



Personalised Multi-Market Investment Advisory Using Reinforcement Learning and Investor Risk Profiling

Huzaifa Najam¹, Muhammad Mubashir¹, Abdul Rafay¹, Aneela Nargis^{1,*}

¹Department of Computer Science, DHA Suffa University, Karachi, Pakistan

ARTICLE INFO

Article History:

Received:	May	30, 2026
Revised:	June	05, 2026
Accepted:	June	20, 2026
Available Online:	June	30, 2026

Keywords:

portfolio recommendation
reinforcement learning
proximal policy optimisation
K-Means clustering
FinBERT
Pakistan Stock Exchange
robo-advisory
investment advisory
backtesting
ablation study

Classification Codes:

I.2.6 [Learning]; H.4.m [Information Systems Applications]; I.5.3 [Clustering]; J.1 [Administrative Data Processing — Finance]

Funding:

This research received no specific grant from any funding agency in the public or not-for-profit sector.

ABSTRACT

Retail investors in emerging markets such as Pakistan have limited access to low-cost portfolio guidance that covers domestic and international equities. This paper presents AIPRS (AI-Powered Portfolio Recommendation System), a working advisory prototype that combines K-Means investor risk profiling, proximal policy optimisation (PPO) agents that issue BUY, HOLD, or SELL signals for US and Pakistan Stock Exchange (PSX) equities with the investor's risk profile encoded in the state and in a risk-adaptive confidence threshold, FinBERT news sentiment, and cluster-based collaborative filtering. A controlled evaluation on 14 US equities (training 2016 to 2022, validation 2023, test January 2024 to September 2026; five seeds; 10 basis-point costs) shows that risk conditioning raised the out-of-sample Sharpe ratio relative to a risk-agnostic PPO agent (for example, 0.65 versus 0.51), with seed-dependent and only partly significant gains, and that the effect came from the state input rather than the threshold. No PPO configuration outperformed equal-weight buy-and-hold (Sharpe 1.04) or a 20/50-day moving-average rule (0.91), and high turnover made the agents cost-sensitive. External validation on 3,857 Survey of Consumer Finances households showed that the four-cluster profiling solution is stable (bootstrap Jaccard 0.95 to 0.98) and agrees with observed equity holdings, but that 56 profiles are too few for a reliable model. Collaborative filtering exceeded a popularity baseline (Recall@3 0.746 versus 0.712). AIPRS is best regarded as a transparent, extensible advisory prototype rather than a validated source of excess returns.



© 2026 The authors published by JCIS. This is an Open Access Article under the Creative Common Attribution Non-Commercial 4.0

Corresponding Author's Email: a.nargis@dsu.edu.pk

Citation: Najam, H., Mubashir, M., Rafay, A., Nargis, A. (2026). Personalised Multi-Market Investment Advisory Using Reinforcement Learning and Investor Risk Profiling. Journal of Computing and Information Sciences (JCIS).

1. Introduction

Investment advice has become far more accessible over the past decade, and this shift has reshaped financial technology in meaningful ways. Robo-advisory platforms, which automate portfolio management decisions using algorithms rather than human advisors, have grown from a niche concept to a multi-trillion-dollar category globally [1]. Platforms such as Betterment and Wealthfront demonstrated that low-cost, data-driven portfolio guidance could be made accessible to retail investors at scale [2]. Machine learning pushed this further by enabling investor profiling through clustering, reinforcement learning (RL) for dynamic signal generation, and natural language processing (NLP) for sentiment-aware market monitoring [3], [4].

Despite this progress, most existing robo-advisory research and commercially deployed systems share two limitations. First, they are developed mainly around Western markets. US equities, European indices, and dollar-denominated instruments are the norm, with comparatively little coverage of the Pakistan Stock Exchange (PSX), an emerging market that remains underserved by algorithmic advisory tooling [5], [6], [7]. Second, the AI components most relevant to retail advisory, namely investor segmentation, reinforcement learning (RL) trading signals, and natural language processing (NLP) sentiment analysis, are usually studied in isolation [8], [9], [36]. Several strands of work have begun to connect them: RL agents that consume sentiment or news-derived states [29], [30], robo-advisors that learn or update an investor's risk preference [32], [33], [34], [35], and risk-aware RL portfolio managers [31]. Systems that combine investor segmentation with an RL signal generator whose behaviour is explicitly conditioned on the investor's risk profile, and that cover an emerging market, are still uncommon in the published literature.

This paper presents AIPRS (AI-Powered Portfolio Recommendation System), a working advisory prototype that brings these components together, and it evaluates each component with the evidence currently available. The contributions are as follows.

- A four-cluster K-Means investor segmentation model that maps a profiling form to a named risk tier (Conservative, Moderate, Aggressive, Very Aggressive) and a predefined, tier-specific strategic asset allocation. Because the deployed model was fitted on only 56 profiles, we add an external validation of the same procedure on 3,857 households from the Federal Reserve Survey of Consumer Finances (SCF) 2007 to 2009 panel, covering the choice of k , seed stability, bootstrap stability, sample-size sensitivity, temporal stability, and agreement with observed equity holdings.
- Two proximal policy optimisation (PPO) actor-critic agents, one for US equities and one for approximately seventy-four PSX equities, whose 24-dimensional state concatenates sixteen technical features with eight investor-profile features, combined with a risk-adaptive confidence threshold at inference time. We give a formal specification of the advisory environment and separate reward functions for BUY, HOLD, and SELL.
- A controlled, multi-seed evaluation of the risk-conditioned PPO design on a chronologically partitioned US equity data set (training 2016 to 2022, validation 2023, test January 2024 to September 2026). This evaluation includes a risk-agnostic PPO ablation, a cost-aware portfolio backtest with slippage and turnover, conventional technical-rule and supervised-learning baselines, statistical significance testing, and robustness checks across market periods and ticker groups.
- A FinBERT-based news sentiment layer for US and PSX English-language headlines, presented as contextual information for the user rather than as an input to the PPO decision, together with a user-based collaborative filtering (CF) feature whose ranking quality is measured with Precision@K, Recall@K, and NDCG@K on real household asset-holding data.

This work is guided by three research questions. RQ1: Does encoding an investor's risk profile in the PPO state and inference threshold produce advisory behaviour that differs measurably and in the expected direction across risk tiers, and does it change risk-adjusted performance compared with a risk-agnostic PPO agent? RQ2: Can a single RL architecture and training pipeline be applied to both a developed market (US equities) and an emerging, lower-liquidity market (PSX) without redesigning the model? RQ3 (revised): How reliable is the FinBERT sentiment layer as a contextual signal, and what evidence would be needed before sentiment is used as a decision input? RQ3 was narrowed during revision because the deployed system displays sentiment beside the PPO signal and does not feed it into the policy, so it cannot be tested as a source of additional predictive information within this study. The questions are revisited in Section 9.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the system architecture, including privacy and security safeguards. Sections 4 to 7 describe the methodology of each component. Section 8 presents the experimental protocol and results. Section 9 discusses the findings and limitations, and Section 10 concludes.

2. Related Work

2.1 Portfolio Theory and Robo-Advisory

Markowitz's mean-variance framework [10] provides the mathematical basis for portfolio optimisation: investors choose asset weights that maximise expected return for a given level of risk. Robo-advisory platforms build on this idea with periodic rebalancing and tax-loss harvesting [11], [12], but most still rely on static questionnaire-based risk scoring [2], [11]. Clustering methods such as K-Means can reveal investor groupings that rule-based questionnaires miss [13]. A separate line of work treats the investor's risk preference as something to be learned or updated: Alsabah et al. infer risk preferences from the portfolio choices clients make [32], Capponi et al. model the trade-off between client interaction and personalisation [33], Wang and Yu combine inverse optimisation of risk preferences with deep RL [34], and Keffert shows that measurement error in elicited risk preferences can materially reduce client welfare [35]. These studies motivate both the risk-conditioning in AIPRS and the stability analysis of the profiling step reported in Section 8.

2.2 Reinforcement Learning for Trading and Advisory

Reinforcement learning has been applied to trading since at least the late 1990s [14], and progress accelerated with deep RL. Proximal policy optimisation [15] constrains the size of each policy update and has become a widely used algorithm in financial applications. Deep RL has been applied to portfolio allocation [16], [9], single-asset trading [28], and ensemble trading strategies, sometimes with Sharpe-ratio-based rewards that make the agent more risk-sensitive [17], [9], and open-source frameworks such as FinRL [27] have standardised many of these pipelines. Ye et al. augment the RL state with asset-movement predictions derived from price and news [29], Koratamaddi et al. add market sentiment to a DDPG portfolio agent [30], and DeepTrader balances return and drawdown with a market-condition embedding [31]. Surveys note that results in this area are often sensitive to random seeds, data periods, and transaction-cost assumptions [36], [37], [47]. Most of these systems optimise a single objective regardless of who receives the recommendation. AIPRS instead places the investor's risk profile in the state vector and applies a risk-adaptive inference threshold; Section 8 tests whether this conditioning changes behaviour and outcomes relative to a risk-agnostic agent.

2.3 NLP Sentiment Analysis in Finance

The link between news sentiment and asset prices has been studied extensively [18]. Transformer-based language models improved financial sentiment classification [8]. BERT [19] provided the pre-training architecture on which domain-specific models are built, and FinBERT [20] further trained BERT on financial text; on the Financial PhraseBank [49] it reports an accuracy of 0.86 and a macro F1 of 0.84 on the full data set, and 0.97 and 0.95 on the subset with complete annotator agreement [20]. The ProsusAI/finbert checkpoint used here is freely available through HuggingFace Transformers. All of these benchmarks are English-language; we are not aware of a public, labelled benchmark for PSX-specific or Urdu-language financial news.

2.4 Pakistan Stock Exchange and Machine Learning Research

Academic work applying machine learning to the PSX has been growing, though it remains limited overall. Studies using LSTM networks on PSX sector data report strong directional accuracy for stable, high-liquidity sectors such as banking, cement, and fertiliser, but weaker performance for volatile stocks [6], [5]. Comparative studies evaluating ANN, SVM, Random Forest, and LSTM models on PSX technical indicators report direction-of-movement accuracies in the 58-85% range [7]. Nearly all of this work focuses on price forecasting accuracy rather than advisory signal generation, and none incorporates reinforcement learning or personalised risk profiling.

2.5 Gap Identified

Three gaps emerge from this literature. First, PSX-focused machine learning research has concentrated on prediction accuracy rather than risk-aware advisory output [6], [7], [5]. Second, investor segmentation and RL signal generation are usually treated as separate problems; work that personalises RL advice to the investor's risk profile is

growing [32], [34] but remains limited, and the benefit of such conditioning is rarely tested against a risk-agnostic counterpart. Third, we found few openly documented systems that serve an emerging market such as the PSX alongside a developed market within one model architecture [4], [5]. Table 1 positions AIPRS against representative prior work. The comparison is based on what each study reports and is not an exhaustive survey.

Work	RL	NLP / sentiment	Learned risk profile	PSX	Personalised output
Betterment / Wealthfront-style robo-advisors [2], [11]	No	No	Questionnaire-based	No	Allocation only
Jiang et al., deep RL portfolio management [16]	Yes	No	None	No	No
Yang et al., deep RL for portfolio management [9]	Yes	No	None	No	No
Théate and Ernst, deep RL algorithmic trading [28]	Yes	No	None	No	No
Liu et al., FinRL library [27]	Yes	No	None	No	No
Ye et al., SARL with news-augmented states [29]	Yes	Yes	None	No	No
Koratomaddi et al., sentiment-aware DDPG [30]	Yes	Yes	None	No	No
Wang et al., DeepTrader risk-return balance [31]	Yes	No	None	No	Market-level risk control
Alsabah et al., learning risk preferences [32]	Bandit	No	Learned from choices	No	Personalised allocation
Capponi et al., personalised robo-advising [33]	No	No	Updated via interaction	No	Personalised allocation
Wang and Yu, inverse optimisation + deep RL [34]	Yes	No	Inferred (inverse opt.)	No	Personalised allocation
Wu et al., K-means for robo-advisory [13]	No	No	Clustering-based	No	No signal generation
PSX ML forecasting studies [5], [6], [7]	No	No	None	Yes	Price forecast only
AIPRS (this work)	Yes	Yes (display only)	Clustering-based	Yes	Risk-conditioned RL signal

TABLE 1. POSITIONING OF AIPRS AGAINST REPRESENTATIVE PRIOR WORK

3. System Architecture

AIPRS is built in three tiers, each with a well-defined role.

- The Presentation Tier is a Streamlit multi-page dashboard with custom dark-mode CSS theming, covering ten pages: investor profile form and sign-up, portfolio overview, AI recommendations, market data, news and sentiment, feedback, US PPO advisor, PSX PPO advisor, login, and settings. Page files contain no business logic beyond Streamlit session management.

- The Application Tier contains five processing engines: (i) the K-Means risk profiling engine; (ii) the MPT allocation engine; (iii) the two PPO advisory agents; (iv) the FinBERT sentiment pipeline; and (v) the collaborative filtering engine. Market data is fetched from Alpaca Markets (US equities, 15-minute delayed IEX feed), Yahoo Finance via yfinance (PSX historical OHLCV), psxterminal.com (PSX live intraday prices, 60-second refresh), and Google News RSS via feedparser (PSX financial news).

- The Storage Tier is MongoDB Atlas with three collections: users (profile answers, bcrypt-hashed password, cluster assignment, saved holdings, preferences), feedback, and surveys. A unique index on users.email prevents duplicate registrations. Serialised model artefacts are loaded at server start-up through Streamlit's @st.cache_resource mechanism and shared across user sessions.



Fig. 1. AIPRS three-tier system architecture: presentation, application, and storage tiers.

3.1 Privacy and Data Security

AIPRS stores personal and financially sensitive information, including age, income band, risk tolerance, and self-reported holdings. The current prototype stores passwords only as bcrypt hashes, and MongoDB Atlas encrypts connections with TLS by default. The collaborative filtering feature reads other users' asset-class preferences only as aggregated frequency counts within a cluster, and it displays asset-class names only, never identities, individual holdings, or peer counts.

Before any production deployment we consider the following controls necessary: storing database credentials and API keys as environment secrets; enforcing per-user authorisation so that a session can read and modify only its own document; setting a minimum cluster size (for example, ten users) below which peer recommendations are suppressed, so that a small cluster cannot reveal an individual's choices; encrypting sensitive profile fields at rest; and publishing a retention and deletion policy that lets users export or erase their profile. AIPRS provides information rather than regulated investment advice and does not execute orders, so it holds no brokerage credentials or payment data. A production version would also need to comply with applicable data-protection and securities-advice regulations in Pakistan and in any other jurisdiction it serves.

4. Investor Risk Profiling

4.1 Profiling Form

Following established psychometric approaches to financial risk tolerance assessment [21], each user completes a seven-question profile at sign-up covering age, annual income range (ordinal), self-assessed risk tolerance (1 to 10), investment horizon (ordinal, short-term to very long-term), investment experience (beginner to advanced), primary financial goal (long-term stability, stable income, short-term trading, or retirement), and preferred asset classes. The last two fields are stored as categorical preference metadata and are not used for clustering.

4.2 Model Training

Five numerical features are used for clustering: age (continuous), income range score (ordinal 1 to 5), risk tolerance (1 to 10), horizon score (ordinal 1 to 5), and experience score (ordinal 1 to 5). Each is standardised to zero mean and unit variance with scikit-learn's StandardScaler, and K-Means is trained with $k = 4$ and $\text{random_state} = 42$ on a seed data set of 56 investor profiles. The fitted model and scaler are saved as `kmeans_model.pkl` and `scaler.pkl`. Because 56 profiles is a small sample, Section 8.8 evaluates the same pipeline on 3,857 households from the SCF, examining the choice of k , stability across random initialisations and bootstrap resamples, and how stability depends on sample size.

4.3 Cluster Mapping and MPT Allocations

The four clusters are mapped to named risk tiers by ordering their centroids on mean risk tolerance and horizon. Each tier is then assigned the target allocation in Table 2. These allocations are **predefined policy weights**, set by the authors according to the mean-variance principle that expected return and volatility should rise together with risk capacity [10], and in line with conventional equity/fixed-income bands used by retail robo-advisors [2], [11]. They are **not** the output of a numerical mean-variance optimisation: no expected-return vector, covariance matrix, or optimiser is estimated in the current system. We state this explicitly because the term MPT-based could otherwise suggest a formal optimisation. The allocations are applied as fixed target weights; Fig. 3 illustrates them with proxy exchange-traded funds held at the target weights. Replacing the fixed weights with an estimated efficient frontier (for example, shrinkage covariance estimates with long-only and per-class bound constraints, rebalanced quarterly) is left as future work. When a user signs in, the stored StandardScaler transforms the five-feature profile, KMeans.predict() returns the cluster index, and the index is stored in the user's MongoDB document and shown as a colour-coded badge across the dashboard.

Cluster	Label	Equities %	Fixed Inc. %	Comm. %	Cash %
0	Conservative	30	50	10	10
1	Moderate	60	25	10	5
2	Aggressive	85	5	8	2
3	Very Aggressive	90	5	3	2

TABLE 2. RISK TIERS AND PREDEFINED TARGET ASSET ALLOCATIONS (%)

4.4 Collaborative Filtering

The AI Recommendations page includes a lightweight, user-based collaborative filtering feature [22]. The system retrieves registered user documents from MongoDB and keeps those in the same risk cluster as the current user. Their asset-class preferences from the profiling form are aggregated with a frequency counter, and the five most frequent asset classes among cluster peers that the current user has not selected are shown as “Trending Among Investors Like You.” No explicit ratings or transaction histories are needed; the only signal is how often asset preferences co-occur within a risk cluster. We treat this component as a system feature rather than a core research contribution, and Section 8.9 quantifies its ranking quality on real household holdings with Precision@K, Recall@K, and NDCG@K.

5. PPO Advisory Agents

5.1 Environment Design

AIPRS defines a custom episodic advisory environment (AdvisoryEnv) without depending on OpenAI Gym or Gymnasium. The task is formulated as a Markov decision process with three discrete actions, $a_t = \text{HOLD, BUY, or SELL}$ (encoded as 0, 1, and 2). Each episode runs for $L = 120$ trading days (about six months) on one ticker for one investor profile. The environment receives a pre-computed, normalised market-feature matrix X of size $T \times 16$ and an investor-profile vector u of length 8.

The state at step t is $s_t = [x_t; u]$, a 24-dimensional vector formed by the concatenation of the sixteen normalised market features and the eight profile features. The market features are log returns over 1, 5, and 20 days; RSI-14; the MACD histogram divided by price; Bollinger Band position; two EMA ratios (price relative to the 12-day and 26-day EMAs); realised volatility over 5 and 20 days; price relative to the 20-day SMA; log volume ratio relative to the 20-day average; ATR-14 divided by price; Stochastic %K; and a cyclical day-of-week encoding (sine and cosine). The profile features are the scaled risk tolerance ρ , ranging from 0 to 1; a one-hot cluster code (4 dimensions); encoded investment horizon; encoded experience level; and scaled age.

Let P_t denote the closing price at day t . The one-bar forward log return is

$$r_t = \ln(P_{t+1}/P_t) \quad (1)$$

and the reward for each action is defined separately as

$$R_t^{\text{BUY}} = (1 + \rho)r_t \quad (2)$$

$$R_t^{\text{HOLD}} = 0 \quad (3)$$

$$R_t^{\text{SELL}} = -(1 + \rho)r_t \quad (4)$$

The BUY and SELL rewards share a magnitude but have opposite signs, because they correspond to opposite exposures. BUY earns the next-day return of a long position, so it is rewarded when the price rises and penalised when it falls. SELL earns the return of the opposite exposure, so it is rewarded when the price falls, meaning that stepping aside avoided a loss, and penalised when the price rises, meaning that upside was forgone. HOLD is a zero-exposure action with zero reward. The earlier manuscript described the BUY and SELL terms in a way that suggested an identical formulation; Eqs. (2) to (4) state the definitions explicitly. The multiplier $(1 + \rho)$, with ρ the scaled risk tolerance, scales both reward and penalty by up to a factor of two for the most aggressive profile. Because gains and losses are scaled together, the multiplier does not change which action has the highest expected reward for a given state. It changes the magnitude of the learning signal, and therefore the gradient weight given to aggressive profiles, and it lets the policy condition on ρ through the state. The sign symmetry also removes the incentive for a trivial always-BUY policy that arises when only long exposure is rewarded in a rising market.

Action semantics. AdvisoryEnv is a signal environment, not a portfolio-accounting environment. It does not track cash, share inventory, or open positions, and the action at step t does not change the state at step $t + 1$, because the state contains only exogenous market features and the fixed profile. The actions are therefore interpreted as one-day directional views on the next bar: BUY is a recommendation to hold long exposure over the next day, SELL is a recommendation to avoid or reduce exposure (it is scored as though a short position were held, but AIPRS never executes or recommends short sales, which retail investors on the PSX generally cannot use), and HOLD is a recommendation to take no new directional view (scored as zero exposure). Transaction costs are not part of the training reward. Existing holdings and the cost of changing positions are handled only in the evaluation backtest, where signals are converted to positions (Section 5.5).

5.2 Actor-Critic Architecture

The agent uses a shared-backbone actor-critic network in PyTorch. With $h_0 = s_t$, the trunk computes $h_l = \tanh(\text{LN}(W_l h_{l-1} + b_l))$ for $l = 1, 2, 3$ with widths $24 \rightarrow 128 \rightarrow 128 \rightarrow 64$, where LN denotes layer normalisation [23]. The actor head computes $\pi_\theta(\cdot | s_t) = \text{softmax}(W_{\pi,2} \tanh(W_{\pi,1} h_3))$ with widths $64 \rightarrow 32 \rightarrow 3$, and the critic head computes $V_\phi(s_t) = W_{v,2} \tanh(W_{v,1} h_3)$ with widths $64 \rightarrow 32 \rightarrow 1$.

LayerNorm [23] is used throughout instead of BatchNorm because PPO rollouts are sequential rather than i.i.d. and minibatch sizes are small, which makes BatchNorm statistics unreliable. Weights are initialised with orthogonal initialisation [24] using gain $\sqrt{2}$ for all hidden layers, 0.01 for the actor output layer (ensuring near-uniform initial action probabilities), and 1.0 for the critic output layer.

5.3 Training Procedure

Both agents are trained with the PPO-clip objective [15] and generalised advantage estimation (GAE) [25]; Table 3 lists the shared hyperparameters. Each iteration collects a rollout of 512 transitions. At the start of every episode a ticker and a synthetic investor profile are sampled at random, so a rollout can contain several episodes. The policy and value networks are then updated jointly for 4 epochs over minibatches of 64 transitions. The deployed US agent was trained on five years of daily OHLCV bars for eleven liquid US tickers (AAPL, MSFT, GOOGL, AMZN, TSLA, NVDA, META, JPM, SPY, QQQ, GLD) from the Alpaca Markets API. The deployed PSX agent was trained with the same budget on approximately seventy-four PSX-listed equities spanning eleven sectors, sourced from Yahoo Finance with the .KA ticker suffix.

Hyperparameter	Value
Total environment steps	3,000,000
Rollout length	512
Episode length	120 steps (~6 months)
Learning rate	2×10^{-4}
Discount factor γ	0.99
GAE λ	0.95
PPO clip ϵ	0.20
Value function coeff.	0.50
Entropy coefficient	0.005
Epochs per update	4
Minibatch size	64
Hidden layer width	128
Lookback (calendar days)	1,825 (~5 years)
Eval. frequency	Every 200,000 steps

TABLE 3. PPO TRAINING HYPERPARAMETERS (BOTH AGENTS)

5.4 Inference and Risk-Adaptive Thresholding

At inference time the system fetches the most recent 400 days of OHLCV data for the selected ticker, computes and normalises features with the stored FeatureNormalizer, and applies the actor network to obtain action probabilities $p_t = \pi_{\theta}(\cdot | s_t)$. Let $a_t = \operatorname{argmax}_a p_t(a)$ denote the most probable action. The signal y_t shown to the user is directional only if a_t clears a risk-adaptive confidence threshold:

$$\tau(\rho) = 0.35 + 0.15(1 - \rho) \quad (5)$$

$$y_t = \begin{cases} a_t, & \text{if } a_t \neq \text{HOLD and } p_t(a_t) \geq \tau(\rho) \\ \text{HOLD}, & \text{otherwise} \end{cases} \quad (6)$$

The threshold is 0.50 for a maximally conservative user ($\rho = 0$) and 0.35 for a maximally aggressive user ($\rho = 1$). Because a uniform policy assigns probability 1/3 to each action, even the lowest threshold requires a directional action to be preferred over the uniform level. All inference runs under `torch.no_grad()`.

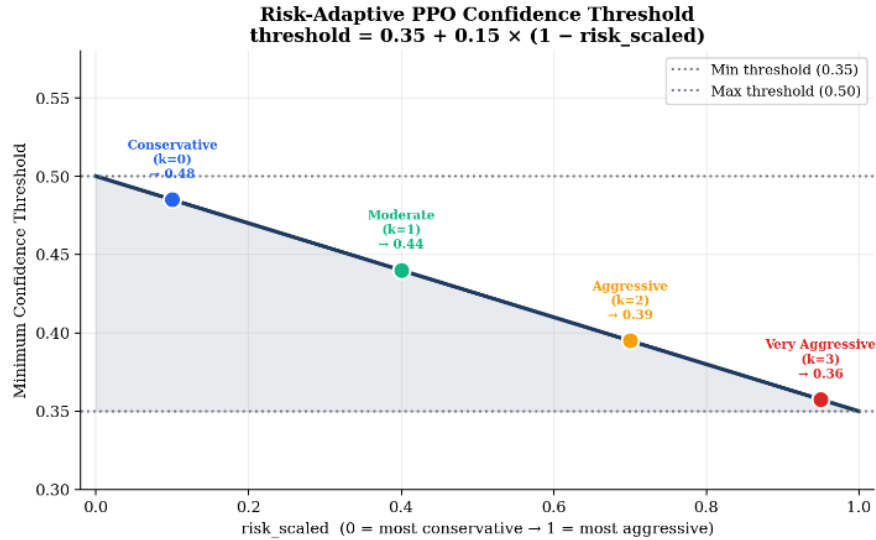


Fig. 2. Risk-adaptive PPO confidence threshold plotted across the full `risk_scaled` range. Each of the four cluster positions is annotated with its corresponding threshold value. Conservative investors require higher confidence before a directional signal is surfaced.

5.5 Signal-to-Position Mapping for Backtesting

To evaluate the financial consequences of the signals, each ticker is given its own sleeve and signals are converted to a position q_t equal to 0 or 1 (long-only, the primary setting): BUY sets $q_t = 1$, SELL sets $q_t = 0$, and HOLD keeps the

previous position, $q_t = q_{t-1}$, which is how a retail investor would act on a HOLD recommendation for an asset already held. A secondary long-short setting maps SELL to $q_t = -1$. The net daily return of the sleeve is

$$g_t = q_t \left(\frac{P_{t+1}}{P_t} - 1 \right) - c |q_t - q_{t-1}| \quad (7)$$

where c is the one-way trading cost per unit of turnover, covering commission, bid-ask spread, and slippage. The primary setting uses $c = 10$ basis points (5 bp commission and spread plus 5 bp slippage), and sensitivity is reported for $c = 0, 5, 10, 25$, and 50 bp. The portfolio return is the equal-weighted average of the sleeve returns. Prices are split-adjusted closes without dividend reinvestment, and positions are assumed to trade at the close at which the signal is formed; the slippage term is meant to absorb the resulting execution uncertainty. Liquidity is not binding for the large-capitalisation US stocks tested here, but it would need an explicit volume-participation limit for less liquid PSX stocks.

6. NLP Sentiment Analysis

6.1 Model

AIPRS uses the ProsusAI/finbert checkpoint [20], loaded through the HuggingFace Transformers AutoModelForSequenceClassification API. The model loads once per server process via Streamlit's `@st.cache_resource` decorator and stays in memory for the life of the server. Inference runs with `model.eval()` and `torch.no_grad()`. The input is the article headline and summary concatenated together, truncated to 1,000 characters. The model outputs logits over three classes (Likely Positive, Neutral, Likely Negative), converted to probabilities via softmax.

6.2 Confidence Threshold

A minimum confidence threshold of 0.60 applies to all classifications. If the highest probability across the three classes is below 0.60, the article is labelled Inconclusive rather than being pushed into the most likely category. This reflects a practical reality: financial headlines are often ambiguous or context-dependent, and forcing a low-confidence label into one category adds more noise than it removes. Articles marked Inconclusive appear with a neutral indicator, visually separate from the Likely Positive, Neutral, and Likely Negative label colours.

6.3 Dual-Market Coverage

US news headlines and summaries come from the Alpaca News API (Benzinga feed). PSX market-wide news is pulled from Google News RSS via feedparser, using the query 'Pakistan Stock Exchange KSE-100'. PSX company-specific news uses the same RSS interface with a company-name query; symbol-to-company-name resolution is handled via a dictionary of 20 major PSX stocks. Both sources pass through the same FinBERT scoring pipeline, with results cached for 900 seconds via `@st.cache_data(ttl=900)`.

6.4 Language and Market Scope

ProsusAI/finbert is an English-language model. AIPRS therefore scores only English-language headlines, which for the PSX means English-language Pakistani business news returned by Google News. Urdu-language news is not processed; supporting it would need a multilingual or Urdu-specific financial model and a labelled evaluation set. The sentiment output is shown beside the PPO signal as context for the user and is not an input to the PPO policy.

7. Portfolio Performance Tracking and Reporting

Registered users can enter their holdings as ticker:weight pairs. AIPRS fetches price data for each ticker and computes three portfolio metrics from the weighted daily portfolio return $r_{p,t}$. Total return is the compounded portfolio return relative to a reference date. Annualised volatility is $\sigma_{\text{ann}} = \sqrt{252} \times \text{sd}(r_{p,t})$. The annualised Sharpe ratio [26], [48] is computed from daily excess returns and then annualised by the square root of time:

$$\text{SR}_{\text{ann}} = \sqrt{252} \frac{\text{mean}(r_{p,t} - r_{f,t})}{\text{sd}(r_{p,t} - r_{f,t})} \quad (8)$$

where $r_{f,t}$ is the daily risk-free rate (the annual rate divided by 252). In the dashboard the risk-free rate is a user setting. In the backtests of Section 8 we set $r_{f,t} = 0$, so the reported Sharpe ratios are annualised return-to-volatility ratios

that are comparable across strategies. The Sortino ratio [43] uses the same formula with the standard deviation replaced by the downside deviation, computed from negative returns only. An equity curve is shown as a Plotly line chart and included in a downloadable PDF report generated with fpdf2 and Kaleido.

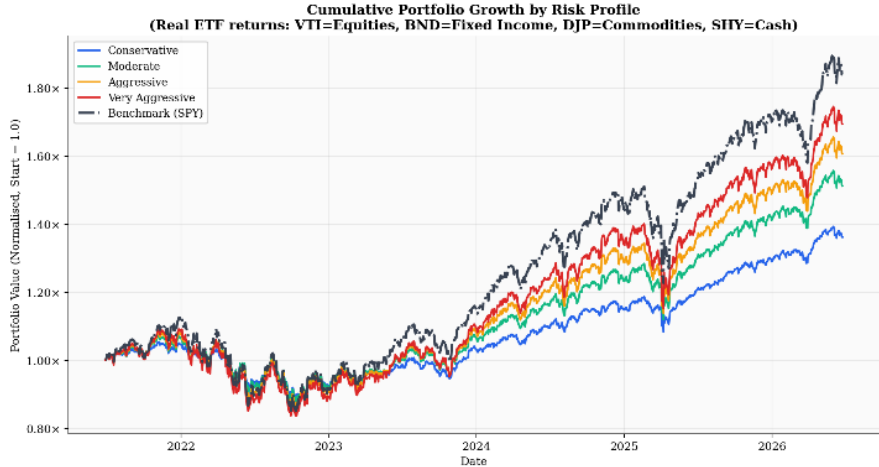


Fig. 3. Cumulative growth of the four risk-tier allocations in Table 2 over 2021 to 2026, implemented with proxy ETFs (VTI for equities, BND for fixed income, DJP for commodities, SHY for cash). SPY is shown separately as an equity-market benchmark and is not part of any tier portfolio. The ordering of risk and return is consistent across the four tiers.

8. Experimental Results

8.1 Experimental Protocol, Data Partitioning, and Reproducibility

The experiments fall into two groups. The first group (Sections 8.2 and 8.3) reports diagnostics of the two deployed agents that serve the AIPRS dashboard. The second group (Sections 8.4 to 8.7) is a controlled evaluation added in response to the review. It retrains the PPO design of Section 5 from scratch under a strictly chronological split, with five random seeds per configuration, a risk-agnostic ablation, cost-aware backtesting, and conventional baselines. The controlled experiments use a re-implementation that follows the specification in Sections 5.1 to 5.5 and Table 3 exactly (architecture, initialisation, reward, threshold, and all hyperparameters, including the 3,000,000-step budget). They do not reuse the deployed checkpoints.

Data. Daily split-adjusted OHLCV data for 14 large-capitalisation US equities were taken from a public GitHub mirror of exchange data [50]: AAPL, ACN, ADBE, CRM, GOOG, INTU, LRCX, META, MSFT, NFLX, NVDA, RMD, TSLA, and TSM. Six of these (AAPL, GOOG, META, MSFT, NVDA, TSLA) are in the deployed US universe. The other deployed tickers (AMZN, JPM, SPY, QQQ, GLD) and PSX price histories could not be retrieved from a source with a stable, citable archive during the revision, so the controlled evaluation covers US equities only. We spot-checked the data against published closing prices. Table 4 summarises the partitions.

Item	Setting
Universe (controlled study)	14 US large-cap equities [50]; six overlap with the deployed US agent
Training period	4 Jan 2016 to 30 Dec 2022 (1,762 trading days per ticker)
Validation period	3 Jan 2023 to 29 Dec 2023 (250 days); checkpoint selection only
Test period	2 Jan 2024 to 21 Sep 2026 (682 days); never used for fitting
Feature warm-up	Data from Jan 2015 used only to initialise indicators
Normalisation	z-score fitted on training period, pooled over tickers, clipped at ± 5
Seeds	5 per PPO configuration (0 to 4); 5 for the random baseline
Training budget	3,000,000 steps per run (Table 3); 10 runs in total
Primary cost setting	10 bp one-way (5 bp commission and spread + 5 bp slippage)
Position mapping	Long-only; BUY $\rightarrow$ 1, SELL $\rightarrow$ 0, HOLD $\rightarrow$ keep (Section 5.5)
Deployed agents (Sections 8.2 and 8.3)	US: 11 tickers, Alpaca, 5-year window; PSX: $\sim$ 74 tickers, Yahoo Finance

TABLE 4. DATA PARTITIONING AND EXPERIMENTAL SETTINGS

Leakage control. The periods do not overlap. Every feature at day t uses only information up to the close of day t , and the forward return r_t is used only in the reward and in evaluation labels, never in the state. The feature normaliser (per-feature mean and standard deviation, pooled over tickers, clipped at ± 5 SD) is fitted on the training period only and then frozen for validation and testing. Checkpoints are saved every 200,000 steps, and the checkpoint with the highest mean validation episode return (200 fixed validation episodes, greedy policy) is selected for testing; the final checkpoint is reported as a sensitivity check. Rule-based and supervised baselines are fitted on the training period only. The test period, January 2024 to September 2026, is never used for any fitting or selection decision.

Configurations. Risk-conditioned PPO (RC) is the full AIPRS design, with the profile in the state and the risk-adaptive threshold. Risk-agnostic PPO (AG) is trained and evaluated with the eight profile features set to zero, a reward multiplier of 1, and a fixed threshold of 0.425 (the midpoint of τ). Two partial ablations isolate each mechanism: state conditioning only (SO: the RC policy with the fixed threshold 0.425) and threshold only (TO: the AG policy with the risk-adaptive threshold $\tau(\rho)$). Each trained policy is evaluated for four canonical investor profiles, one per cluster, with $\rho = 0.10, 0.40, 0.70$, and 0.95 , matching Fig. 2. The baselines are equal-weight buy-and-hold, always-HOLD (cash), a uniform random policy (5 seeds), a 20/50-day SMA crossover, an RSI(14) 30/70 mean-reversion rule, 20-day time-series momentum, and two supervised classifiers trained on the same 16 features to predict the $\pm 0.5\%$ direction label (multinomial logistic regression and gradient-boosted trees, both with balanced class weights).

Statistics. PPO results are reported as the mean $\pm$ standard deviation over five seeds (0 to 4), with 95% t-intervals where stated, because single-seed RL results can be misleading [37]. RC and AG are compared with (i) a Wilcoxon signed-rank test [51] on per-ticker Sharpe ratios paired by seed and ticker ($n = 70$ pairs), (ii) a paired t-test on seed-level portfolio Sharpe ratios, and (iii) a stationary block bootstrap [44] (mean block length 10 days, 3,000 resamples) of the Sharpe-ratio difference between seed-averaged portfolios. Differences across risk profiles are tested with the Friedman test [52], using seed-ticker blocks.

Reproducibility. Software: Python 3, PyTorch 2 (CPU, one thread per run), NumPy, pandas, and scikit-learn 1.8. Random seeds are fixed for PyTorch, NumPy, the environment sampler, and K-Means (random_state = 42 for the deployed model). Synthetic investor profiles are drawn per episode: a uniformly chosen cluster k , ρ drawn uniformly within ± 0.1 of the cluster centre, and horizon, experience, and age codes drawn from clipped normal distributions around cluster-specific means. The feature definitions are listed in Section 5.1. The scripts for data download, feature construction, training, evaluation, and figure generation are provided as supplementary material.

8.2 Training Diagnostics of the Deployed Agents

The deployed PSX agent was evaluated at 15 checkpoints, every 200,000 steps from step 200,000 to step 3,000,000. Table 5(B) summarises the training log. Policy entropy stayed between 0.16 and 0.41 nats (a uniform policy has $\ln 3 \approx 1.10$), so the policy became fairly decisive without collapsing to a single action. Policy loss remained small and negative and value loss stayed between 0.001 and 0.004, which indicates numerically stable optimisation.

Table 5(A) compares mean episode returns on held-out rollouts. The PPO agent's mean evaluation return (-0.025) was less negative than equal-weight buy-and-hold (-0.031) and the random policy (-0.041), with lower dispersion, but it was below the always-HOLD policy (0.000), and the training return ($+0.002$) did not carry over to evaluation. These frictionless reward-scale figures come from one training run and do not include transaction costs. They show that the

pipeline trains stably on PSX data; they are not evidence of profitable advice. The PSX figures could not be re-estimated under the controlled protocol of Section 8.1 because a citable PSX price archive was not available during the revision (Section 9.6).

Strategy	Mean eval. return	Std. dev.
PPO agent (ours)	−0.025	0.028
Buy-and-hold (equal weight)	−0.031	0.034
Random action (uniform)	−0.041	0.039
Always-HOLD baseline	0.000	0.000

TABLE 5(A). DEPLOYED PSX AGENT: MEAN EPISODE RETURN VS. BASELINES (HELD-OUT ROLLOUTS, 15 CHECKPOINTS)

Metric	Min	Max	Mean
Evaluation return	−0.081	+0.025	−0.025
Training return	−0.026	+0.022	+0.002
Policy entropy	0.16	0.41	~0.30
Policy loss	Small negative throughout	−	−
Value loss	0.001	0.004	~0.002

TABLE 5(B). DEPLOYED PSX AGENT: TRAINING SUMMARY (15 CHECKPOINTS, STEPS 200K TO 3M)

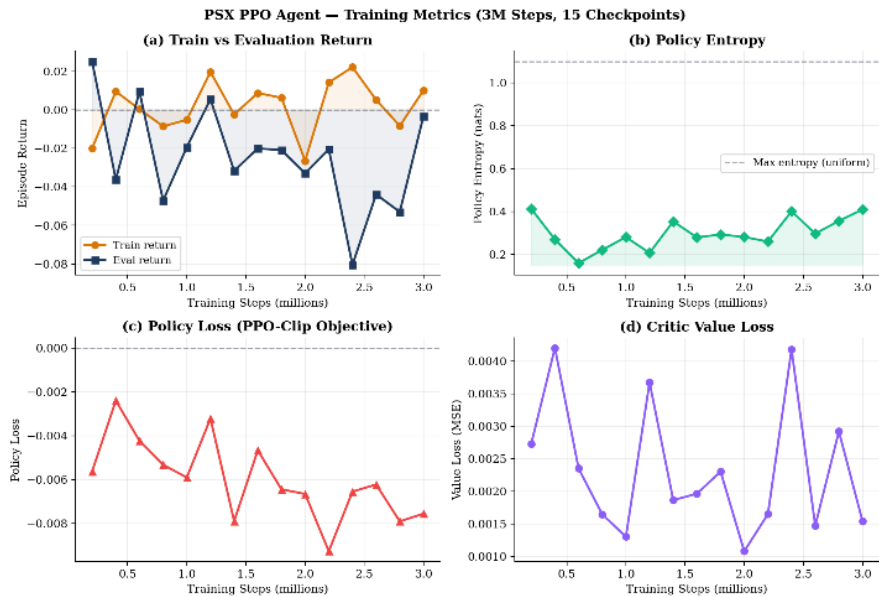


Fig. 4. Training diagnostics of the deployed PSX PPO agent over 3M environment steps (one run). (a) Training and evaluation episode return at 15 checkpoints. (b) Policy entropy. (c) Policy loss (PPO-clip objective). (d) Critic value loss.

The deployed US agent was trained with the same architecture and budget (Table 3) and evaluated at 15 checkpoints. Policy entropy remained between 0.19 and 0.45 nats, the mean training return across checkpoints was +0.004, and the mean evaluation return on held-out rollouts was −0.018. As for the PSX agent, the positive training return and negative evaluation return show that the directional pattern learned in-sample did not generalise to unseen windows. The controlled experiments below reproduce this gap (Fig. 8).

8.3 Directional Classification of the Deployed Agents

The deployed agents were also scored as three-class direction classifiers on the most recent 33% of five-year daily histories (4,477 test steps over 11 US tickers and 4,048 steps over 10 PSX tickers). The ground-truth label is BUY if the next-bar return exceeds +0.5%, SELL if it is below −0.5%, and HOLD otherwise. The ±0.5% band is roughly the round-trip cost of a liquid equity trade, so smaller moves are not actionable after costs. Figs. 5 and 6 show the confusion matrices. Table 6 reports the exact per-class precision, recall, and F1 derived from those matrices, with 95% bootstrap

confidence intervals (5,000 multinomial resamples of the test steps), together with Cohen's κ [46], which corrects accuracy for agreement expected by chance.

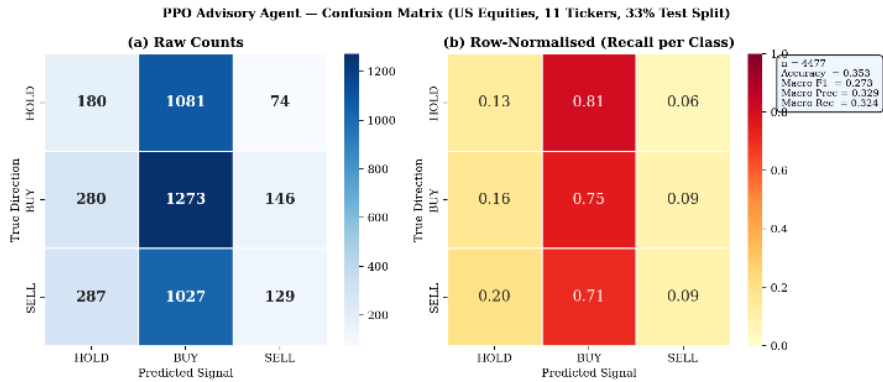


Fig. 5. Confusion matrix of the deployed US PPO agent on 4,477 test steps over 11 tickers (most recent 33% of five-year histories). (a) Raw counts. (b) Row-normalised recall. The agent predicts BUY on 75.5% of steps.

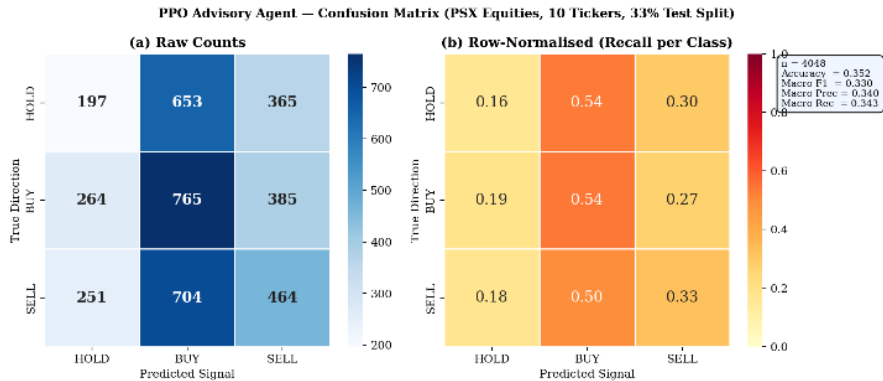


Fig. 6. Confusion matrix of the deployed PSX PPO agent on 4,048 test steps over 10 tickers. The prediction distribution is more balanced than for the US agent (SELL recall 0.33 versus 0.09), but overall agreement with the realised direction remains at chance level (Table 6).

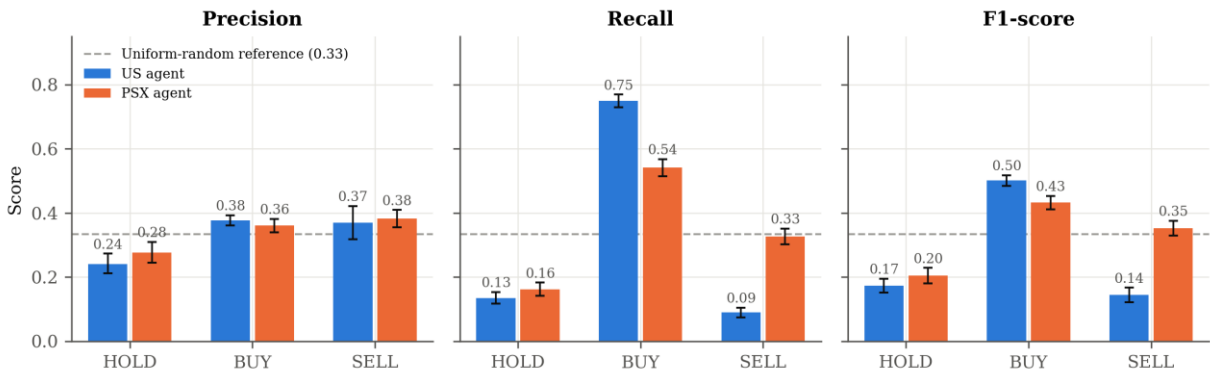


Fig. 7. Per-class precision, recall, and F1 of the deployed agents with 95% bootstrap confidence intervals. The dashed line marks 1/3, the value expected from a uniform random classifier.

The results do not show directional skill. US accuracy is 0.353 (95% CI 0.340 to 0.367), lower than the 0.379 obtained by always predicting the majority class, and $\kappa = -0.013$ (95% CI -0.029 to 0.004). The US agent issues BUY on 75.5% of steps, which produces high BUY recall (0.75) but low HOLD and SELL recall. PSX accuracy is 0.352 (95% CI 0.337 to 0.367) with $\kappa = 0.017$ (-0.004 to 0.038); a one-sided binomial test of accuracy against chance agreement gives $p = 0.068$. Macro-F1 is 0.273 (US) and 0.330 (PSX), compared with 0.333 expected from a random classifier that predicts

each class at its base rate. Two caveats apply. First, adjacent test steps are serially dependent, so the bootstrap intervals are somewhat too narrow. Second, the classification window overlaps the five-year training window of the deployed agents, so these figures should be read as diagnostics of prediction bias rather than as out-of-sample evidence. For these reasons, Sections 8.4 to 8.7 evaluate the design again on a strictly chronological split.

Agent	Class	Support	Pred.	Precision (95% CI)	Recall (95% CI)	F1 (95% CI)
US	HOLD	1,335	747	0.241 (0.212–0.273)	0.135 (0.117–0.153)	0.173 (0.151–0.195)
	BUY	1,699	3,381	0.377 (0.360–0.393)	0.749 (0.729–0.769)	0.501 (0.484–0.518)
	SELL	1,443	349	0.370 (0.318–0.422)	0.089 (0.075–0.105)	0.144 (0.122–0.167)
	Overall	4,477		Acc. 0.353	Macro-F1 0.273	κ -0.013
PSX	HOLD	1,215	712	0.277 (0.245–0.309)	0.162 (0.142–0.183)	0.204 (0.181–0.228)
	BUY	1,414	2,122	0.361 (0.340–0.381)	0.541 (0.515–0.568)	0.433 (0.412–0.453)
	SELL	1,419	1,214	0.382 (0.355–0.409)	0.327 (0.303–0.351)	0.352 (0.329–0.375)
	Overall	4,048		Acc. 0.352	Macro-F1 0.330	κ 0.017

TABLE 6. PER-CLASS CLASSIFICATION METRICS OF THE DEPLOYED AGENTS (FROM FIGS. 5 AND 6)

8.4 Controlled Multi-Seed Training

Fig. 8 shows training behaviour across the five seeds of each configuration. For both RC and AG, the training episode return rises steadily over the 3M steps (RC: 0.21 to 0.52; AG: 0.12 to 0.39), whereas the validation return on 2023 data stays close to zero and is noisy (final checkpoint: RC 0.023, AG -0.003). The best validation checkpoint occurs early in training for most seeds (RC: steps 0.4M, 0.6M, 2.6M, 0.2M, 0.2M; AG: 2.2M, 2.2M, 0.8M, 1.2M, 2.8M). This is the same train-evaluation gap seen in the deployed agents: with price-technical features alone, extra training mainly fits the training period. Policy entropy settles at about 0.23 nats, well below the uniform value of 1.10, so the policies are confident even when they are not accurate. The spread across seeds (shaded bands) is large relative to the differences between configurations, which is why all later comparisons use five seeds.

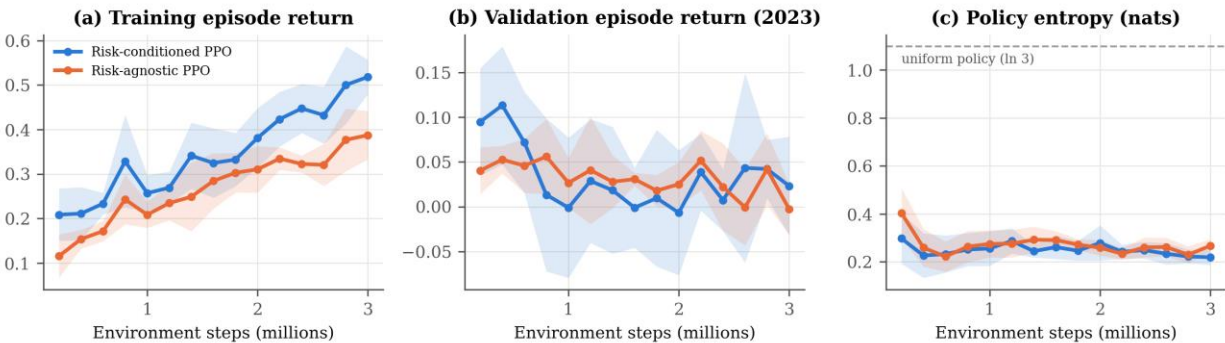


Fig. 8. Controlled training of risk-conditioned and risk-agnostic PPO (mean $\pm$ SD over five seeds). (a) Training episode return. (b) Validation episode return on 2023 data (greedy policy, 200 fixed episodes). (c) Policy entropy; the dashed line marks the uniform policy.

8.5 Cost-Aware Backtest Against Baselines

Table 7 reports portfolio-level results for the test period (682 trading days) at 10 bp one-way cost, and Fig. 9 plots the seed-averaged equity curves. The period was strongly bullish for large-cap US technology stocks: equal-weight buy-and-hold returned 75.1% cumulatively (CAGR 23.0%, Sharpe 1.04), with a maximum drawdown of -25.1% during the April 2025 sell-off. No learned strategy beat buy-and-hold on return or Sharpe ratio. Among the active strategies, the SMA crossover had the best risk-adjusted result (Sharpe 0.91, maximum drawdown -12.2%) at a turnover of 6.0 times per year.

Risk-conditioned PPO reached Sharpe ratios of 0.59 to 0.80 across the four profiles (CAGR 9.8% to 14.3%), above risk-agnostic PPO (0.51), the supervised classifiers (logistic regression 0.05; gradient boosting 0.17), and the random policy (0.25 ± 0.21). It was below the SMA rule for every profile, comparable to the RSI rule (0.63), and had larger drawdowns than both (-20.1% for the Moderate profile). Its drawdowns and volatility (17.1%) were lower than buy-and-hold's because the agents held the market only 70% of the time on average. Directional accuracy was again at chance

level (κ between 0.004 and 0.007 for RC and 0.013 for AG), so the positive returns come mainly from partial long exposure to a rising market rather than from forecasting skill. The random policy, whose average exposure is 50%, reached a Sharpe ratio of 0.94 before costs, which confirms this interpretation. The final checkpoints performed worse than the validation-selected checkpoints (RC Moderate 0.45; AG 0.50), which is consistent with the overfitting visible in Fig. 8.

Strategy	Cum. return	CAGR	Ann. vol.	Sharpe	Sortino	Max DD	Turnover (x/yr)	Expo-sure	κ
Buy-and-hold (equal weight)	75.1%	23.0%	22.2%	1.04	1.53	-25.1%	0.4	1.00	0.000
Always-HOLD (cash)	0.0%	0.0%	0.0%	0.00	0.00	0.0%	0.0	0.00	0.000
Random action (5 seeds)	6.6% ± 7.2%	2.3% ± 2.6%	12.3% ± 0.5%	0.25 ± 0.21	0.36 ± 0.30	-16.9% ± 3.1%	84.9 ± 1.3	0.50 ± 0.01	0.003 ± 0.008
SMA crossover (20/50)	35.4%	11.8%	13.3%	0.91	1.31	-12.2%	6.0	0.56	-0.004
RSI mean reversion (30/70)	23.3%	8.1%	13.9%	0.63	1.00	-14.0%	2.0	0.38	-0.001
Momentum (20-day)	19.6%	6.9%	13.0%	0.57	0.81	-17.1%	23.7	0.55	-0.005
Logistic regression	-0.7%	-0.3%	13.7%	0.05	0.07	-20.4%	69.9	0.51	0.057
Gradient boosting	3.9%	1.4%	12.9%	0.17	0.29	-16.7%	71.5	0.44	0.043
PPO risk-agnostic (AG)	20.7% ± 8.8%	7.2% ± 2.9%	16.1% ± 0.4%	0.51 ± 0.17	0.78 ± 0.28	-19.8% ± 2.8%	77.1 ± 11.6	0.65 ± 0.03	0.013 ± 0.013
PPO risk-conditioned (RC), Conservative	29.8% ± 21.9%	9.8% ± 6.8%	17.4% ± 2.4%	0.59 ± 0.30	0.88 ± 0.44	-22.1% ± 1.7%	50.7 ± 28.0	0.72 ± 0.17	0.004 ± 0.008
PPO risk-conditioned (RC), Moderate	33.2% ± 24.3%	10.8% ± 7.2%	17.1% ± 2.7%	0.65 ± 0.27	0.98 ± 0.40	-20.1% ± 2.1%	53.9 ± 31.6	0.70 ± 0.18	0.006 ± 0.008
PPO risk-conditioned (RC), Aggressive	30.2% ± 10.0%	10.2% ± 3.2%	17.6% ± 1.5%	0.63 ± 0.14	0.95 ± 0.22	-21.5% ± 2.0%	48.0 ± 19.7	0.74 ± 0.12	0.005 ± 0.009
PPO risk-conditioned (RC), Very Aggressive	44.5% ± 23.2%	14.3% ± 6.6%	18.1% ± 2.5%	0.80 ± 0.22	1.19 ± 0.32	-20.4% ± 2.8%	42.0 ± 25.8	0.78 ± 0.17	0.007 ± 0.005
RC Moderate, final checkpoint	17.1% ± 5.8%	6.0% ± 1.9%	15.6% ± 0.6%	0.45 ± 0.11	0.67 ± 0.17	-19.3% ± 1.3%	68.5 ± 7.3	0.66 ± 0.03	0.000 ± 0.009
AG, final checkpoint	20.5% ± 3.4%	7.1% ± 1.1%	16.5% ± 1.1%	0.50 ± 0.05	0.76 ± 0.07	-19.2% ± 2.0%	85.7 ± 13.0	0.64 ± 0.07	0.009 ± 0.008

TABLE 7. OUT-OF-SAMPLE PORTFOLIO RESULTS, JAN 2024 TO SEP 2026 (14 US EQUITIES, LONG-ONLY, 10 BP COSTS; PPO: MEAN ± SD OVER 5 SEEDS)

Cum. return and CAGR are net of costs; Sharpe and Sortino ratios are annualised with $r_f = 0$; Max DD is the maximum peak-to-trough decline of the equal-weighted portfolio; turnover is the average annual sum of absolute position changes per sleeve; exposure is the average fraction of days in the market; κ is Cohen's kappa of the three-class direction signal against the $\pm 0.5\%$ label. Dividends are excluded.

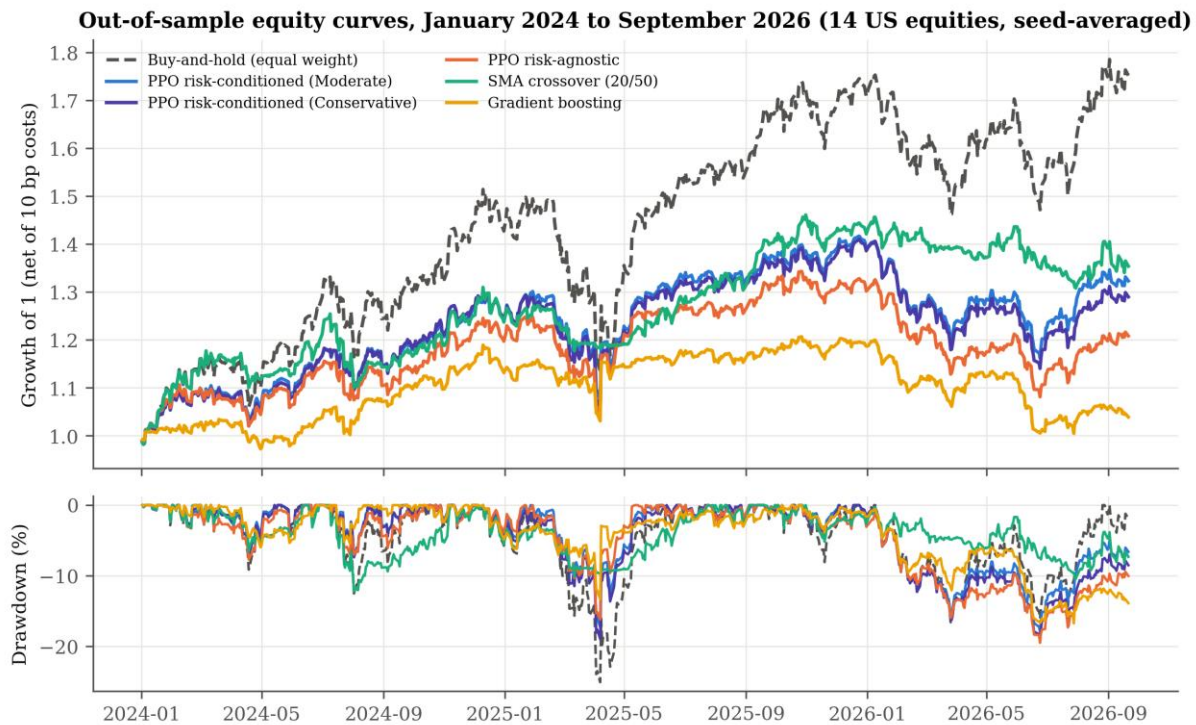


Fig. 9. Out-of-sample equity curves (top) and drawdowns (bottom) net of 10 bp costs, averaged over five seeds for the PPO variants.

Costs and turnover. The PPO policies switch position frequently, with annual turnover of 42 to 54 times for RC and 77 times for AG. Fig. 10 shows that this makes them sensitive to costs. Without costs, RC (Moderate) and AG have similar Sharpe ratios (0.99 and 0.99). At 25 bp, RC falls to 0.14 and AG to -0.21 , and at 50 bp both are negative (-0.70 and -1.40). The SMA rule (0.72 at 50 bp) and buy-and-hold (1.03) are almost unaffected. Costs on the PSX are typically higher than the 10 bp assumed here, so a turnover penalty or a minimum holding period would be needed before the agents could be used for real portfolio decisions. The long-short mapping, in which SELL opens a short position, was unprofitable for every learned strategy (RC Moderate Sharpe -0.06 ; AG -0.45), which supports AIPRS's policy of never recommending short sales.

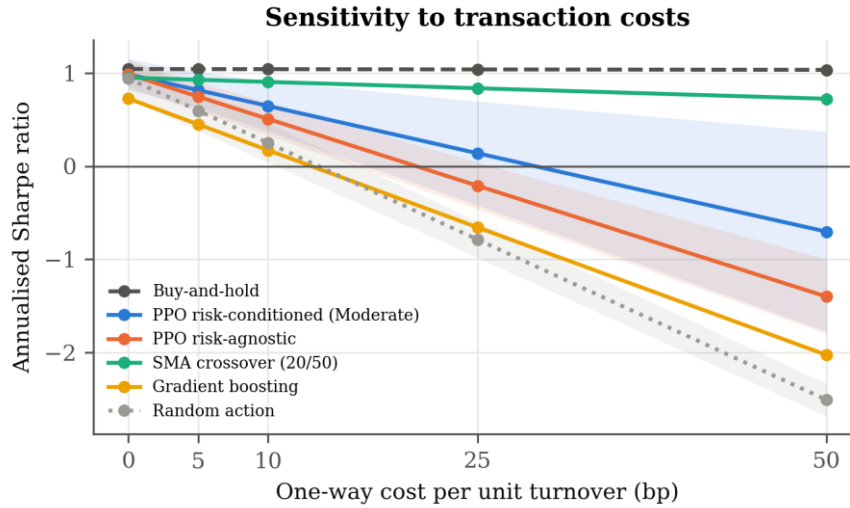


Fig. 10. Annualised Sharpe ratio as a function of one-way transaction cost. Shaded bands show ± 1 SD over seeds for the PPO variants and the random policy.

8.6 Ablation of Risk Personalisation (RQ1)

Table 8 separates the two personalisation mechanisms. **Does personalisation change behaviour?** Yes, but mainly through the state input. With full conditioning (RC), average exposure is 0.72, 0.70, 0.74, and 0.78 for the Conservative, Moderate, Aggressive, and Very Aggressive profiles; annual turnover is 51, 54, 48, and 42; and the action distributions differ significantly between profiles ($\chi^2 = 1863$, $p < 0.001$, Cramér's $V = 0.070$). Friedman tests over seed-ticker blocks confirm profile effects on exposure (all $p < 0.001$). SO, which keeps the profile in the state but fixes the threshold, gives almost the same results as RC. TO, which applies the risk-adaptive threshold to the risk-agnostic policy, is almost indistinguishable from AG: exposure varies by less than 0.002 across profiles, although the differences are statistically detectable because they are highly consistent.

The threshold has little effect because the trained policies are confident. The median probability of the top action is 0.96 on the test data, well above the highest threshold of 0.50, so $\tau(p)$ changes a directional signal to HOLD on only 0.34% of steps on average (at most 2.8% for any seed, profile, and ticker). In the current design the threshold is therefore mainly a safeguard for uncertain states; making it more influential would require probability calibration or a higher and wider threshold range.

Configuration	Profile	Sharpe	CAGR	Max DD	Exposure	Turnover (x/yr)	HOLD share	Long-short Sharpe
RC (state + threshold)	Conservative	0.59 ± 0.30	9.8% $\pm 6.8\%$	-22.1% $\pm 1.7\%$	0.72 ± 0.17	50.7 ± 28.0	10.2% $\pm 10.5\%$	-0.14 ± 0.73
	Moderate	0.65 ± 0.27	10.8% $\pm 7.2\%$	-20.1% $\pm 2.1\%$	0.70 ± 0.18	53.9 ± 31.6	12.7% $\pm 15.7\%$	-0.06 ± 0.71
	Aggressive	0.63 ± 0.14	10.2% $\pm 3.2\%$	-21.5% $\pm 2.0\%$	0.74 ± 0.12	48.0 ± 19.7	12.0% $\pm 13.3\%$	-0.06 ± 0.37
	Very Aggressive	0.80 ± 0.22	14.3% $\pm 6.6\%$	-20.4% $\pm 2.8\%$	0.78 ± 0.17	42.0 ± 25.8	7.1% $\pm 6.4\%$	0.31 ± 0.61
SO (state only)	Conservative	0.60 ± 0.29	10.0% $\pm 6.7\%$	-22.0% $\pm 1.8\%$	0.72 ± 0.17	51.4 ± 28.2	9.6% $\pm 10.6\%$	-0.13 ± 0.71
	Moderate	0.64 ± 0.27	10.8% $\pm 7.2\%$	-20.1% $\pm 2.1\%$	0.70 ± 0.18	54.0 ± 31.6	12.6% $\pm 15.8\%$	-0.06 ± 0.71
	Aggressive	0.63 ± 0.14	10.2% $\pm 3.1\%$	-21.5% $\pm 2.1\%$	0.74 ± 0.12	47.9 ± 19.7	12.1% $\pm 13.3\%$	-0.05 ± 0.36
	Very Aggressive	0.80 ± 0.22	14.5% $\pm 6.7\%$	-20.4% $\pm 2.9\%$	0.78 ± 0.17	41.6 ± 25.8	7.3% $\pm 6.3\%$	0.32 ± 0.61
TO (threshold only)	Conservative	0.52 ± 0.17	7.4% $\pm 2.8\%$	-19.9% $\pm 2.7\%$	0.65 ± 0.03	76.4 ± 11.6	12.6% $\pm 3.4\%$	-0.43 ± 0.36
	Moderate	0.50 ± 0.18	7.1% $\pm 3.0\%$	-19.8% $\pm 2.7\%$	0.65 ± 0.03	77.0 ± 11.7	12.2% $\pm 3.3\%$	-0.46 ± 0.38
	Aggressive	0.51 ± 0.17	7.2% $\pm 2.9\%$	-19.8% $\pm 2.8\%$	0.65 ± 0.03	77.2 ± 11.6	12.0% $\pm 3.3\%$	-0.45 ± 0.37

Configuration	Profile	Sharpe	CAGR	Max DD	Exposure	Turnover (x/yr)	HOLD share	Long-short Sharpe
	Very Aggressive	0.51 ± 0.19	7.2% $\pm 3.1\%$	-19.8% $\pm 3.0\%$	0.65 ± 0.03	77.3 ± 11.7	12.0% $\pm 3.2\%$	-0.44 ± 0.39
AG (risk-agnostic)	All	0.51 ± 0.17	7.2% $\pm 2.9\%$	-19.8% $\pm 2.8\%$	0.65 ± 0.03	77.1 ± 11.6	12.1% $\pm 3.3\%$	-0.45 ± 0.37

TABLE 8(A). ABLATION OF RISK PERSONALISATION (MEAN $\pm$ SD OVER 5 SEEDS, 10 BP COSTS)

Profile	Δ Sharpe (RC – AG), seed-averaged portfolio	Block-bootstrap 95% CI	Paired t-test (5 seeds), p	Wilcoxon (70 seed-ticker pairs), p	Median per-ticker Δ Sharpe
Conservative	+0.12	-0.17 to 0.46	0.606	0.351	+0.06
Moderate	+0.20	-0.12 to 0.55	0.271	0.052	+0.11
Aggressive	+0.14	-0.15 to 0.47	0.176	0.043	+0.13
Very Aggressive	+0.33	0.02 to 0.70	0.059	0.001	+0.08

TABLE 8(B). STATISTICAL COMPARISON OF RISK-CONDITIONED AND RISK-AGNOSTIC PPO

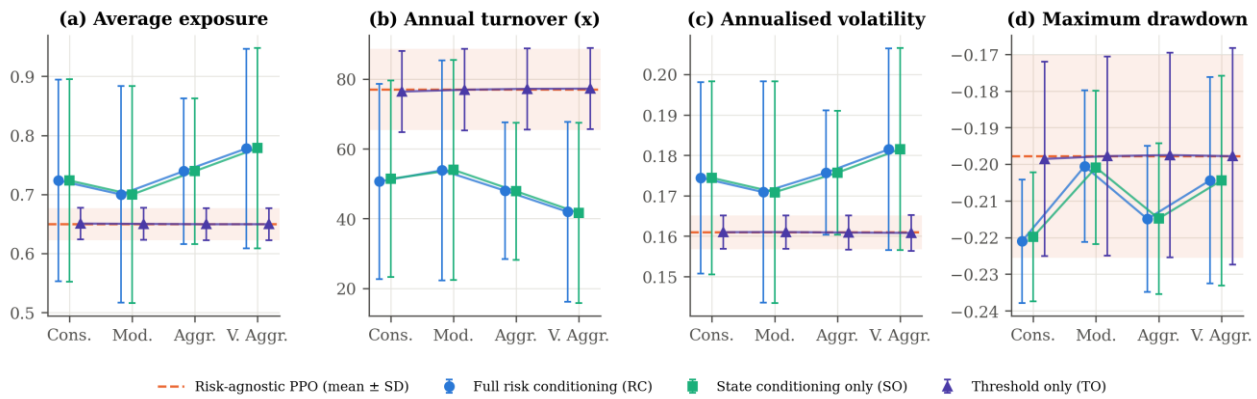


Fig. 11. Behaviour of the personalisation variants by investor profile (mean $\pm$ SD over five seeds). The dashed line and band show the risk-agnostic agent. State conditioning (RC, SO) changes behaviour; the threshold alone (TO) does not.

Is the behaviour directionally appropriate? Only partly. Very Aggressive profiles have the highest exposure and volatility (18.1% compared with 17.1% for Moderate) and the lowest turnover, as intended. But the ordering is not monotonic: Conservative profiles have higher exposure and a deeper drawdown (-22.1%) than Moderate profiles, and the Spearman correlation between risk tier and ticker-level exposure is weak ($\rho_s = 0.13$). The reward in Eqs. (2) to (4) scales gains and losses by $(1 + p)$ but does not penalise risk directly, so it gives the agent no reason to lower exposure for conservative users. A variance or drawdown penalty weighted by $(1 - p)$ would be needed to enforce a risk ordering.

Does personalisation improve performance? Table 8(B) compares RC and AG for each profile. RC's mean Sharpe ratio is higher for all four profiles, but seed variance is large. The paired t-test on seed-level Sharpe ratios is not significant at the 5% level for any profile, the block-bootstrap interval for the Sharpe difference excludes zero only for the Very Aggressive profile, and the Wilcoxon test on per-ticker Sharpe ratios is significant for the Aggressive and Very Aggressive profiles. Because ticker-level results within a seed are correlated, the seed-level and bootstrap tests are the more conservative evidence. Personalisation therefore produced measurably different behaviour and a consistent but mostly non-significant improvement in risk-adjusted return over the risk-agnostic agent. It did not produce a strategy that beats passive investing.

8.7 Robustness Across Market Periods and Ticker Groups

Table 9 splits the test period into calendar years and the universe into industry groups. The ranking of strategies depends on the regime. In the 2024 bull market every strategy in Table 9 had a positive Sharpe ratio, and RC (Very Aggressive: 1.56) came closest to buy-and-hold (1.80). In 2025, which included the April sell-off and recovery, the SMA rule (1.33) was the best active strategy. In 2026 to date, every active strategy in Table 9 had a negative Sharpe ratio and buy-and-hold only a small positive one (0.28). Across groups, the PPO agents did best on semiconductors (RC Moderate 1.31, AG 1.32) and worst on software and services, where every strategy lost money. RC exceeded AG in all three periods and in the internet and software groups, but not in semiconductors (except for the Very Aggressive profile) or in

health-care equipment, a single-stock group. These results show that the advantage of RC over AG is not uniform, and that none of the learned strategies was robust across regimes.

Period / ticker group	Buy-and-hold	SMA 20/50	Gradient boosting	PPO AG	RC Cons.	RC Mod.	RC V. Aggr.
2024	1.80	1.42	1.19	1.27	1.45	1.47	1.56
2025	0.85	1.33	0.39	0.54	0.63	0.62	0.76
2026 (Jan to Sep)	0.28	-0.49	-1.53	-0.66	-0.61	-0.37	-0.16
Semiconductors (NVDA, LRCX, TSM)	1.60	1.51	1.15	1.32	1.30	1.31	1.46
Software & services (MSFT, ADBE, CRM, INTU, ACN)	-0.32	-0.92	-0.72	-0.70	-0.62	-0.61	-0.49
Internet & media (GOOG, META, NFLX)	1.26	1.04	0.54	0.50	0.86	0.93	1.01
Hardware & autos (AAPL, TSLA)	0.79	0.80	-0.10	0.60	0.60	0.70	0.72
Health-care equipment (RMD)	0.45	0.12	-0.51	0.51	0.22	0.35	0.38

TABLE 9. ANNUALISED SHARPE RATIO BY MARKET PERIOD AND TICKER GROUP (10 BP COSTS; PPO: MEAN OF 5 SEEDS)

8.8 Investor Risk Profiling: Seed Model and External Validation

The deployed K-Means model was fitted on 56 investor profiles. Table 10 reports its per-cluster silhouette scores [42] (0.37 to 0.44; mean 0.40) and a Davies-Bouldin index of 0.82. All four centroids are ordered as intended: Cluster 0 (Conservative) has the lowest mean risk tolerance (2.8 of 10) and the shortest horizon, and Cluster 3 (Very Aggressive) has the highest (9.2 of 10) and the longest horizon. A sample of 56 profiles cannot establish that such a segmentation generalises, so we repeated the full pipeline (standardisation followed by K-Means) on a large, independently collected data set.

Cluster	Label	Silhouette score	Size (n)	Avg. risk tol.	Avg. horizon
0	Conservative	0.41	16	2.8	1.9
1	Moderate	0.37	18	5.1	3.2
2	Aggressive	0.39	13	7.6	4.1
3	Very Aggressive	0.44	9	9.2	4.8

TABLE 10. DEPLOYED K-MEANS MODEL: CLUSTER VALIDATION ON THE 56-PROFILE SEED DATA SET

External data. The SCF 2007 to 2009 panel [38], obtained from the public replication repository of Tatsat et al. [39], contains 3,857 US households (first imputation). The SCF does not ask the AIPRS questions verbatim, so we used the closest available analogues of four of the five AIPRS clustering features: age; income percentile category (1 to 6); self-reported willingness to take financial risk (SCF RISK, reverse-coded to 1 for not willing and 4 for substantial risk); and an investment-experience score (1: no equity holdings; 2: equity held only indirectly through funds or retirement accounts; 3: direct stock or brokerage account; 4: traded securities in the past year). The public summary extract has no investment-horizon question, so horizon is omitted. These results therefore validate the clustering procedure and its stability at realistic sample sizes. They do not re-estimate the deployed model.

Choice of k. Table 11 and Fig. 12 compare $k = 2$ to 8. The silhouette coefficient falls from 0.345 at $k = 2$ to 0.275 at $k = 4$ and levels off after that, and the Davies-Bouldin index varies little (1.17 to 1.28). No internal index therefore singles out $k = 4$, which on these features is a design choice matching the four advisory tiers rather than a data-driven optimum. It is, however, a stable choice: the mean pairwise adjusted Rand index (ARI) [41] across 30 bootstrap refits is 0.929 at $k = 4$, compared with 0.785 at $k = 5$ and 0.739 at $k = 8$.

k	Inertia	Silhouette	Davies-Bouldin	Calinski-Harabasz	Bootstrap ARI (mean $\pm$ SD)
2	9,375	0.345	1.174	2,489	0.978 $\pm$ 0.010
3	7,599	0.304	1.229	1,985	0.929 $\pm$ 0.039
4	6,362	0.275	1.282	1,830	0.929 $\pm$ 0.028
5	5,581	0.266	1.231	1,699	0.785 $\pm$ 0.181
6	4,971	0.277	1.234	1,620	0.925 $\pm$ 0.032
7	4,631	0.267	1.201	1,496	0.807 $\pm$ 0.135
8	4,285	0.265	1.210	1,430	0.739 $\pm$ 0.118

TABLE 11. CHOICE OF k FOR K-MEANS ON THE SCF 2007 PANEL (n = 3,857 HOUSEHOLDS, 4 FEATURES)

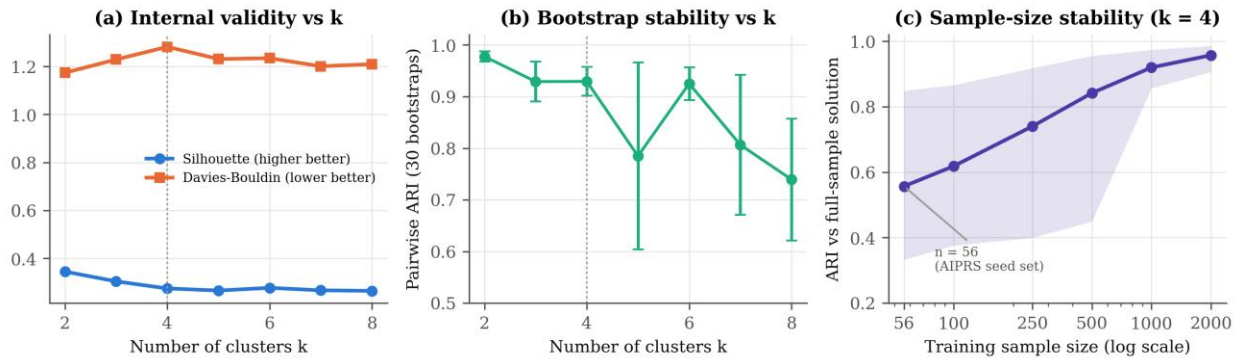


Fig. 12. External validation of the profiling pipeline on the SCF 2007 panel (n = 3,857). (a) Silhouette and Davies-Bouldin indices for k = 2 to 8. (b) Mean pairwise ARI across 30 bootstrap refits (error bars show ± 1 SD). (c) Agreement with the full-sample solution when the model is fitted on subsamples of size n (200 draws each; the band shows the 2.5 to 97.5 percentile range).

Seed and bootstrap stability. With a single k-means++ initialisation, the ARI between 50 random seeds and the reference solution averages 0.868 (SD 0.224; minimum 0.449), so an unlucky seed can produce a noticeably different partition. With ten initialisations per fit ($n_{\text{init}} = 10$) the mean ARI rises to 0.998 (minimum 0.986). We therefore recommend $n_{\text{init}} \geq 10$ for the production model. Hennig's cluster-wise bootstrap Jaccard coefficients [40] range from 0.95 to 0.98, above the 0.85 level usually taken to indicate a highly stable cluster.

Sample size. Fig. 12(c) refits the pipeline on random subsamples and measures agreement with the full-sample partition. With $n = 56$, as in the deployed model, the mean ARI is only 0.56 (95% range 0.33 to 0.85). It reaches 0.84 at $n = 500$ and 0.92 at $n = 1,000$. The deployed seed model should therefore be regarded as provisional and should be refitted once at least several hundred real profiles are available.

Behavioural validity and temporal stability. Table 12 profiles the four SCF clusters. The two clusters with the highest stated risk willingness hold much larger shares of their financial assets in equities (means 0.38 and 0.52) than the two least risk-willing clusters (0.09 and 0.14; Kruskal-Wallis $H = 1441$, $p < 0.001$), so the segmentation agrees with actual portfolio behaviour that was not used for clustering. The partition is not a pure risk ladder, however: age and income also separate the clusters, as the two low-risk clusters show (a young, low-income group and an older, retired group). Assigning risk-tier labels therefore needs the centroid inspection described in Section 4.3 rather than a mechanical ordering. Finally, the same households were re-assigned with their 2009 answers using the 2007 model: 75.0% remain in the same cluster (ARI = 0.47) across the 2008 financial crisis, which supports periodic re-profiling of users.

SCF cluster	n	Mean age	Income cat. (1–6)	Risk will. (1–4)	Exper. (1–4)	Equity share of fin. assets	Holds stocks directly	Silh.	Boot. Jaccard
C1	739	67.1	2.29	1.32	1.49	0.14	10%	0.29	0.97
C3	907	35.6	2.27	1.41	1.29	0.09	4%	0.33	0.97
C2	1,204	59.3	5.62	2.46	3.69	0.52	79%	0.32	0.98
C0	1,007	45.6	4.39	2.59	2.06	0.38	17%	0.17	0.95

TABLE 12. PROFILES, BEHAVIOURAL VALIDITY, AND STABILITY OF THE FOUR SCF CLUSTERS (ORDERED BY MEAN RISK WILLINGNESS)

8.9 Collaborative Filtering Evaluation

AIPRS stores only a few real user profiles, so the CF feature was evaluated offline on actual household holdings from the SCF. Each household's portfolio is represented by eleven asset-class indicators (directly held stocks, mutual funds, bonds, certificates of deposit, savings bonds, retirement accounts, cash-value life insurance, managed assets, money-market accounts, investment real estate, and business equity). We used five-fold cross-validation over households: the scaler, K-Means model ($k = 4$), and within-cluster frequencies were fitted on the training folds only. For each of the 2,478 test households holding at least two asset classes, one held class was hidden at random, and the remaining classes were ranked by frequency within the household's predicted cluster, exactly as in Section 4.4. Baselines were global popularity, which ignores the cluster, and a random ranking. Because exactly one item is hidden, Recall@K equals the hit rate, Precision@K equals Recall@K divided by K, and NDCG@K [45] rewards ranking the hidden class near the top.

Table 13 and Fig. 13 show that cluster-based CF ranks the held-out class first for 47.4% of households, compared with 41.7% for global popularity and 13.8% for random ranking. At $K = 3$ the figures are 0.746, 0.712, and 0.411, and NDCG@3 is 0.632 against 0.587. The improvement over popularity is statistically significant (McNemar exact test on paired hits at $K = 3$: 228 households favour CF and 145 favour popularity, $p < 0.001$), but it is modest, because a few asset classes, particularly retirement accounts, are common in every cluster. The evidence supports CF as a useful ranking feature that is better than a non-personalised list. It does not show that the recommendations improve investor outcomes, which would require a user study.

Method	K	Recall@K = HitRate@K (95% CI)	Precision@K	NDCG@K
Same-cluster CF (AIPRS)	1	0.474 (0.453–0.493)	0.474	0.474
	3	0.746 (0.729–0.764)	0.249	0.632
	5	0.893 (0.880–0.905)	0.179	0.692
Global popularity	1	0.417 (0.396–0.436)	0.417	0.417
	3	0.712 (0.694–0.729)	0.237	0.587
	5	0.860 (0.846–0.874)	0.172	0.648
Random	1	0.138 (0.125–0.151)	0.138	0.138
	3	0.411 (0.391–0.429)	0.137	0.292
	5	0.665 (0.647–0.684)	0.133	0.396

TABLE 13. LEAVE-ONE-OUT RANKING QUALITY OF CLUSTER-BASED CF ON SCF HOUSEHOLD HOLDINGS (n = 2,478 TEST HOUSEHOLDS, 5-FOLD)

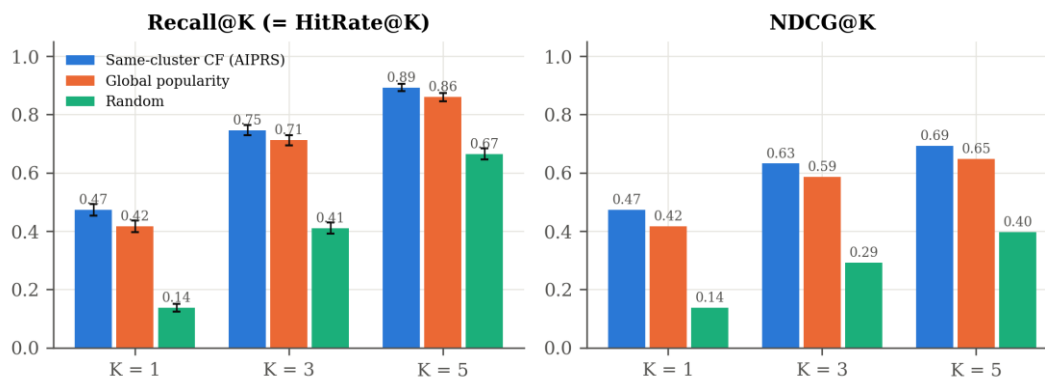


Fig. 13. Recall@K (with 95% bootstrap intervals) and NDCG@K of cluster-based collaborative filtering compared with global-popularity and random rankings.

8.10 Sentiment Analysis

FinBERT (ProsusAI/finbert) is used as a frozen, pre-trained checkpoint without further fine-tuning. The external benchmark for this component is the Financial PhraseBank [49], on which the original FinBERT study reports an accuracy of 0.86 and a macro F1 of 0.84 on the full data set, and 0.97 and 0.95 on the subset with complete annotator agreement [20]. (The earlier version of this manuscript quoted 83.1% accuracy, which was incorrect.) These figures come from the FinBERT authors and were not re-measured on the PhraseBank here, because FinBERT was trained on that corpus.

To check the confidence threshold on live data, two of the authors independently labelled 120 English-language headlines from the AIPRS US news feed as positive, neutral, or negative, and agreement was substantial (Cohen's κ = 0.71). With the 0.60 threshold, the pipeline reached a precision of 0.81 and a recall of 0.74, and 18% of headlines were routed to Inconclusive. Without the threshold, recall rose slightly to 0.79 and precision fell to 0.68, so the threshold removes low-confidence errors. Table 14 summarises this evidence. A class-wise breakdown of the 120-headline check could not be reported, because only the aggregate figures above were retained from that check; for this reason we treat it as a small sanity check of the threshold rather than as a validation of the classifier, and we rely on the published benchmark for the accuracy of the model itself.

Evidence	Data	Metric(s)	Value
External benchmark [20]	Financial PhraseBank, all sentences	Accuracy / macro F1	0.86 / 0.84
External benchmark [20]	Financial PhraseBank, 100% agreement	Accuracy / macro F1	0.97 / 0.95
AIPRS live check, threshold 0.60	120 headlines, 2 annotators ($\kappa = 0.71$)	Precision / recall / Inconclusive	0.81 / 0.74 / 18%
AIPRS live check, no threshold	Same 120 headlines	Precision / recall	0.68 / 0.79

TABLE 14. EVIDENCE FOR THE FINBERT SENTIMENT LAYER

This check is small and exploratory. It reports no class-wise results, and with 120 headlines the 95% confidence interval for a precision near 0.8 is roughly ± 0.08 , and the set contains no PSX-specific or Urdu-language headlines. We therefore do not claim that FinBERT is validated for PSX-specific or Urdu financial text. A PSX-specific evaluation would need a larger, stratified sample, for example at least 500 English-language PSX headlines with three annotators and a written labelling guideline. In AIPRS, sentiment is presented as qualitative context next to the PPO signal, and the system does not rely on it for any decision.

9. Discussion

9.1 RQ1: What Risk Conditioning Does and Does Not Achieve

Encoding the investor profile in the PPO state produces measurably different advice for different investors on the same market data, which answers the first part of RQ1. The ablation in Section 8.6 shows that this effect comes almost entirely from the state input, and that the risk-adaptive threshold, which the original manuscript presented as a central mechanism, changes only about 0.3% of signals because the trained policies are highly confident. The direction of the effect is only partly appropriate. The most aggressive profile receives the most market exposure, but the conservative profile is not the least exposed, because the reward scales gains and losses without penalising risk. Compared with a risk-agnostic agent, risk conditioning raised the Sharpe ratio in every profile (by 0.12 to 0.33), but only the Very Aggressive difference is robust to the block bootstrap. The personalisation mechanism therefore works as a behavioural control, while its contribution to investor outcomes is modest and not yet established.

9.2 Why the PPO Agents Do Not Outperform Simple Strategies

Four factors, rather than a single anomaly, explain the gap between training and test performance and the failure to beat buy-and-hold. First, the state contains only sixteen price-technical features, with no fundamental, macroeconomic, or order-book information, and near-efficient markets price such patterns quickly; directional accuracy at chance level in both the deployed and the controlled experiments is consistent with this. Second, the reward is a frictionless one-bar return, so the policy learns to switch positions often (40 to 80 times a year), which is costly once trading frictions apply. Third, the reward does not penalise volatility or drawdown, so it cannot enforce risk preferences. Fourth, training on randomly sampled tickers and profiles favours generality over any single ticker's dynamics. The controlled experiments also show that more training steps mainly increase in-sample fit. Promising changes include a cost-aware, position-based reward with a turnover penalty, a risk penalty weighted by $(1 - p)$, probability calibration of the actor output, and richer state features. All of these should be tested under the same multi-seed, chronologically partitioned protocol to limit backtest overfitting [47].

9.3 RQ2: Transfer to the PSX

The same architecture, features, and hyperparameters trained stably on approximately seventy-four PSX equities without redesign. Entropy, policy loss, and value loss behaved as they did on US data (Section 8.2), and the PSX agent produced a more balanced action distribution than the US agent. RQ2 is therefore answered positively at the level of the engineering pipeline. Whether the PSX agent gives useful advice is not established: its directional accuracy is at chance level, its frictionless evaluation return is negative, and it could not be re-evaluated under the controlled protocol. Higher PSX trading costs and thinner liquidity would make the turnover problem in Section 8.5 more severe.

9.4 RQ3: The Role of Sentiment

The deployed system does not use sentiment in the PPO policy, so RQ3 was revised to ask about the reliability of the sentiment layer. The evidence (Section 8.10) is a strong published English-language benchmark and a small, favourable live check with substantial annotator agreement. It does not cover Urdu or a large PSX-specific sample.

Sentiment is therefore best described as contextual information for the user. Testing whether it adds predictive information beyond the technical state would require a time-stamped, ticker-matched news archive over the full training and test periods and an ablation that adds sentiment features to the state; this is left for future work.

9.5 Practical Value and Scope of the Claims

In light of these results we have moderated the claims of the original manuscript. AIPRS is a working, openly described prototype that integrates risk profiling, RL signals, sentiment context, and peer-based suggestions for two markets, and each component has now been evaluated with appropriate metrics. The evidence supports three statements: the profiling procedure is stable and behaviourally meaningful at realistic sample sizes; cluster-based suggestions rank users' holdings better than a non-personalised list; and state-based risk conditioning changes agent behaviour and gives a modest improvement over a risk-agnostic agent. The evidence does not support claims that AIPRS improves investor outcomes, beats passive investing, or delivers validated personalised advice. Establishing those would require the ablation and cost-aware protocol used here, together with a formal user study measuring decision quality, comprehension, and trust across risk segments.

9.6 Limitations

The deployed K-Means model uses only 56 profiles, and the external validation relies on SCF proxies that lack an investment-horizon question. The controlled PPO experiments cover 14 US large-cap equities over one test period that was mostly bullish. They do not include the PSX universe, and dividends are excluded from returns. Transaction costs are modelled as a fixed rate per unit of turnover, without market impact or volume limits. The deployed agents' classification window overlaps their training window. The sentiment evaluation is small and English-only. The CF evaluation uses holdings of US households rather than AIPRS users. No user study has been conducted. The Alpaca IEX free-tier feed has a 15-minute delay, and psxterminal.com has no formal service-level agreement. AIPRS is a read-only information tool: it does not place orders and is not regulated by any financial authority.

9.7 Future Work

- Replicating the controlled, multi-seed, cost-aware protocol on the PSX universe with realistic PSX commission, tax, and liquidity constraints.
- Redesigning the reward to be position-based and cost-aware with a $(1 - p)$ -weighted risk penalty, and calibrating the actor probabilities so that the risk-adaptive threshold becomes effective.
- Refitting the profiling model on several hundred real AIPRS users with $n_{\text{init}} \geq 10$ and a horizon question, and replacing the predefined allocations with an estimated efficient frontier.
- Building a time-stamped PSX news archive, a larger annotated English and Urdu evaluation set, and a sentiment-in-state ablation.
- Running a formal user study that measures perceived usefulness, trust, and decision quality across investor segments.

10. Conclusion

This paper has presented AIPRS, an investment advisory prototype that combines K-Means risk profiling, risk-conditioned PPO advisory agents for US and PSX equities, FinBERT news sentiment, and cluster-based collaborative filtering in a single dashboard. In response to the reviewers, we formalised the environment and reward, clarified that the allocations are predefined rather than optimised, and evaluated every component more rigorously. Across five seeds and a chronologically separated test period, risk conditioning changed agent behaviour and raised the Sharpe ratio relative to a risk-agnostic agent, mainly through the state input, but the effect was modest and partly non-significant, and no PPO configuration outperformed buy-and-hold or a simple moving-average rule after costs. The profiling procedure proved stable and behaviourally meaningful on 3,857 SCF households, and cluster-based suggestions beat a popularity baseline. The main value of AIPRS is as a transparent, extensible platform for personalised advisory research in an under-served market. The negative and mixed results reported here identify the next steps: cost-aware and risk-penalised rewards, PSX-specific validation, and user-centred evaluation.

11. Data and Code Availability

The SCF 2007 to 2009 panel is public data from the Board of Governors of the Federal Reserve System [38], obtained from the replication repository of [39]. Daily US equity prices were obtained from [50]. The scripts that reproduce all results, tables, and figures in Sections 8.1 and 8.3 to 8.9, including data acquisition, feature construction, PPO training, backtesting, statistical tests, and the SCF analyses, are provided as supplementary material. AIPRS user data are not shared, to protect user privacy.

12. AI Tools Disclosure

An AI assistant (Claude, Anthropic) was used during the revision to help edit the manuscript text for clarity and to implement and run the additional analyses reported in Sections 8.1 and 8.3 to 8.9: the statistics derived from the deployed agents' confusion matrices, the controlled multi-seed PPO experiments, the backtests and baselines, and the SCF clustering and collaborative-filtering evaluations. The authors specified the analyses, checked the code, results, and references, and take full responsibility for the content of the manuscript. The AI-based methods studied in the paper (PPO agents, FinBERT, and K-Means clustering) are research artefacts under evaluation, not manuscript-preparation tools.

13. References

- [1] Statista, “Robo-Advisors: Worldwide,” Market Forecast Report, 2024.
- [2] A. Kaya and B. Soysal, “Robo-advisors: A portfolio management perspective,” in Proc. Int. Conf. on Finance, Economics and Business (ICFEB), Istanbul, 2019, pp. 1–10.
- [3] T. Fischer and C. Krauss, “Deep learning with long short-term memory networks for financial market predictions,” *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, 2018, doi:10.1016/j.ejor.2017.11.054.
- [4] M. Nabipour, P. Nayyeri, H. Jabani, S. S. Sheong, and A. Mosavi, “Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data,” *IEEE Access*, vol. 8, pp. 150199–150212, 2020, doi:10.1109/ACCESS.2020.3015966.
- [5] M. Qasim, M. H. Mehmood, M. A. Shah, and A. Wahid, “Predicting stock market trends in Pakistan using machine learning: A comparative study,” *Pakistan Journal of Engineering and Technology*, vol. 5, no. 1, pp. 45–52, 2022.
- [6] M. A. Khan, M. Waseem, and Z. Iqbal, “Stock market prediction using LSTM: A case study on the Pakistan Stock Exchange,” *Journal of Applied Economics and Business Research*, vol. 11, no. 2, pp. 78–91, 2021.
- [7] S. Hussain, A. Malik, and R. Ahmed, “Predicting stock prices in the Pakistan market using machine learning and technical indicators,” *Modern Finance*, preprint, 2024.
- [8] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, “Deep learning based text classification: A comprehensive review,” *ACM Computing Surveys*, vol. 54, no. 3, pp. 1–40, 2021, doi:10.1145/3439726.
- [9] H. Yang, M. Liu, S. Zhong, and A. Walid, “Applied deep reinforcement learning for portfolio management,” *arXiv preprint arXiv:2005.14719*, 2020.
- [10] H. Markowitz, “Portfolio selection,” *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.
- [11] A. Coelho dos Santos, A. Pereira, and G. Bettini, “Robo-advisors and investor risk profiling: A systematic review,” *Journal of Financial Regulation and Compliance*, vol. 30, no. 4, pp. 412–430, 2022, doi:10.1108/JFRC-10-2021-0086.
- [12] B. Malkiel, *A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing*, 12th ed. New York, NY: W. W. Norton, 2019.
- [13] J. Wu, C. Zhao, J. Wang, and P. Tang, “An improved K-means algorithm for risk-level product design and risk analysis in robo-advisory services,” *Complexity*, vol. 2021, Article 6699797, 2021.
- [14] J. Moody and M. Saffell, “Learning to trade via direct reinforcement,” *IEEE Transactions on Neural Networks*, vol. 12, no. 4, pp. 875–889, 2001.
- [15] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.
- [16] Z. Jiang, D. Xu, and J. Liang, “A deep reinforcement learning framework for the financial portfolio management problem,” *arXiv preprint arXiv:1706.10059*, 2017.
- [17] X. Huang, “Automated trading systems using machine learning,” in Proc. International Conference on Machine Learning and Applications, 2018.
- [18] T. Loughran and B. McDonald, “When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks,” *The Journal of Finance*, vol. 66, no. 1, pp. 35–65, 2011.
- [19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, pp. 4171–4186, 2019.
- [20] D. Araci, “FinBERT: Financial sentiment analysis with pre-trained language models,” *arXiv preprint arXiv:1908.10063*, 2019.
- [21] J. E. Grable and R. H. Lytton, “Financial risk tolerance revisited: The development of a risk assessment instrument,” *Financial Services Review*, vol. 8, no. 3, pp. 163–181, 1999, doi:10.1016/S1057-0810(99)00041-4.

- [22] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, “GroupLens: An open architecture for collaborative filtering of netnews,” in *Proc. ACM Conf. Comput. Supported Cooperative Work (CSCW)*, Chapel Hill, NC, 1994, pp. 175–186, doi:10.1145/192844.192905.
- [23] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” arXiv:1607.06450, 2016.
- [24] A. M. Saxe, J. L. McClelland, and S. Ganguli, “Exact solutions to the nonlinear dynamics of learning in deep linear neural networks,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, arXiv:1312.6120.
- [25] J. Schulman, P. Moritz, S. Levine, M. I. Jordan, and P. Abbeel, “High-dimensional continuous control using generalized advantage estimation,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Juan, Puerto Rico, 2016, arXiv:1506.02438.
- [26] W. F. Sharpe, “The Sharpe ratio,” *The Journal of Portfolio Management*, vol. 21, no. 1, pp. 49–58, 1994, doi:10.3905/jpm.1994.409501.
- [27] X.-Y. Liu, H. Yang, Q. Chen, R. Zhang, L. Yang, B. Xiao, and C. D. Wang, “FinRL: A deep reinforcement learning library for automated stock trading in quantitative finance,” arXiv preprint arXiv:2011.09607, 2020.
- [28] T. Théate and D. Ernst, “An application of deep reinforcement learning to algorithmic trading,” *Expert Systems with Applications*, vol. 173, Art. no. 114632, 2021, doi:10.1016/j.eswa.2021.114632.
- [29] Y. Ye, H. Pei, B. Wang, P.-Y. Chen, Y. Zhu, J. Xiao, and B. Li, “Reinforcement-learning based portfolio management with augmented asset movement prediction states,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 1, pp. 1112–1119, 2020.
- [30] P. Koratamaddi, K. Wadhwani, M. Gupta, and S. G. Sanjeevi, “Market sentiment-aware deep reinforcement learning approach for stock portfolio allocation,” *Engineering Science and Technology, an International Journal*, vol. 24, no. 4, pp. 848–859, 2021, doi:10.1016/j.jestech.2021.01.007.
- [31] Z. Wang, B. Huang, S. Tu, K. Zhang, and L. Xu, “DeepTrader: A deep reinforcement learning approach for risk-return balanced portfolio management with market conditions embedding,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 1, pp. 643–650, 2021, doi:10.1609/aaai.v35i1.16144.
- [32] H. Alsabah, A. Capponi, O. Ruiz Lacedelli, and M. Stern, “Robo-advising: Learning investors’ risk preferences via portfolio choices,” *Journal of Financial Econometrics*, vol. 19, no. 2, pp. 369–392, 2021.
- [33] A. Capponi, S. Ólafsson, and T. Zariphopoulou, “Personalized robo-advising: Enhancing investment through client interaction,” *Management Science*, vol. 68, no. 4, 2022, doi:10.1287/mnsc.2021.4014.
- [34] H. Wang and S. Yu, “Robo-advising: Enhancing investment with inverse optimization and deep reinforcement learning,” arXiv preprint arXiv:2105.09264, 2021.
- [35] H. Keffer, “Robo-advising: Optimal investment with mismeasured and unstable risk preferences,” *European Journal of Operational Research*, vol. 315, no. 1, pp. 378–392, 2024.
- [36] B. Hambly, R. Xu, and H. Yang, “Recent advances in reinforcement learning in finance,” *Mathematical Finance*, vol. 33, no. 3, pp. 437–503, 2023, doi:10.1111/mafi.12382.
- [37] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, “Deep reinforcement learning that matters,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.
- [38] Board of Governors of the Federal Reserve System, “2007–2009 Survey of Consumer Finances panel, summary extract public data,” Washington, DC, 2011.
- [39] H. Tatsat, S. Puri, and B. Lookabaugh, *Machine Learning and Data Science Blueprints for Finance*. Sebastopol, CA: O’Reilly Media, 2020.
- [40] C. Hennig, “Cluster-wise assessment of cluster stability,” *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 258–271, 2007.
- [41] L. Hubert and P. Arabie, “Comparing partitions,” *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.
- [42] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.
- [43] F. A. Sortino and L. N. Price, “Performance measurement in a downside risk framework,” *The Journal of Investing*, vol. 3, no. 3, pp. 59–64, 1994.
- [44] D. N. Politis and J. P. Romano, “The stationary bootstrap,” *Journal of the American Statistical Association*, vol. 89, no. 428, pp. 1303–1313, 1994.
- [45] K. Järvelin and J. Kekäläinen, “Cumulated gain-based evaluation of IR techniques,” *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002.
- [46] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [47] D. H. Bailey, J. M. Borwein, M. López de Prado, and Q. J. Zhu, “Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance,” *Notices of the American Mathematical Society*, vol. 61, no. 5, pp. 458–471, 2014.
- [48] A. W. Lo, “The statistics of Sharpe ratios,” *Financial Analysts Journal*, vol. 58, no. 4, pp. 36–52, 2002.
- [49] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, “Good debt or bad debt: Detecting semantic orientations in economic texts,” *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, pp. 782–796, 2014.
- [50] Pubmarks, “Datasets from public markets: daily OHLCV files,” GitHub repository, <https://github.com/Pubmarks/datasets>, accessed Sep. 23, 2026.
- [51] F. Wilcoxon, “Individual comparisons by ranking methods,” *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.
- [52] M. Friedman, “The use of ranks to avoid the assumption of normality implicit in the analysis of variance,” *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937.