`numpy`, `scipy`, `statsmodels`). All are freely available and widely used, lowering the barrier to replication.

Full hyperparameter reporting. All hyperparameters used in the experiments are reported in full in the experimental setup chapter (Table 1 and Table 2), including network architectures, learning rates, PPO-specific coefficients, calibration parameters, and data generation settings.

5 Conclusions and Future Work

5.1 Summary and Conclusions

This thesis investigated whether a regime-aware reinforcement learning agent outperforms a regime-neutral counterpart in pairs trading. Because daily price data for a single pair provides only around 2,500 observations - far fewer than a PPO agent typically requires to converge - the thesis develops a regime-switching simulation pipeline combining a Hidden Markov Model and an Ornstein-Uhlenbeck process to generate 3,000 synthetic training episodes, and designs a regime-adjusted z-score that encodes per-regime equilibrium information directly into the agent's observation. Both agents are additionally trained and evaluated on real historical data to assess how well the advantage transfers beyond the simulation.

The answer to the primary research question is yes. Addressing the three sub-questions in turn: *(i)* the regime-aware agent outperforms the regime-neutral PPO baseline on all evaluated pairs in both simulation (ΔSharpe = +0.616 on V/MA and +0.619 on V/JKHY) and real data (ΔSharpe = +0.739 on V/MA and +0.167 on V/JKHY). *(ii)* The regime-aware agent outperforms the classical z-score threshold strategy in all simulation experiments and on both real data pairs, where the z-score loses money on V/MA while the regime agent generates a positive return. *(iii)* The performance advantage observed in simulation persists on real historical data for both pairs, confirming that the regime signal provides genuine value beyond the controlled setting, though the magnitude of the advantage is pair-dependent and diminishes when the cointegrating relationship is weaker.

The advantage stems from the regime-adjusted z-score signal, which centers on the per-regime equilibrium rather than a stale rolling mean, identifying high-confidence entry points that the regime-neutral agent cannot distinguish from noise. The result is consistent across 10 training seeds and robust to $\pm 20\%$ perturbations of the spread volatility in simulation, suggesting it is not an artefact of a single training run or the exact calibrated parameters.

The thesis makes three concrete contributions. First, a complete simulation pipeline covering pair selection, OU calibration, HMM regime detection, per-regime parameter estimation, and regime-switching data generation. Second, the regime-adjusted z-score as a compact three-dimensional observation signal that makes regime information directly actionable without increasing model capacity relative to the baseline. Third, an online EMA calibration mechanism that adapts the per-regime equilibrium estimate over time, reducing dependence on stale historical calibrations.

5.2 Future Work

Several directions emerge directly from the limitations identified in this work.

Non-Gaussian HMM emissions. The current HMM assumes Gaussian emission distributions for the three spread features. Real financial features - particularly log-volatility and the logit reversion rate - exhibit fat tails and occasional extreme values during market dislocations that a Gaussian model systematically underestimates. A possible improvement is to use Student-t distributions, expecting that the regime classifier would become more robust to outliers. Improved regime detection accuracy would directly give us a more reliable z_adj signal, since the per-regime equilibrium $\tilde{\mu}_{r_t}$ is only useful when the regime label $\hat{r}_t$ is correct.

Simulation with structural breaks and regime degradation. The current simulation assumes that the spread is always cointegrated: at every step the Markov chain transitions between regimes but the OU process never loses its mean-reverting property entirely. Real markets, however, go through extended periods where the cointegrating relationship weakens or breaks down temporarily. Extending the simulator to include occasional structural break episodes and volatility clustering would substantially close the sim-to-real gap while retaining the key advantage of simulation: the ability to generate thousands of diverse training episodes. Whether such an enriched simulator would outperform direct historical training remains an open question. A promising direction is a hybrid approach that combines real historical episodes with distorted synthetic ones, giving the agent both authentic market dynamics and sufficient volume to learn robust behavior across a wide range of market conditions.

Implicit regime learning via recurrent policies. Rather than detecting regimes explicitly with an HMM, a recurrent policy (e.g. LSTM) could learn to track regime-like dynamics implicitly through its hidden state. This eliminates all assumptions about the number and nature of regimes, removes the need for offline calibration, and may generalize better to out-of-sample periods where the regime structure has changed.

Capital reallocation during non-cointegrated periods. The current framework trades the spread continuously throughout the test period, even when an online stationarity test indicates that cointegration has temporarily broken down. Future work could extend the action space to include a cash allocation, allowing the agent to shift capital to a risk-free bank account during such periods rather than remaining exposed to a non-reverting spread. This would reduce drawdowns during structural break periods and produce a more realistic assessment of the strategy's risk-adjusted returns over multi-year horizons.

Short-selling margin requirement. Taking a short position on the spread requires borrowing the shorted asset and posting collateral as a guarantee, typically a fraction of the short position's notional value. In the current formulation the agent can freely take positions in $[-1, 1]$ without reserving any capital for this margin requirement, effectively overstating the available capital for trading. Future work could incorporate a margin constraint into the environment by reducing the effective capital by a fraction $m \cdot |w_t|$ when $w_t < 0$, where m is the required margin rate. This would constrain the maximum short exposure the agent can sustain and produce a more conservative and realistic evaluation of the strategy.

References

[1] Robert F. Engle and C. W. J. Granger. Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):251–276, 1987.

[2] Oyvind Foshaug. *Implementation of Pairs Trading Strategies*, 2010.

[3] Evan Gatev, William N. Goetzmann, and K. Geert Rouwenhorst. *Pairs Trading: Performance of a Relative Value Arbitrage Rule*, 2006.

[4] gimnasium developers. Gimnasium. `https://gymnasium.farama.org/index.html`, 2023.

[5] James D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, NJ, 1994.

[6] hmmlearn developers. hmmlearn: Hidden markov models in python, with scikit-learn like api. `https://hmmlearn.readthedocs.io/`, 2010.

[7] Zhengyao Jiang, Dixing Xu, and Jinjun Liang. *A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem*, 2017.

[8] Eun-chong Kim, Han-wook Jeong, and Nak-young Lee. *Global Asset Allocation Strategy Using a Hidden Markov Model*, 2019.

[9] Siyu Lin and Peter A. Beling. *An End-to-End Optimal Trade Execution Framework Based on Proximal Policy Optimization*, 2021.

[10] John Moody and Matthew Saffell. *Learning to trade via direct reinforcement*, 2001.

[11] Rizwan Mushtaq. *Augmented Dickey Fuller Test*, 2011.

[12] Gabriel Nixon Raj. Adaptive and regime-aware RL for portfolio optimization. *arXiv preprint arXiv:2509.14385*, 2025.

[13] Cristian Quintero, Diego Leon, Javier Sandoval, and German Hernandez. *Deep Reinforcement Learning in Continuous Action Spaces for Pair Trading: A Comparative Study of A2C and PPO*, 2025.

[14] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[15] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-Baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.

[16] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. *Proximal Policy Optimization Algorithms*, 2017.

[17] Yahoo Finance. *Yahoo Finance*, 2024. `https://finance.yahoo.com`.

A Dollar P&L Derivation

Since the spread is computed from log-prices, Δz_t consists of log-returns. Using the first-order Taylor approximation $\log(1+x) \approx x$ for small x gives:

$$\Delta \log p_t^A = \log\left(1 + \frac{p_t^A - p_{t-1}^A}{p_{t-1}^A}\right) \approx \frac{p_t^A - p_{t-1}^A}{p_{t-1}^A} = \tilde{r}_t^A$$

and analogously for asset B, so that:

$$\Delta z_t \approx \tilde{r}_t^A - \hat{\beta} \cdot \tilde{r}_t^B$$

With capital C_t and position weight w_t, the agent holds $w_t C_t/(1+\hat{\beta})$ dollars long in asset A and $w_t\hat{\beta}C_t/(1+\hat{\beta})$ dollars short in asset B, so the total notional across both legs equals $w_t C_t$. The dollar P&L from each leg is:

$$\mathrm{PnL}^A_{t+1} = \frac{w_t C_t}{1+\hat{\beta}} \cdot \tilde{r}^A_{t+1}$$

$$\mathrm{PnL}^B_{t+1} = -\frac{w_t \hat{\beta} C_t}{1+\hat{\beta}} \cdot \tilde{r}^B_{t+1}$$

Since $c \cdot |w_{t+1} - w_t|$ is expressed in spread units, it is converted to dollars using the same scaling factor $C_t/(1+\hat{\beta})$. Summing both legs and subtracting the transaction cost:

$$\text{Dollar PnL}_{t+1} = \frac{w_t C_t}{1+\hat{\beta}} \left(\tilde{r}^A_{t+1} - \hat{\beta} \cdot \tilde{r}^B_{t+1}\right) - \frac{C_t \cdot c \cdot |w_{t+1} - w_t|}{1+\hat{\beta}}$$

$$\approx \frac{w_t C_t \cdot \Delta z_{t+1}}{1+\hat{\beta}} - \frac{C_t \cdot c \cdot |w_{t+1} - w_t|}{1+\hat{\beta}} = \frac{C_t}{1+\hat{\beta}} \left(w_t \cdot \Delta z_{t+1} - c \cdot |w_{t+1} - w_t|\right) = \frac{C_t \cdot r_{t+1}}{1+\hat{\beta}}$$

B Calibrated Model Parameters

Global Parameters

Table 5: Global hedge ratio and OU parameters.

Parameter	**V/MA**	**V/JKHY**
Hedge ratio $\hat{\beta}$	0.7891	1.0357
Mean-reversion speed κ	9.8595	5.7261
Equilibrium mean μ_{ou}	0.7288	-0.1337
Volatility σ_{global}	0.1115	0.1843
Half-life (trading days)	17.7	30.5

Per-Regime OU Parameters

Table 6: Per-regime OU parameters for V/MA.

Parameter	**Stable** ($r=0$)	**Neutral** ($r=1$)	**Crisis** ($r=2$)
Equilibrium mean μ_r	0.7360	0.7321	0.7152
Reversion speed κ_r	5.59	6.48	28.97
Volatility σ_r	0.07331	0.09001	0.14130
Half-life (trading days)	31	27	6
In-regime samples N_r	571	618	430

Table 7: Per-regime OU parameters for V/JKHY.

Parameter	**Stable** ($r = 0$)	**Neutral** ($r = 1$)	**Crisis** ($r = 2$)
Equilibrium mean μ_r	-0.1367	-0.1314	-0.1391
Reversion speed κ_r	3.89	6.62	5.88
Volatility σ_r	0.13612	0.13882	0.20482
Half-life (trading days)	45	26	30
In-regime samples N_r	519	583	499

HMM Transition Matrices

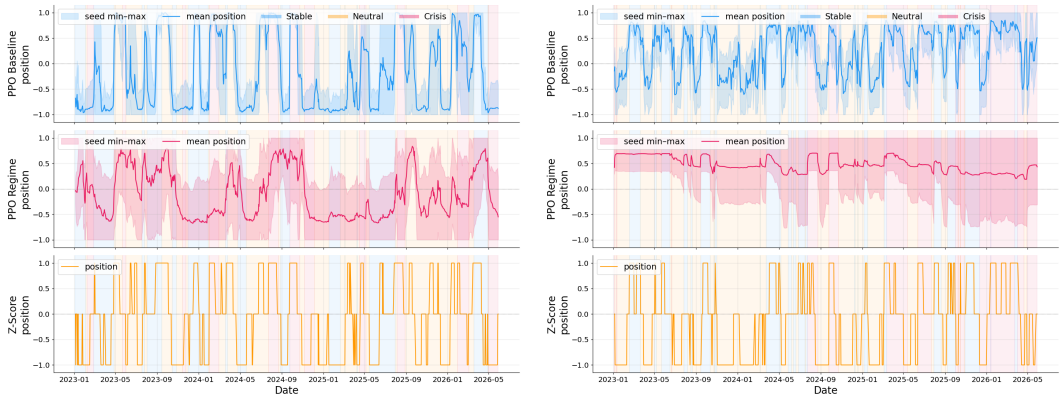
Table 8: HMM transition matrix $\mathbf{A}$ for V/MA, where $A_{rr'} = P(s_t = r' \mid s_{t-1} = r)$.

	Stable	**Neutral**	**Crisis**
Stable	0.9589	0.0411	0.0000
Neutral	0.0324	0.9366	0.0311
Crisis	0.0110	0.0336	0.9554

Table 9: HMM transition matrix $\mathbf{A}$ for V/JKHY.

	Stable	**Neutral**	**Crisis**
Stable	0.9226	0.0523	0.0252
Neutral	0.0472	0.9399	0.0129
Crisis	0.0263	0.0156	0.9581

C Strategy Positions Over the Test Period



(a) V/MA. (b) V/JKHY.

Figure 5: Mean position over the test period for each strategy across 10 training seeds. Background colors indicate the regime inferred by the causal HMM forward filter: blue = stable, orange = neutral, pink = crisis. The shaded band shows the seed min-max range.