Enhancing Transparency and Interpretability in AI-Driven Algorithmic Trading

Nathan Portelli

Supervisor: Dr Vincent Vella

January, 2026

Submitted in partial fulfilment of the requirements for the degree of M.Sc. in Artificial Intelligence.

L-Università ta' Malta
Faculty of Information & Communication Technology

L-Università ta' Malta

www.um.edu.mt

First edition, Friday 30th January, 2026

To my beloved family

whose unconditional love and steadfast belief in me provided the foundation and strength necessary for this endeavour. Your encouragement was my constant source of motivation.

To my cherished friends

whose well-timed distractions, good humour, and unwavering support kept me grounded and provided the necessary relief during the most demanding times.

And to my dedicated supervisor

whose guidance, patience, and profound knowledge illuminated the path to understanding and made this achievement possible.

Acknowledgements

I am extremely grateful to Dr. Vincent Vella for his exceptional supervision, support, and continuous guidance throughout this project. I also extend my sincere thanks to my family and friends for their continued support, as well as all survey participants for their valuable time and insightful contributions.

Abstract

Algorithmic trading increasingly relies on AI-driven decision systems, yet opaque models limit trust and accountability. This study investigates how Reinforcement Learning (RL) can be made more transparent by combining a model-agnostic explainability framework with Large Language Model (LLM) based narrative synthesis. The framework comprises four layers linking trading behaviour to feature attribution and temporal dynamics, stability and regime sensitivity, policy surrogacy, and reward decomposition.

In Experiment 1, we apply state-of-the-art RL algorithms to constituents of the Dow Jones Industrial Average and evaluate performance using standard return and risk measures. We investigate modern explainability techniques across market regimes to characterise which indicators drive decisions, how stable attributions are over time, and how policies can be approximated by compact surrogate rules. The results indicate convergent feature drivers, expected masking behaviour, smoother attributions, indicating that explanations are temporally stable, and low-complexity surrogates with credible fidelity.

In Experiment 2 we extend these findings by investigating how explainability artefacts can be translated into grounded natural-language narratives. Explanations generated from the framework are assessed through automated text metrics and a human-centred study with participants at varying levels of trading experience. Our results show that structured prompting improves lexical quality and adherence to factuality constraints relative to a zero-shot baseline. An important finding is that participants view the combined visual–narrative explanations as clear, moderately trustworthy, and practically helpful, with narratives particularly useful for reconstructing the agents' reasoning and feature contributions.

The contributions of this work are threefold. First, the study introduces a unified, model-agnostic explainability framework for financial RL linking feature-level attributions to policy behaviour and realised rewards. Second, it proposes and validates a pipeline for grounded natural-language synthesis anchored in quantitative explainability outputs. Third, in daily DJIA trading and a small human study, it provides evidence that technical faithfulness and human-centred accessibility can be advanced together, offering a template for transparent decision support in finance.

Contents

List of Figures

List of Tables

List of Abbreviations

A2C Advantage Actor Critic 120
AI Artificial Intelligence 114
AIA Artificial Intelligence Act 97
AIM Accuracy on Important Features Masked by Reference Padding 105
ANN Artificial Neural Network 14
ARIMA Autoregressive Integrated Moving Average 7
AUM Accuracy on Unimportant Features Masked by Reference Padding 105
BLEU Bilingual Evaluation Understudy 121
CCI Commodity Channel Index 41
DDPG Deep Deterministic Policy Gradient 120
DJIA Dow Jones Industrial Average 120
DQN Deep Q-Learning 109
DX Directional Movement Index 41
EMH Efficient-Market Hypothesis 6
FinRL Financial Reinforcement Learning 120
HFT High-Frequency Trading 7
IG Integrated Gradients 95
LIME Local Interpretable Model-Agnostic Explanations 45
LLM Large Language Model 125
LSTM Long Short-Term Memory 23
MACD Moving Average Convergence/Divergence 41
METEOR Metric for Evaluation of Translation with Explicit Ordering 121
ML Machine Learning 54
NLP Natural Language Processing 36
OHLCV Open, High, Low, Close and Volume 38
PPO Proximal Policy Optimization 120

1 Introduction

1.1 Problem Definition

Financial markets are among the most dynamic and unpredictable environments in which AI can be applied (Islam et al. (2021); Théate and Ernst (2021); Yeo et al. (2023)). Trading decisions are made under uncertainty and shaped by complex interdependencies between macroeconomic events, technical signals, and investor behaviour (Benhamou et al. (2021); Kindleberger and Aliber (2011)). Within this setting, Reinforcement Learning (RL) has attracted growing attention for its ability to optimise sequential decision-making (Bai et al. (2024); Zou et al. (2023a)). Unlike supervised learning, which relies on static labelled datasets, RL agents adapt their strategies by interacting with market environments and updating their policies based on realised trading outcomes (Agarwal et al. (2021); Arsenault et al. (2024); Ferreira et al. (2021); Izzo (2022); Zou et al. (2023b)).

Despite this promise, the use of RL in finance is constrained by a central challenge: model opacity. The decision-making processes of RL agents involve high-dimensional state spaces, stochastic policies, and deep neural approximators, rendering their internal reasoning inaccessible to most users (Agarwal et al. (2021); Ali et al. (2023); Islam et al. (2021); Théate and Ernst (2021); Yeo et al. (2023); Çetin et al. (2023)). In financial contexts, where decisions affect capital allocation and portfolio risk, this black-box character is a critical barrier. Traders, investors, and regulators require not only optimisation performance but also visibility into why trading actions are taken (Arrieta et al. (2019); Quinn (2023); Théate and Ernst (2021); Weber et al. (2024); Yeo et al. (2023)). A model that cannot justify its buy, hold, or sell recommendations is unlikely to be adopted, particularly in regulated financial environments (Arrieta et al. (2019); Weber et al. (2024)).

The situation is further complicated by the nature of existing explainability techniques. Current XAI techniques tend to improve transparency at the model level, but fail to close the interpretability gap for human end-users (Jin et al. (2023); Rong et al. (2024)). Research in Explainable AI (XAI) has produced a wide array of methods, ranging from feature attribution, such as SHAP and Integrated Gradients, to policy visualisation, saliency

maps, and surrogate models. Although these tools can clarify model behaviour for researchers and practitioners, their outputs often remain technical and difficult for many stakeholders to interpret in practice. A SHAP summary plot may highlight the importance of indicators such as RSI, MACD, or trading volume, but still provide limited guidance to a trader deciding whether to enter or exit a position. Likewise, surrogate trees or attribution heatmaps may improve transparency without providing explanations that are readily understandable or actionable for financial professionals or regulators.

Opacity has regulatory implications. Under the EU Artificial Intelligence Act (AIA), the The European Parliament (2024) emphasises that high-risk AI systems must be transparent, accountable, and explainable (Regulation (EU) 2024/1689). Institutions deploying RL-based trading systems without robust interpretability mechanisms risk non-compliance and may erode investor confidence, particularly in retail markets where novice traders perceive algorithmic systems as inaccessible or untrustworthy (Arrieta et al. (2019); Quinn (2023); Théate and Ernst (2021); Weber et al. (2024); Yeo et al. (2023)). The dual challenge is not only to open the black box of RL but also to ensure that the explanations are meaningful and usable for diverse audiences and levels of expertise.

This research is positioned within this problem space, with the goal of making RL models for algorithmic trading more transparent, trustworthy, and accessible. It addresses both the technical dimension of model explainability and the human dimension of explanation accessibility, with the broader goal of contributing to financial trading systems that are effective, yet also understandable, accountable, and aligned with human-centred needs.

1.2 Motivation

This work is motivated by three factors. First, financial trading is increasingly shaped by algorithmic systems that automate decision-making and optimise returns. Within this space, RL offers a powerful framework for adapting to market volatility and changing regimes, but growing adoption amplifies concerns about opacity. Trust in financial systems is critical, as traders and institutions cannot rely on models whose behaviour cannot be scrutinised without risking financial losses and reputational damage.

Second, regulatory frameworks are evolving towards stricter expectations of transparency. Emerging legal instruments increasingly require high-risk AI systems, including those used in finance, to provide meaningful explanations of their decision-making processes. Without effective interpretability mechanisms, financial RL applications may face barriers to deployment or compliance, limiting their practical impact.

Third, there is a persistent usability gap in the outputs of current XAI methods. Techniques such as SHAP values, feature attributions, and surrogate models can clarify how an RL model functions, yet their outputs often remain technical and difficult for end-users to interpret. Traders may require clear narratives, regulators may demand audit trails that link decisions to measurable indicators, and non-expert investors may need intuitive descriptions that demystify algorithmic strategies rather than dense visualisations aimed at AI practitioners. Recent advances in Natural Language Processing (NLP), particularly Large Language Models (LLMs), offer an opportunity to bridge this gap by transforming technical XAI artefacts into natural language explanations that are accessible and actionable for diverse user groups. This potential motivates the integration of XAI and NLP in this project to address both the technical and human dimensions of interpretability.

1.3 Aims and Objectives

The overarching aim of this project is to examine the extent to which the transparency and interpretability of RL models in financial algorithmic trading can be improved. It does so by combining XAI techniques with NLP-driven explanations to address both the technical opacity of RL and the usability gap in current explainability methods. The specific objectives of the research are as follows:

1. Develop and implement a model-agnostic explainability framework to assess the extent to which transparency and interpretability can be improved for RL models in algorithmic trading.

2. Investigate whether LLM-based natural-language explanations improve human-centred interpretability when compared with traditional visual and numerical XAI outputs, using automated text-based metrics and a user study focused on non-expert users, particularly novice traders.

This dissertation prioritises novice traders because they are the stakeholder group most likely to be affected by the usability gap created by technical explainability outputs, and they face a higher risk of misinterpreting model behaviour when explanations are unclear. While expert traders and regulators are also relevant audiences, the empirical user evaluation focuses on novices to provide a stringent test of explanation accessibility within the scope of this project.

These objectives collectively frame the dual focus of the research, ensuring that RL models can be made technically interpretable through XAI, and that these interpretations can be translated into explanations meaningful to human users.

1.4 Proposed Solution

To address the challenges outlined above, this project proposes a two-stage research approach that integrates XAI and NLP techniques.

The first stage is dedicated to developing and implementing a model-agnostic explainability framework to enhance the transparency of RL models in algorithmic trading. In this stage, representative algorithms, namely PPO, DDPG, A2C and TD3, are trained on DJIA data spanning the period 2009-2021, following Liu et al. (2022b) as the baseline. Their outputs are subjected to a comprehensive suite of XAI methods, including SHAP (global, rolling), Integrated Gradients, GradientSHAP, saliency maps, decision tree surrogates, reward attribution, and trajectory analysis. These techniques are applied to uncover which features drive trading decisions, how explanations vary across temporal and market regimes, and how policies can be approximated in interpretable forms.

Crucially, the explainability methods are benchmarked using quantitative criteria, with full definitions and computation provided in Section 3.1.5.

Together, these layers of analysis form a structured, model-agnostic framework whose end goal is to deliver transparent, robust, and interpretable insight into the decision-making processes of RL agents in financial markets.

The second stage extends this framework into the human domain by integrating NLP-driven explanation generation. Using outputs from the previous experiment, LLMs generate natural language explanations tailored to expert and non-expert audiences. These are validated for fidelity to the underlying XAI outputs and evaluated through a user study involving participants without technical expertise in AI. Participants compare traditional explainability artefacts and narratives in terms of trust, clarity, cognitive load, and decision-making support. This experiment directly addresses the usability gap, testing whether NLP-enhanced explanations improve user understanding and confidence.

Together, these stages constitute a comprehensive solution that addresses both technical and human challenges of interpretability. By integrating XAI benchmarking with NLP-enhanced narratives, the framework seeks to bridge the gap between algorithmic opacity and human-centred transparency in financial trading.

1.5 Document Structure

This document is structured as follows. Chapter 2 provides background and reviews the literature on RL in algorithmic trading, XAI methods, and the emerging role of NLP in explanation generation. Chapter 3 details the experimental design, including data, model training, XAI integration, and the user study. Chapter 4 presents the empirical results

and evaluation of both experiments, combining quantitative metrics, qualitative feedback, and integrative discussion. Finally, Chapter 5 summarises the contributions, discusses limitations, and outlines directions for future work.

2 Background & Literature Overview

This chapter provides a detailed overview of the foundational and contemporary research that underpins this project. It begins with a discussion of financial markets and the evolution of algorithmic trading, followed by an examination of Reinforcement Learning (RL) as an emerging paradigm for decision-making in trading. The main focus of this review is the growing role of Explainable AI (XAI) in improving transparency within financial Artificial Intelligence (AI) systems and an exploration of how Natural Language Processing (NLP) can enhance the transparency and interpretability of these explanations for non-technical users.

2.1 Financial Markets and Algorithmic Trading

Financial markets serve as the backbone of the global economy, enabling capital allocation, risk transfer, and liquidity provision. Early organised exchanges, such as the Amsterdam Stock Exchange, which was established in the early seventeenth century, introduced tradable ownership through equity shares, laying the foundations for modern markets. Over time, markets expanded to include a wide array of instruments, including equities, bonds, derivatives, and commodities, and became global centres of speculation, investment, and hedging (Allen and Gale (2000); Kindleberger and Aliber (2011); Neal (1991)).

At the core of the market function lies the process of price discovery and liquidity provision. The Efficient-Market Hypothesis (EMH), first formalised by Fama (1970), proposes that market prices fully reflect available information, implying that persistent excess returns are unattainable. Subsequent research, however, has documented empirical anomalies such as momentum, mean reversion, and seasonal effects that challenge strict forms of efficiency and point to exploitable inefficiencies (Jegadeesh and Titman (1993); Lo and MacKinlay (1999)). These findings catalysed the development of quantitative and data-driven trading strategies, where systematic exploitation of such patterns became a central theme in modern trading systems.

2.1.1 From Quantitative Analysis to Algorithmic Trading

The late twentieth century marked a shift from manual to automated trading. For much of their history, trades were executed through brokers on exchange floors, with orders matched via open outcry. The digitisation of financial infrastructure and the emergence of electronic communication networks transformed this landscape, giving rise to algorithmic trading, whereby pre-programmed models execute orders based on data-driven rules (Harris (2003); Hasbrouck (2007)).

Early algorithmic strategies focused on order execution, employing methods such as Volume-Weighted Average Price (VWAP) and Time-Weighted Average Price (TWAP) to minimise market impact by slicing large orders into smaller transactions (Bialkowski et al. (2008)). As computational power and data availability expanded, algorithms evolved from static rule-based systems to adaptive, predictive models capable of responding to very short-horizon changes in market conditions. High-Frequency Trading (HFT) emerged as a defining feature of this evolution, exploiting small price discrepancies across venues to generate profit through speed and precision (Aldridge (2013); Cartea et al. (2015)).

Although algorithmic trading has contributed to improved liquidity and tighter spreads, it has also introduced systemic vulnerabilities. Events such as the 2010 'Flash Crash', where the DJIA experienced a rapid and temporary collapse of nearly 1,000 points, underscored the risks associated with automated feedback loops and high-speed strategies (Easley et al. (2011); Kirilenko et al. (2017)). In response, regulators introduced circuit breakers and enhanced reporting requirements to mitigate such risks (IOSCO (2011); U.S. Securities and Exchange Commission (2013)).

2.1.2 Forecasting and Data-Driven Strategies

A defining feature of modern algorithmic trading is its reliance on forecasting models. Traditional approaches relied on fundamental and technical analysis, using indicators such as moving averages, momentum oscillators, and price-volume trends to form views on future movements (Murphy (1999); Pring (2002)). Computational finance then expanded this toolkit to statistical time-series models such as Autoregressive Integrated Moving Average (ARIMA) for conditional mean dynamics, GARCH for volatility, and VAR for inter-market dependencies (Teymurzade and Åšlepaczuk (2023); Vo and Ślepaczuk (2022)).

Machine Learning (ML) further enhanced forecasting capacity by uncovering non-linear relationships in data. Techniques such as Support Vector Machine (SVM), random forests, and deep neural networks demonstrated strong predictive performance in identifying patterns within complex financial time series (Fischer and Krauss (2018); Tsay

(2010)). However, these approaches also introduced challenges, particularly over-fitting, lack of interpretability, and sensitivity to regime shifts, which often limited their reliability in live trading environments (Islam et al. (2021); Park et al. (2022); Théate and Ernst (2021); Yeo et al. (2023)).

2.1.3 From Predictive Modelling to Adaptive Decision-Making

Although ML improved market forecasting, it remained essentially static, with models trained on historical data and then applied to future observations without active adaptation to changing market conditions. Financial markets, by contrast, are dynamic and sequential, and each trading decision influences the future states of the portfolio and, to some extent, the market itself. Traditional predictive models do not directly optimise such intertemporal dependencies.

This limitation paved the way for Reinforcement Learning (RL), which formulates trading as a sequential decision-making problem in which an agent interacts with the market environment, receives feedback in the form of rewards or losses, and refines its strategy to maximise long-term performance. RL can, in principle, unify prediction, execution, and risk management within a single adaptive framework, enabling strategies that learn policies under uncertainty (Deng et al. (2017); Moody and Saffell (2001)).

2.2 Reinforcement Learning in Algorithmic Trading

The application of RL in financial trading has gained significant momentum in recent years (Bai et al. (2024); Liu et al. (2022a); Théate and Ernst (2021); Zou et al. (2023a)), offering a data-driven approach to dynamic decision-making under uncertainty. Early works, such as Neuneier (1996) and Moody et al. (1998), demonstrated that reinforcement-based optimisation could outperform traditional buy-and-hold benchmarks through adaptive asset allocation. However, these early methods relied on linear function approximations and were limited in handling the high dimensionality of financial data. The emergence of deep learning enabled RL agents to process complex, non-linear patterns, paving the way for modern deep RL systems capable of learning from sequential interactions within financial environments.

RL formalises the trading process as a sequential decision-making problem in which an agent interacts with a market environment, observes its state, performs an action, and receives a reward based on trading performance. The agent's objective is to learn a policy that maximises expected cumulative reward over time. This formulation aligns naturally

with financial trading, where every decision directly influences future opportunities and portfolio states (Théate and Ernst (2021)).

2.2.1 Key Algorithms and Applications in Finance

Among the family of RL algorithms, several have become particularly prominent in algorithmic trading due to their stability, scalability, and adaptability:

- **Proximal Policy Optimization (PPO)** is an on-policy method designed for stability and efficiency through clipped policy gradients. It has been used to model trading policies in volatile markets, balancing exploration and exploitation effectively (Pippas et al. (2025); Yuan et al. (2020); Zou et al. (2023a)).
- **Deep Deterministic Policy Gradient (DDPG)** is an off-policy algorithm that supports continuous action spaces, suitable for dynamically adjusting portfolio weights or fine-tuning order sizes (Bai et al. (2024); Pippas et al. (2025); Yuan et al. (2020); Zou et al. (2023a)).
- **Twin-Delayed Deep Deterministic Policy Gradient (TD3)** is an enhanced version of DDPG that reduces overestimation bias via twin critics, improving performance stability in noisy financial environments (Pippas et al. (2025); Zou et al. (2023a)).
- **Advantage Actor-Critic (A2C)** is a synchronous actor-critic approach that balances learning speed and stability, often used as a baseline for comparative studies in trading performance (Bai et al. (2024); Li et al. (2025); Pippas et al. (2025); Yuan et al. (2020); Zou et al. (2023a)).

These algorithms represent different trade-offs between exploration and exploitation, on-policy and off-policy learning, and discrete versus continuous action spaces. Comparative analyses, such as those by Yang et al. (2020) and Mohammadshafie et al. (2024), show that no single algorithm universally dominates. This is consistent with broader benchmarks and surveys, which report competitive results for A2C, PPO, DDPG and TD3 across a range of stock and high-frequency trading tasks and emphasise that performance depends on the interaction between the learning algorithm and the market environment. In Liu et al. (2022a), for instance, these agents are benchmarked on Dow-30 equities, with A2C achieving the highest return among individual agents and an ensemble over PPO, A2C and DDPG that selects the best-performing model in each test window attaining the strongest overall performance. Meanwhile, the survey by Zou et al. (2023b) summarises applications where DDPG-based traders achieve exceptional profitability on

equity portfolios and a model-based PPO agent yields stable profits in non-stationary limit-order-book markets, alongside adaptive methods such as iRDPG and DeepTrader that explicitly embed market conditions in the state representation.

There have been several attempts at standardising RL research workflows in finance, such as the FinRL library, introduced by Liu et al. (2022b). Built atop Stable-Baselines3[1], FinRL provides a unified framework that streamlines the implementation of RL agents for trading. As shown in Figure 2.1, the framework integrates three primary layers, namely the market environments, DRL agents, and the applications. The market environment layer supports both historical and live data sources, including Yahoo Finance[2], WRDS[3] and Alpaca[4], enabling agents to interact with realistic financial scenarios.

The agent layer incorporates various RL algorithms such as DQN, DDPG, TD3, A2C, SAC, and PPO, built on top of popular libraries including ElegantRL[5], RLlib[6], and Stable-Baselines3. Finally, the application layer connects these components to real-world use cases such as stock and cryptocurrency trading, portfolio optimisation, and market regulation studies, facilitating end-to-end research and deployment in financial RL.

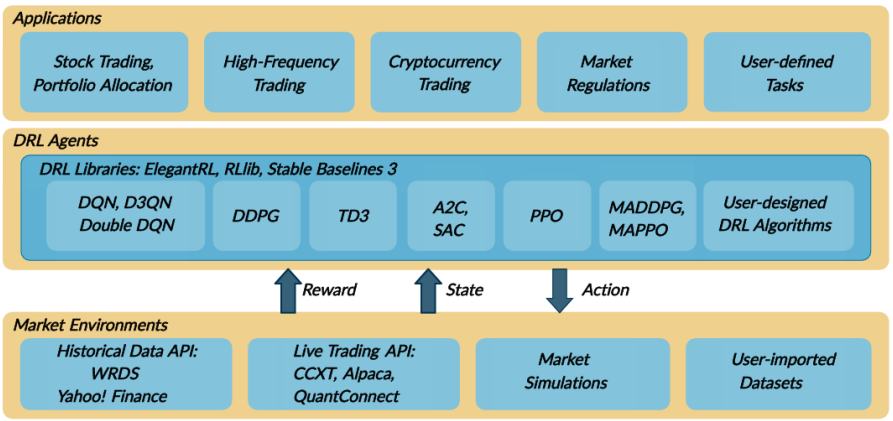

Figure 2.1: Overview of FinRL (Taken from Liu et al. (2022b))

2.2.2 Challenges and the Need for Explainability

Despite strong simulation performance, RL-based trading systems remain opaque (Bai et al. (2024); Ferreira et al. (2021); Pippas et al. (2025)). High-dimensional policies and

[1]GitHub Repository for Stable Baselines3: `https://github.com/DLR-RM/stable-baselines3`
[2]Yahoo Finance API: `https://developer.yahoo.com/api/`
[3]WRDS Python Package: `https://pypi.org/project/wrds/`
[4]Alpaca Markets API: `https://alpaca.markets/`
[5]GitHub Repository for ElegantRL: `https://github.com/AI4Finance-Foundation/ElegantRL`
[6]RLlib Documentation: `https://docs.ray.io/en/latest/rllib/index.html`

non-linear decision boundaries obscure the rationale behind trades and portfolio shifts. This 'black-box' characteristic undermines trust and limits adoption in regulated financial environments (Arrieta et al. (2019); Weber et al. (2024); Yeo et al. (2023)).

The lack of interpretability also presents practical risks. Without explainability, it is difficult for human analysts to verify whether a model's decisions align with risk management objectives or regulatory constraints. Moreover, the sensitivity of deep RL agents to data shifts and market anomalies can lead to over-fitting, erratic behaviour, or hidden biases (Izzo (2022); Kumar et al. (2022)). These issues highlight the urgent need for transparency and accountability within financial RL systems.

Consequently, since at least 2021, research has shifted towards Explainable Reinforcement Learning (XRL), integrating post-hoc interpretability techniques such as saliency maps, reward attribution, and feature importance visualisation to make agent decisions more transparent (Heuillet et al. (2021); Zytek et al. (2024)).

As noted in Heuillet et al. (2021) and Zytek et al. (2024), existing XAI approaches often struggle to convert technical outputs into explanations that non-expert stakeholders can readily understand and act upon. The core issue is not that post-hoc XAI provides no insight, but that the insight is typically presented in forms that assume technical literacy and therefore do not align well with the needs of traders, auditors, and regulators (Giorgi et al. (2025)). Researchers stress that explanations need to be tailored to the audience's expertise, with developers often creating explanations from a researcher's perspective rather than a layperson's, limiting their usefulness to business users (Bello et al. (2025)).

In particular, many methods produce representation-heavy artefacts such as attribution plots, saliency maps, or surrogate structures that can be difficult to interpret without prior knowledge and may not answer practical questions such as what most influenced the action, how strong that evidence was, or which considerations dominated the decision (Arrieta et al. (2019); Weber et al. (2024); Yeo et al. (2023)). Even when influential indicators are identified, explanations can remain context-poor because they do not clearly express directionality, regime dependence, or feature interactions in a way that supports intuitive reasoning about why a given trade was taken (Weber et al. (2024); Yeo et al. (2023)).

Moreover, different explainers can highlight different drivers and may vary across time and market conditions, which increases cognitive load and can lead non-experts to over interpret a single explanation as a definitive rationale (Arrieta et al. (2019); Weber et al. (2024)). This usability gap is widely recognised, motivating approaches that translate or augment technical XAI artefacts with audience-aware natural language narratives that preserve factual grounding while improving accessibility for non-specialists (Bello et al. (2025); Giorgi et al. (2025); Zytek et al. (2024)).

A detailed summary of key studies following the applications of RL in the financial industry can be found in Appendix B, Table B.1.

2.3 Explainable AI

2.3.1 Introduction and History of XAI

The origins of explainability in AI trace back to early interpretable models such as decision trees and linear regressions, which inherently allowed users to follow decision paths and parameter effects (Samek et al. (2017)). However, as AI evolved towards complex architectures such as deep neural networks and SVMs, interpretability diminished. This opacity led to a growing need for methods that could uncover how black-box systems derive their decisions (Agarwal et al. (2021); Ali et al. (2023); Hassija et al. (2023); Islam et al. (2021); Théate and Ernst (2021); Yeo et al. (2023); Çetin et al. (2023)).

The modern emergence of XAI was driven by the use of AI in high-stakes sectors like healthcare, law, and finance, where accountability and understanding are paramount (Mersha et al. (2024); Saranya and Subhashini (2023)). Regulatory frameworks such as the European Union's Artificial Intelligence Act (AIA) have reinforced this movement by strengthening requirements for transparency, documentation, and explainability in high-risk AI systems (Regulation (EU) 2024/1689). According to Arrieta et al. (2019), XAI now serves multiple roles, fostering user trust, validating model causality, ensuring informativeness, and aligning predictions with ethical standards.

2.3.2 Transparency, Interpretability, and Explainability

Although the need for model intelligibility is well established, the notions of transparency, interpretability, and explainability are often conflated, and no single formal definition has achieved broad consensus (Lipton (2017)). For this study, the terms are defined as follows;

- **Transparency** denotes how directly the internal mechanisms of a model can be inspected and understood. Transparent models, such as decision trees or linear models, reveal how features contribute to outputs, enabling verification and auditability. In algorithmic trading, this entails tracing how indicators or macroeconomic variables influence trading actions or portfolio shifts.
- **Interpretability** concerns the degree to which a human can logically follow a model's behaviour without full access to its internals (Maree and Omlin (2022)). It remains

the most elusive of the three concepts, often described as a property that is intuitively recognised rather than formally defined (Doshi-Velez and Kim (2017)). Ultimately, interpretable systems enable experts to relate model outputs to familiar financial notions such as momentum, volatility, or mean reversion.

- **Explainability** typically refers to post-hoc techniques that render complex, non-transparent models intelligible (Arrieta et al. (2019); Maree and Omlin (2022)). Methods such as SHAP, Integrated Gradients, or saliency maps attribute importance to inputs, identifying which features most strongly drive trading decisions.

As shown in Figure 2.2 three dimensions form a continuum. Transparency enables interpretability in inherently understandable models, and explainability extends similar understanding to otherwise opaque models, balancing insight with predictive power.

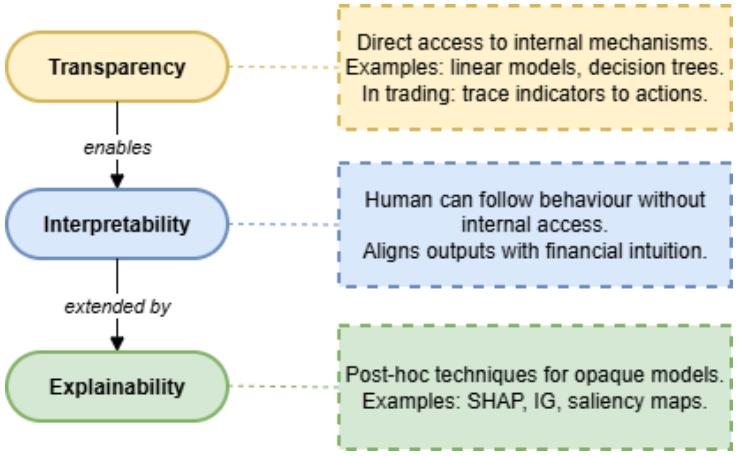

Figure 2.2: Author's conceptualisation of the relationship between the terms 'transparency', 'interpretability', and 'explainability' for the purposes of this study.

2.3.3 Explainable AI in Financial and Algorithmic Trading Contexts

Integration of AI and RL into financial systems has intensified the demand for transparency, interpretability, and accountability. Financial markets are highly regulated and risk-sensitive, where model-driven actions can carry significant economic consequences. Stakeholders such as traders, investors, and regulators therefore require not only high-performing models but also intelligible systems capable of justifying their reasoning processes (Arrieta et al. (2019); Weber et al. (2024); Yeo et al. (2023)). Transparency is therefore a central prerequisite for trustworthy AI. Opaque models can jeopardise compliance and erode user confidence, posing both financial and reputational risks. Within

autonomous trading systems, XAI functions as a governance mechanism that supports model auditing, error diagnosis, and the assurance that actions are in accordance with human and regulatory expectations.

These explainability requirements are particularly acute in algorithmic trading, where AI models autonomously allocate capital and respond to market dynamics at high speed. Recent studies increasingly apply XAI techniques to demystify such decision processes (Weber et al. (2024); Černevičienė and Kabašinskas (2024)). Common tasks such as stock price prediction and portfolio optimisation often employ complex models, including ANNs and ensemble learners like XGBoost, where post-hoc techniques, notably feature attribution and rule extraction, play a central role (Yeo et al. (2023)). Explainability in this context is typically multi-layered, spanning input-level feature importance, strategy-level interpretability, and performance attribution. This layered perspective meets the needs of financial professionals, offering insight into why a model issues a buy or sell decision, how stable those explanations remain over time, and whether the identified factors genuinely drive profitability.

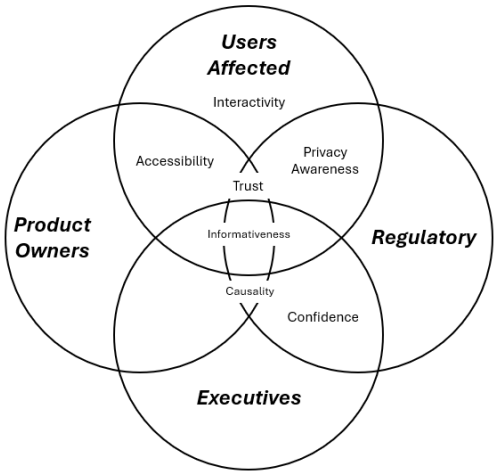

Figure 2.3: Venn Diagram of XAI Goals for Non-Technical Audiences (Adapted from Arrieta et al. (2019))

2.3.3.1 Enhancing Transparency, Accountability, and Fairness

Studies such as Yeo et al. (2023) and Ali et al. (2023) show how explainability improves the traceability of decisions in algorithmic trading, portfolio optimisation, and fraud detection. By employing techniques such as visualisation, feature attribution, and counterfactual reasoning, XAI enables analysts to link model behaviour to underlying financial logic,

thereby producing interpretable and auditable decision processes. Beyond improving operational clarity, XAI strengthens governance and regulatory compliance (Goodman and Flaxman (2017)). By adding an interpretability layer, it helps align automated decision-making with ethical and legal standards, supporting accountability and transparency in responsible AI deployment (Arrieta et al. (2019)).

In regulated financial domains, XAI is also increasingly discussed as a tool for diagnosing and mitigating bias, particularly in applications such as credit scoring, customer screening, and investment advisory, where algorithmic outcomes directly affect individuals and market participants (Demajo et al. (2020); Weber et al. (2024)). By exposing model reasoning, practitioners can identify problematic feature dependencies and clarify trading or allocation rationales (Théate and Ernst (2021)), which in turn can inform fairness-aware model design and oversight. Fairness is therefore framed as a broader normative requirement that XAI can help interrogate, even though the empirical work in this study focuses on transparency and interpretability rather than on dedicated fairness metrics.

2.3.3.2 Improving Risk Management and Model Robustness

XAI also enhances risk management and model robustness by uncovering the key factors driving trading actions and market responses. Attribution-based techniques such as Integrated Gradients and SHapley Additive exPlanations (SHAP) reveal feature-output relationships that help analysts interpret model behaviour under different market conditions (Arsenault et al. (2024); Zytek et al. (2024)). This transparency enables proactive risk mitigation and strategy refinement, ensuring that models remain interpretable, adaptable, and trustworthy in volatile financial environments.

2.3.4 Types of XAI Approaches

XAI methodologies can be broadly classified into intrinsic and post-hoc approaches

- **Intrinsic approaches** use interpretable models by design, such as linear models or rule-based systems, which provide transparent decision paths but can sacrifice predictive power in complex domains such as trading (Lipton (2017); Yeo et al. (2023)).
- **Post-hoc approaches** explain black-box models after training through feature attribution, visualisation, or counterfactual reasoning (Arrieta et al. (2019)). Within finance, post-hoc methods are predominant, offering flexibility across deep and ensemble architectures.

Among post-hoc techniques, several categories have become especially relevant;

- **Feature attribution methods**, such as SHAP and Local Interpretable Model-Agnostic Explanations (LIME), assign importance scores to input features (Arsenault et al. (2024); Yeo et al. (2023)).
- **Visualisation methods**, including saliency and heatmaps, highlight the inputs or regions of the input space that most influence predictions (Singh et al. (2024)).
- **Counterfactual explanations** evaluate 'what-if' scenarios to assess how small changes affect predictions (Gomez et al. (2020)).

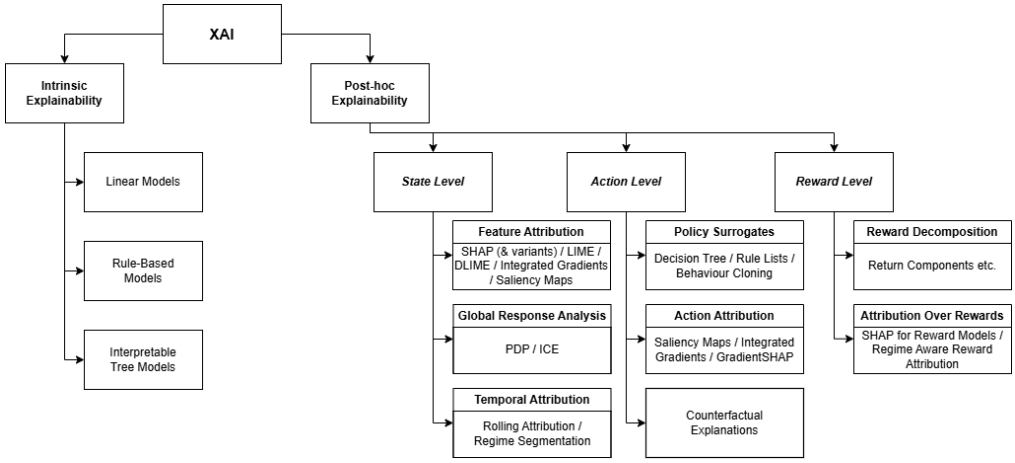

Figure 2.4: Non-exhaustive overview of XAI techniques. Intrinsic models and post-hoc state-level methods (feature attribution and response analysis) are general XAI approaches, whereas action-level and reward-level methods (policy surrogates, action attribution, and reward decomposition) are particularly relevant to RL because they explain agent policies and reward mechanisms.

The following subsections examine how feature attribution, explanation stability, policy interpretability, and reward attribution are applied within algorithmic trading to improve transparency, reliability, and financial faithfulness. While the broader XAI literature spans local surrogate methods such as LIME, rule lists, and counterfactual reasoning, the empirical work in this study concentrates on attribution-based, stability-oriented, surrogate-policy, and reward-focused methods.

2.3.4.1 Feature Attribution in Trading Models

Feature attribution methods explain model outputs by assigning importance scores to input variables, revealing which indicators or market conditions prompted a trading de-

cision. In financial contexts, these techniques translate abstract predictions into interpretable relevance rankings. The principal approaches include SHAP, LIME, IG, and saliency maps, each offering distinct theoretical and practical advantages.

Tree- and Ensemble-Based Models: For ensemble and gradient-boosted methods, SHAP, grounded in cooperative game theory, has become the most prevalent attribution tool owing to its consistency and ability to capture feature interactions (Lundberg and Lee (2017); Yeo et al. (2023)). Benhamou et al. (2021) employ SHAP to analyse gradient-boosted trees predicting market regime shifts, identifying volatility and credit spreads as leading indicators of market crashes. Their subsequent study on S&P 500 data provided a retrospective explanation of the March 2020 COVID-19 downturn, revealing contrarian signals from the technology sector that preceded the crash (Benhamou et al. (2021)). Similarly, Dikmen and Burns (2022) integrate domain expertise into SHAP analyses for peer-to-peer lending, aligning algorithmic reasoning with human financial intuition. Collectively, these studies illustrate how attribution can increase transparency and support user confidence by aligning model drivers with domain intuition, while also exposing systemic market signals (Arrieta et al. (2019); Babaei et al. (2022); Dikmen and Burns (2022)).

Deep and Reinforcement Learning Models: In neural and RL-based trading systems, gradient-based attribution is widely used. Kumar et al. (2022) apply SHAP to a Deep Q-Network agent, showing that moving averages and the relative strength index drive most trading actions. Gradient-based methods such as IG (Sundararajan et al. (2017)) and saliency maps (Simonyan et al. (2013)) provide complementary insights by computing path-integrated or raw sensitivities between inputs and outputs, clarifying how technical indicators influence policy decisions in continuous state spaces. These approaches maintain interpretability without constraining model complexity, a crucial balance in dynamic financial environments.

Local Surrogates and Model-Agnostic Tools: LIME explains individual predictions by fitting local surrogate models, often linear regressions, around target observations (Yeo et al. (2023)). Its model-agnostic design makes it suitable for diverse financial datasets but limits its scalability for large, sequential data. In comparative analyses, SHAP is often preferred for its theoretical guarantees and dual global-local interpretability (Yeo et al. (2023)).

Limitations and Hybrid Approaches: Despite their explanatory power, attribution scores alone lack contextual depth. They reveal which features matter but not how they interact

or under which conditions their influence changes. Arrieta et al. (2019) therefore advocates the combination of numerical attributions with textual or visual explanations. Such hybrid techniques bridge quantitative reasoning with human interpretability, reinforcing trust among both analysts and regulators (Černevičienė and Kabašinskas (2024)).

Overall, Yeo et al. (2023) identify SHAP as the most widely adopted XAI approach in financial machine learning, valued for its scalability across regression, classification, and RL tasks. When complemented by contextual and domain-aware interpretation, these methods collectively enhance the transparency and credibility of algorithmic trading systems.

2.3.4.2 Explanation Stability and Regime Sensitivity

Beyond identifying influential features, explainability in finance demands stability and contextual awareness. Explanation stability refers to the degree to which small changes in input or model configuration produce consistent attributions. In high-noise domains like finance, erratic explanations undermine user trust. To assess robustness, researchers perturb input data and measure shifts in feature importance rankings. Müller et al. (2022) introduce stability metrics based on fidelity and sensitivity, quantifying divergence between perturbed and baseline explanations. Methods that maintain consistent feature rankings under slight variations are deemed more reliable for high-stakes domains.

Equally important is regime sensitivity, the capacity of explanations to reflect market context. Markets alternate between distinct regimes such as bullish, bearish, and high-volatility phases, and the relevance of explanatory features often changes accordingly. Regime-conditioned analyses split data into subsets (e.g., pre-crash vs crash periods) and compute attributions separately. Benhamou et al. (2021) demonstrate this approach by training crash-prediction models across different regimes, revealing that indicators like tech-sector momentum inverted their importance once the market entered a downturn (Benhamou et al. (2021)). Similarly, Weber et al. (2024) describe dividing datasets into bullish and bearish periods before applying XAI methods to each, finding that models tend to emphasise trend-following indicators in bullish conditions but volatility and safety signals in bearish ones. This contextual differentiation enables a more faithful interpretation of trading logic.

Théate and Ernst (2021) observe that many deep RL trading agents lack explicit regime awareness and advocate using XAI to monitor how an agent's policy changes across volatility conditions. Complementary work by Singh et al. (2024) argues that explanations should be explicitly conditioned on context to avoid misleading generalisations. Together, these studies suggest that robust financial explanations must be both stable across noise

perturbations and sensitive to macroeconomic regimes to ensure trustworthy decision support.

2.3.4.3 Policy Interpretability for Trading Agents

While attribution methods clarify local feature effects, a deeper goal of explainability in finance is to interpret the global policy or strategic behaviour of trading agents. Policy interpretability seeks to uncover the rules and logic governing an agent's actions across market conditions.

A common technique is surrogate modelling, where an interpretable model such as linear regression is trained to approximate the decision function of a complex agent. The surrogate translates black-box logic into human-readable rules, for example, 'IF volatility is high AND trend is downward THEN sell'. Attanasio et al. (2020) demonstrate this approach using associative classifiers to produce explicit buy/sell rules for stock trading, enabling expert validation of model logic and increasing accountability.

Another prominent technique is action attribution, which explains why an agent chose one action over alternatives in a given state. Kumar et al. (2022) apply SHAP to the Q-values of a Deep Q-Network to determine which features drive each decision, revealing that high momentum and low volatility typically prompted 'buy' actions. Guan and Liu (2021) use IG along trading trajectories to attribute rewards to features, offering fine-grained insight into the reasoning behind each trade. These granular explanations assist portfolio managers in verifying that an agent's actions remain economically sound.

Policy-level methods often combine multiple strategies, including:

- **Decision-tree surrogates** use simplified trees derived from neural policies to summarise strategic logic via IF–THEN conditions.
- **Linear or rule-based distillation** uses sparse linear models (e.g., LASSO) to approximate agent decisions through interpretable coefficients.
- **Action-specific attributions** apply per-action SHAP or IG analyses to explain why one action was chosen over another.

Cong et al. (2021) apply policy distillation and polynomial sensitivity analysis to a deep RL portfolio agent, revealing a handful of key macroeconomic drivers, for example interest rates and volatility, responsible for most decisions while maintaining strong out-of-sample performance. Decision-tree surrogates have also proven to be effective in summarising trading policies as interpretable flow diagrams (Dispoto et al. (2025); Mern et al. (2021)). Such policy-level insights enable stakeholders to understand, validate, and refine complex trading strategies in human terms.

2.3.4.4 Reward Attribution and Performance Analysis

The final layer of explainability in algorithmic trading concerns linking a model's internal reasoning to realised financial performance. Reward attribution techniques identify which features or decisions contribute most to cumulative profit or loss, paralleling traditional portfolio performance attribution but applied to model actions rather than human ones.

One approach aggregates feature attributions over many trades to identify consistent profit drivers. Cong et al. (2021) employ polynomial sensitivity analysis to estimate each feature's effect on final returns, revealing that interest rate and volatility signals dominate profitability. Other studies correlate attribution scores with realised returns. Guan and Liu (2021) use IG to compute feature-level reward attributions and compare them to observed gains, while Babaei et al. (2022) use SHAP to dissect cryptocurrency asset allocation models, identifying Bitcoin momentum and market capitalisation as primary profit contributors. Strong alignment between attributions and realised performance provides evidence of financial faithfulness, that is, explanations correspond to genuine reward-generating factors.

In addition, researchers validate explanation quality using outcome-based metrics. If a simplified strategy built on top-ranked attributed features yields competitive performance, such as a strong Sharpe ratio, it serves as practical support for explanation reliability. However, reward attribution remains challenging due to the cumulative and path-dependent nature of financial returns. Despite its complexity, this area shows great promise for connecting model explainability to tangible economic reasoning, closing the interpretability loop between algorithmic logic and realised outcomes.

A summary of studies that apply different XAI approaches within the financial industry can be found in Appendix B, Table B.2.

2.4 Integrated Explanations and Human-Level Insight

While individual XAI techniques such as feature attribution, policy distillation, and reward decomposition each illuminate specific aspects of model behaviour, explainability remains incomplete if these insights are not communicated in a manner that human stakeholders can readily understand. Effective interpretability requires combining quantitative, visual, and linguistic perspectives into a coherent narrative that aligns with the mental models of traders, analysts, and decision makers. This integrated view transforms fragmented technical outputs into actionable, human-readable insight.

2.4.1 Combining Methods for Coherent Understanding

An integrated explanation seeks to combine different layers of understanding, namely feature importance, regime sensitivity, policy logic, and reward attribution, into a unified representation. Demajo et al. (2020) demonstrate such integration in a credit risk context by merging global feature importance with interpretable decision rules, thereby linking numerical drivers with qualitative reasoning. In algorithmic trading, a similar methodology can synthesise local and global attributions into rule based scenarios, for example summarising that a spike in volatility prompted a portfolio risk reduction action consistent with the model's bearish policy.

Building on the integrated perspective discussed by Arsenault et al. (2024), one can consider an illustrative scenario. Feature attribution first identifies that a sudden trade was driven by heightened volatility. Regime analysis then situates this event within a bearish market phase. A policy surrogate confirms that the agent's learned response in such regimes is to reduce exposure. Finally, reward attribution quantifies how this de-risking decision mitigated potential losses and improved risk-adjusted performance.

2.4.2 Visualisation as a Communication Medium

Visual interfaces play a vital role in making multi-layered explanations comprehensible. Numerous studies employ heatmaps, saliency overlays, and feature timelines to situate model reasoning within familiar market contexts. Visualising XAI outputs, for example feature attributions or saliency scores, directly over price or volatility charts allows users to see where the model identified a trend, reacted to a regime shift, or reduced exposure in response to risk cues. These visual narratives enable traders to cross reference model explanations with observed market events, thereby enhancing interpretability. A concrete example is the DI2XAI dashboard, which presents investment decision support with XAI outputs through an interactive visual interface (García-Magariño and Bravo-Agapito (2024)). As Weber et al. (2024) observe, visual explanation tools such as feature timelines can bridge communication gaps between quantitative modellers and discretionary traders, translating complex model logic into recognisable patterns that resemble traditional technical analysis.

2.4.3 From Visual to Verbal: Natural Language Summaries

A complementary and increasingly influential trend involves converting technical explanations into natural language narratives. Mersha et al. (2024) highlight that tailoring explanations to specific user roles, including traders, portfolio managers, and clients, improves

both comprehension and perceived relevance. In this setting, templated narratives are derived by converting structured outputs from feature attribution, policy analysis, and reward attribution into coherent textual summaries, provided that the mapping from numerical outputs to text is clearly specified.

Recent advances explore the use of LLM to assemble and refine these narratives automatically. Singh et al. (2024) propose using LLMs to translate raw logical outputs (for example, 'indicator X was high, hence action A was taken') into fluent, contextually enriched sentences. This hybrid pipeline bridges quantitative analysis and linguistic communication, creating explanations that are both accurate and accessible. However, LLM-assisted explanations must be validated rigorously to prevent the introduction of 'explanation hallucinations.' Each generated statement must remain traceable to empirical model evidence to preserve factual integrity and regulatory compliance.

2.4.4 Towards Human-Centred Interpretability

Ultimately, integrated explainability aims to provide stakeholders with a holistic, human-centred understanding of model behaviour (Arrieta et al. (2019); Doshi-Velez and Kim (2017); Lipton (2017)). By combining multiple XAI modalities (visual, symbolic, and linguistic), it enables users to perceive not only what a model decided, but also why and how those decisions align with market context and investment rationale (Arsenault et al. (2024); Weber et al. (2024); Yeo et al. (2023)).

This synthesis marks a natural transition toward the use of natural language processing techniques for explanation, where linguistic models can further humanise XAI outputs by articulating the underlying rationale and consequences of algorithmic behaviour in accessible and domain-relevant language (Mersha et al. (2024); Singh et al. (2024); Zytek et al. (2024)).

2.5 Accessible Explanation through NLP Techniques

The progression of explainability in AI has entered a new phase with the introduction of natural language generation capabilities. While earlier layers of this research have shown how visual and quantitative methods can clarify the internal logic of RL trading agents, the accessibility of such explanations to non-technical stakeholders remains limited.

In financial contexts, where the ultimate users of model output often include investors, regulators, and analysts without advanced machine-learning expertise, purely visual or numeric explanations may not convey the underlying rationale of a trading system. The integration of NLPs and LLMs into explainability frameworks offers a potential solution by

translating quantitative explanation artefacts into coherent, human-readable narratives that align with domain-specific reasoning (Singh et al. (2024); Zytek et al. (2024)).

2.5.1 Historical Evolution of NLP in Explainability

Rule-based expert and decision-support systems have long employed templates to verbalise reasoning chains, producing statements such as 'The applicant was rejected because income was below the threshold and debt ratio exceeded 40%' (Srinivasan et al. (2019)). Although these symbolic systems promoted transparency, they relied on rigid grammars and lacked adaptability to data-driven contexts.

The advent of neural language models marked a turning point. Recurrent architectures such as Long Short-Term Memory (LSTM) improved fluency and contextual awareness, but their sequential design constrained narrative complexity. The introduction of the Transformer architecture (Vaswani et al. (2017)) revolutionised this field through attention-based modelling and parallel sequence processing, giving rise to pre-trained models such as GPT and BERT that underpin modern LLM-based explainability. Within XAI, this evolution shifted text generation from fixed templates towards more flexible forms of model rationalisation. Rajani et al. (2019) showed that Transformer models can produce self rationalisations, that is natural language justifications conditioned on internal model states, thereby helping to connect symbolic style reasoning with data driven interpretation.

In finance, NLP-based explanations initially emerged in sentiment driven prediction systems, where models provided textual rationales for forecasts (Yang et al. (2023)). These early approaches primarily focused on describing features extracted from text. More recent LLM developments extend this paradigm by contextualising quantitative artefacts, including SHAP values and policy trajectories, into coherent, domain consistent narratives that financial professionals can readily interpret.

2.5.2 Modern LLMs and Their Role in Explainability

The latest generation of LLMs, trained on large corpora of text and code, has expanded what is practically feasible for natural-language generation within explainability pipelines. In this context, their main contribution is not to provide new financial knowledge, but to re-express structured evidence (e.g., feature attributions and regime tags) into coherent narratives tailored to different audiences. However, LLM outputs are not inherently reliable. They can produce fluent statements that are not supported by the supplied context, particularly when prompts invite elaboration. For this reason, throughout this study LLMs are treated as constrained synthesis components whose outputs are grounded in

an explicit XAI substrate and evaluated using rule-based faithfulness checks and human-centred assessment.

2.5.2.1 OpenAI GPT

The GPT lineage remains one of the most extensively studied, with GPT-4 introducing multimodal capabilities that integrate text and images, while treating tabular data as structured text or code. In the context of explainability, GPT models have been used to translate outputs from XAI tools into coherent narratives. Zytek et al. (2024) show that GPT-4 can aggregate multiple local explanations into higher level rationales, positioning GPT as a bridge between algorithmic reasoning and human interpretation.

2.5.2.2 Anthropic Claude

Anthropic's Claude family emphasises transparency and ethical alignment through constitutional AI. When prompted for chain of thought reasoning, the model can provide step by step explanations that support more auditable reasoning traces. It is designed with a comparatively conservative generation style that aims to reduce the risk of explanation hallucination, a persistent challenge in LLM-based interpretability systems.

2.5.2.3 Domain-Specific LLMs in Finance

Beyond general-purpose models, several domain-specialised LLM have been developed specifically for finance, such as FinBERT, FinGPT, and BloombergGPT (Wu et al. (2023); Yang et al. (2023)). FinBERT adapts the BERT architecture to financial text, excelling at sentiment classification of news and reports (Araci (2019)). FinGPT and BloombergGPT extend this to broader financial tasks, including reasoning over structured market data and summarising macroeconomic reports. These models serve as foundational tools for integrating NLP with quantitative XAI artefacts, translating numerical explanations into semantically rich descriptions that align with the discourse of financial analysts.

For instance, Yang et al. (2023) demonstrate FinGPT's use in summarising algorithmic outputs by combining XAI metrics such as feature importance with textual rationales from financial reports. Similarly, Zhang et al. (2023) show that instruction-tuned variants of FinGPT can generate explanations tailored to specific user roles, such as retail investors versus institutional analysts.

However, this study focuses on general-purpose LLM, i.e. GPT and Claude, rather than these specialised models. Domain-specific LLM are primarily optimised for textual sentiment analysis and report summarisation, and their support for multimodal reason-

ing over numerical XAI artefacts, such as SHAP tables or reward trajectories, is still limited compared to general-purpose multimodal models. Moreover, their limited availability, proprietary training data, and inconsistent public deployment hinder reproducibility and comparability between experiments (Nauta et al. (2023)). In contrast, GPT, Claude, and similar models offer broader contextual understanding, accessible evaluation interfaces, and, in recent versions, cross-modal reasoning capabilities that are valuable for generating accessible, evidence-grounded explanations in financial RL contexts (Bona et al. (2024)).

2.5.3 NLP and LLM-Enhanced Explainability in Financial Contexts

The integration of LLMs into explainability pipelines addresses two persistent challenges in finance. These are the cognitive barrier of technical XAI outputs, and the communication gap between quantitative models and human decision-makers. By translating structured explanation artefacts into domain-relevant narratives, LLMs make complex model reasoning both interpretable and actionable without compromising accuracy.

2.5.3.1 Bridging Technical and Human Understanding

Financial decision systems involve multiple stakeholder groups, such as data scientists, portfolio managers, compliance officers, and clients, each requiring explanations at differing levels of abstraction. Traditional XAI tools such as SHAP or Integrated Gradients reveal what drives a model's decision, but they do so numerically or visually. LLM, on the other hand, can integrate these results and generate interpretive summaries that map them to domain knowledge. For example, a feature importance plot highlighting volatility and moving averages as key drivers may be summarised linguistically as 'The model interprets the current rise in volatility as a risk signal and accordingly reduces stock exposure.' Such interpretive narratives make the reasoning chain traceable to human intuition (Singh et al. (2024); Zytek et al. (2024)).

This approach is especially valuable in regulatory frameworks such as the EU AIA, which strengthen requirements for transparency and comprehensible information about high-risk automated decision systems (Regulation (EU) 2024/1689). In compliance contexts, natural language explanations facilitate auditability and accountability, ensuring that trading systems can justify their behaviour in human-understandable terms.

2.5.3.2 NLP for Diverse Stakeholder Roles

Different user groups require varying explanation granularity and linguistic framing. For instance, a data scientist may value causal attribution detail, while a regulator prioritises

transparency and traceability. Instruction-tuned LLM enable role-specific explanation generation, adapting tone and complexity accordingly. By tailoring outputs to cognitive and professional needs, LLM-based systems can substantially improve clarity, usability, and trust in automated financial decision-making (Weber et al. (2024); Zytek et al. (2024)).

In this context, studies such as Carta et al. (2022) illustrate that linguistic structure and contextualisation directly influence user comprehension and perceived reliability of explanations. By tailoring explanations to the domain expertise and decision horizon of each stakeholder, NLP-based systems provide both accessibility and precision, fulfilling complementary roles to numerical XAI methods.

2.5.3.3 Integrative Frameworks for Financial XAI

Recent literature has begun to explore frameworks that integrate multiple XAI layers into unified narratives. Arsenault et al. (2024) discuss multilevel architectures in which feature attribution, policy interpretation, and performance attribution are synthesised through a natural language generation module. In other words, these multilevel architectures are the implementation pattern used to realise integrative frameworks, since they organise distinct explanation components into a single end-to-end pipeline that can be synthesised into a unified narrative. This form of explanatory storytelling links model signals to broader market contexts, improving interpretability and user engagement without sacrificing technical fidelity (Weber et al. (2024)).

Zytek et al. (2024) extend this concept by training LLM to align XAI outputs with predefined reasoning templates grounded in financial ontology. Their results indicate that such integration can reduce cognitive load for users and improve explanation usability in investor-facing settings. The success of these frameworks suggests that natural language synthesis represents not only a communication aid, but also a methodological extension of explainability itself.

2.6 Metrics and Measurement to Evaluate XAI

While the development of XAI techniques has advanced considerably, their credibility ultimately depends on how effectively their outputs can be evaluated. Measuring the quality of explanations is essential for ensuring that AI-driven trading systems are not only interpretable but also reliable, consistent, and useful to human stakeholders. In this sense, evaluation of XAI serves two intertwined purposes. It assesses the technical fidelity of explanations to the underlying model and determines their human-centred effectiveness in promoting understanding, trust, and actionable insight. The literature on XAI evaluation

has evolved to address both dimensions through quantitative and qualitative metrics, yet consensus on standard methodologies remains elusive (Müller et al. (2022); Weber et al. (2024); Zytek et al. (2024)).

2.6.1 Quantitative versus Qualitative Evaluation

Quantitative evaluation refers to the objective assessment of explanation quality using mathematical or statistical measures. These metrics assess whether the explanations accurately represent model behaviour, remain stable under perturbation, and correlate with meaningful outcomes. They are often computed directly from model outputs and can be applied systematically across datasets, supporting reproducibility and comparison between different XAI methods (Müller et al. (2022)). Common examples include fidelity, faithfulness, completeness, stability, and robustness.

Qualitative evaluation, by contrast, focuses on the human perspective. It seeks to understand whether explanations improve user comprehension, confidence, and ability to make informed decisions. This strand of evaluation involves user studies, surveys, and interviews with domain experts such as traders, analysts, and regulators. These assessments provide insight into how explanations are perceived and whether they truly improve trust and interpretability (Babaei et al. (2022); Zytek et al. (2024)). In practice, rigorous XAI validation often combines both perspectives, with quantitative tests to ensure technical correctness and human-centred studies to evaluate interpretive utility.

2.6.2 Quantitative Metrics

2.6.2.1 Fidelity and Faithfulness

Fidelity measures how accurately an explanation reproduces the decision logic of the underlying model. In surrogate-based interpretability, it reflects how well an interpretable proxy replicates a complex agent's actions, particularly in RL trading, where surrogate accuracy directly indicates policy alignment (Théate and Ernst (2021)). Faithfulness extends this by focusing on causal validity. A feature is faithful if manipulating it leads to predictable changes in the model's output. Perturbation-based approaches test this by removing or masking top-ranked features and observing their impact on performance (Bussmann et al. (2020)). In finance, faithful attributions provide evidence that drivers, such as volatility or credit spreads, are genuinely influential for the model's decisions, rather than purely artefacts of spurious correlations.

2.6.2.2 Completeness and Additivity

Completeness assesses whether all relevant contributors to a prediction are captured in the explanation. Additive methods such as SHAP and Integrated Gradients guarantee that feature contributions sum to the model's output (or to the difference between the output and a chosen baseline) (Lundberg and Lee (2017); Sundararajan et al. (2017)), preventing selective or partial explanations. In finance, this property ensures that critical risk variables are not obscured. Quantitatively, completeness can be expressed as a ratio of total prediction variance explained by the top-k features. Babaei et al. (2022) show that complete explanations correlate strongly with financial interpretability, offering a holistic view of market drivers.

2.6.2.3 Stability and Robustness

Reliable XAI requires that similar inputs yield similar explanations. Stability captures this consistency under small input or parameter perturbations, often measured using cosine similarity or rank correlation between attribution vectors (Müller et al. (2022)). Robustness, closely related, evaluates resilience to noise or adversarial shifts. This is an essential criterion in volatile, non-stationary markets. Stable explanations across volatility regimes can be indicative of more robust market understanding, whereas unstable patterns suggest over-fitting. In practice, stability is assessed through rolling attribution windows or temporal consistency indices (Weber et al. (2024)).

2.6.2.4 Consistency and Regime Sensitivity

In financial domains, interpretability must remain consistent across structural market changes. Consistency measures the alignment of feature attributions under different regimes, such as bullish or bearish phases. Domain-specific metrics such as Accuracy on Important Features Masked by Reference Padding (AIM) and Accuracy on Unimportant Features Masked by Reference Padding (AUM) quantify how the predictions of a model change when important or unimportant features are masked via reference padding. Reference padding replaces the selected feature entries with a fixed baseline reference value so that the input or state vector remains well formed and comparable across different masking patterns (Xiong et al. (2024)). These masking-based metrics can be computed within different market regimes to assess whether an agent's explanations remain stable across structural changes, while complementary stability scores (RIS) further characterise robustness.

2.6.2.5 Reward Attribution and Performance Alignment

For RL in trading, explanation quality should align with realised performance. Reward attribution metrics test whether influential features correlate with outcomes such as profit, Sharpe ratio, or drawdown. For example, Kumar et al. correlate attribution scores with realised returns to assess whether highly ranked features coincide with profitable states (Kumar et al. (2022)). Several studies further validate explanations by constructing simplified strategies that emphasise features identified as important by XAI methods and checking whether these reduced strategies retain comparable risk-adjusted performance, for instance in terms of Sharpe ratio, operationalised here as remaining within a small tolerance (within 5–10%) of the original agent's Sharpe ratio (Cong et al. (2021)). Such outcome-linked evaluation patterns connect interpretability with financial validity, ensuring that explanations correspond to genuinely value-generating factors.

2.6.2.6 Fidelity-Performance Trade-off

Quantitative evaluation must also acknowledge the balance between model fidelity and interpretability. Highly faithful surrogates may remain too complex for end-users, while oversimplified ones can misrepresent model logic. Bussmann et al. (2020) demonstrate that interpretable tree-based gradient boosting models with SHAP explanations achieved near-identical predictive accuracy to black-box baselines, showing that transparency does not need to reduce performance. Nonetheless, maintaining the equilibrium between quantitative accuracy and qualitative comprehensibility remains essential, especially in regulated financial environments.

2.6.3 Qualitative and Human-Centred Evaluation Metrics

Quantitative validity alone does not guarantee that explanations are meaningful or actionable to human stakeholders. In financial XAI, evaluation must also consider human-centred qualities such as clarity, trust, cognitive alignment, and perceived usefulness. Therefore, recent studies integrate subjective user assessments with quantitative measures to form a holistic understanding of interpretability.

2.6.3.1 Clarity and Readability

Clarity measures how easily users can comprehend an explanation's purpose and logic. Explanations that are linguistically concise, visually structured and use familiar financial terminology are generally perceived as clearer in user studies of XAI, where participants rate the comprehensibility of textual or visual outputs using Likert-scale questionnaires

(Hoffman et al. (2018); Mohseni et al. (2021); Zytek et al. (2024)). In trading contexts, clarity reflects whether practitioners can identify the main factors driving a model's decision or whether regulators can trace a model's reasoning from inputs to outputs.

2.6.3.2 Trust and Actionability

Trust represents a user's confidence in an AI system after viewing its explanations, while actionability measures whether those explanations support informed or correct decisions. Studies frequently assess both through pre- and post-explanation trust surveys such as the Explanation Satisfaction Scale or domain-specific trust indices (Hoffman et al. (2018); Zytek et al. (2024)). In cryptocurrency portfolio settings, Babaei et al. (2022) argue that SHAP-based decompositions can enhance investors' and regulators' confidence by revealing the factors driving allocation decisions, even though no formal user study is conducted. Explanations that improve confidence but fail to inform behaviour are of limited utility. Therefore, trust and actionability are best evaluated jointly to capture practical decision impact.

2.6.3.3 Human Alignment and Cognitive Load

Human alignment assesses how closely an explanation reflects human reasoning or domain heuristics, while cognitive load captures the mental effort required to interpret it. High cognitive load reduces usability, particularly for non-technical users. Recent human-centred XAI user studies employ instruments such as the SUS and task-specific rating scales to assess perceived usability, effort, and comfort when working with explanations (Ma et al. (2024); Rong et al. (2024)). In financial environments, explanations aligned with intuitive reasoning, such as 'buy on positive momentum, sell on rising volatility', are likely to be perceived as more cognitively efficient, whereas opaque or contradictory explanations increase strain and erode trust. This pattern is consistent with broader human-centred XAI findings on human-friendliness and trust (Paraschou et al. (2025)).

2.6.3.4 Usefulness and Satisfaction

Usefulness captures the perceived decision-making value of an explanation, while satisfaction reflects how well it meets user expectations for clarity and completeness. These dimensions are typically measured through structured surveys or interviews after interaction with the model. Weber et al. (2024) note that usefulness varies by stakeholder. For auditors, it concerns whether explanations enable verification of regulatory compliance. For traders, whether they yield new strategic insights. Together, usefulness and

satisfaction metrics ensure that explanations serve their intended human and operational purposes rather than existing as purely technical artefacts.

2.6.4 Layer-Specific Evaluation in Financial XAI Frameworks

Because explanation quality is multi-dimensional, financial XAI and explainable RL studies commonly assess interpretability through multiple complementary criteria rather than a single aggregate score (Kumar et al. (2022); Weber et al. (2024); Xiong et al. (2024)). In this study, this layer-specific perspective is operationalised through the four-layer explainability framework introduced in Section 3.1.5, and evaluated empirically in Section 4.1.6.

Recent work in financial XAI highlights recurring concerns regarding feature attribution, temporal and regime wise stability, policy level transparency, and the economic grounding of explanations (Kumar et al. (2022); Weber et al. (2024)). Building on these themes, this study organises the evaluation into four conceptual layers; feature attribution, explanation stability and regime sensitivity, policy interpretability, and reward attribution. These layers provide a unifying structure for discussing metrics and findings across the framework.

2.6.4.1 Feature Attribution

At the feature level, the primary concern is the relevance and faithfulness of the computed feature importance scores. Quantitative evaluation typically relies on measures of fidelity and faithfulness, for instance perturbation-based tests that verify whether masking or altering highly ranked features produces the expected degradation in performance, together with completeness-style checks that ensure the attributions account for the model output. Qualitatively, explanations are judged in terms of their clarity and economic interpretability, that is, whether domain experts can plausibly relate the highlighted features to known financial mechanisms and whether non-experts can follow the rationale without excessive cognitive load (Bussmann et al. (2020), Babaei et al. (2022)).

2.6.4.2 Explanation Stability and Regime Sensitivity

The second layer concerns the stability of explanations across time and market regimes. Here the focus is on whether similar conditions give rise to similar explanations and how sensitive attributions are to small perturbations in the input. Quantitative assessment uses stability indices, such as the similarity of attribution vectors under perturbation or across contiguous time windows, robustness scores, and masking-based measures such as AIM and AUM introduced in recent RL benchmarks. On the qualitative side, evaluation

centres on human perception of consistency, both across different time periods and under varying market conditions, and on whether shifts in explanations align with plausible regime changes (Müller et al. (2022); Weber et al. (2024); Xiong et al. (2024)).

2.6.4.3 Policy Interpretability

A third layer targets the interpretability of the learned policy itself, often via simplified surrogate models. The evaluative focus is on the faithfulness and transparency of these surrogates when they approximate complex RL policies. Quantitatively, this involves surrogate fidelity measures such as accuracy or agreement between the surrogate and the original policy decisions, together with indicators of rule compactness, for example tree depth or number of leaves in decision-tree surrogates. Qualitative assessment emphasises rule comprehensibility and expert validation, asking whether the inferred trading rules are understandable, plausible, and consistent with domain knowledge (Attanasio et al. (2020); Cong et al. (2021); Kumar et al. (2022)).

2.6.4.4 Reward Attribution

Finally, reward attribution connects explanatory drivers to economic and performance outcomes. The evaluative focus is on whether explanations meaningfully account for realised returns and risk characteristics. Quantitative metrics include correlations between feature attributions and realised rewards, the performance of simplified strategies that trade only on top attributed features in terms of measures such as Sharpe ratio or maximum drawdown, and explicit P&L decompositions by feature family or motif. Qualitative evaluation considers the perceived explanatory power and financial plausibility of the resulting narratives, asking whether the explanation storylines provide a coherent account of how particular signals contributed to gains or losses (Babaei et al. (2022); Cong et al. (2021); Guan and Liu (2021)).

Taken together, this multi-dimensional evaluation structure ensures that both the technical and experiential validity of XAI outputs are systematically assessed. Quantitative metrics provide reproducible rigour at each layer, while human-centred measures safeguard interpretive quality and practical usability for financial decision-makers.

2.6.5 Evaluation of NLP and LLM-Enhanced Explanations

The evaluation of NLP-generated explanations introduces distinct methodological challenges compared with traditional quantitative XAI metrics. Whereas fidelity or stability assess numerical alignment with model behaviour, linguistic explanations must also be

evaluated for semantic accuracy, coherence, and human interpretability. Accordingly, recent frameworks combine automated text metrics with human-centred assessment to ensure both factual and communicative validity (Singh et al. (2024); Zytek et al. (2024)).

2.6.5.1 Automated Metrics

Automated evaluation draws on methods originally developed for summarisation and machine translation. Standard metrics include BLEU (Papineni et al. (2002)), ROUGE (Lin (2004)), and Metric for Evaluation of Translation with Explicit Ordering (METEOR) (Lavie and Agarwal (2007)), which measure lexical or n-gram overlap between generated and reference texts. Embedding-based approaches such as BERTScore (Zhang et al. (2020a)) and SentenceMover's Similarity (Reimers and Gurevych (2019)) capture contextual semantics more effectively. In XAI, these measures quantify how closely generated explanations match expert-written rationales or ground-truth commentary. For example, Liang et al. (2025) employ BERTScore to quantify the semantic similarity between FinGPT's generated trading explanations and human-written reference texts, illustrating how embedding-based metrics can be applied in financial explanation settings.

2.6.5.2 Hybrid and Fidelity-Oriented Evaluation

Beyond linguistic quality, evaluation must ensure that textual explanations remain faithful to model reasoning. Techniques such as counterfactual consistency checks (Dhurandhar et al. (2018)) and explanation grounding (Lu et al. (2023)) can be used to check whether each statement corresponds to underlying model evidence. As noted by Singh et al. (2024), LLM-based systems introduce risks of explanation hallucination or overgeneralisation, making fidelity verification essential to preserve reliability and regulatory integrity.

2.6.5.3 Human-Centred Evaluation

Automated metrics alone cannot determine whether explanations are understandable or appropriately trusted, so human centred evaluation remains indispensable. Doshi-Velez and Kim (2017) propose a taxonomy that distinguishes application grounded, human grounded, and functionally grounded studies. In particular, human grounded experiments use real users and simplified tasks to assess general explanation quality without embedding participants in the full target application. Hoffman et al. (2018) further decompose XAI evaluation into families of measures targeting explanation goodness, including clarity and precision, explanation satisfaction and users' mental models, trust and appropriate

reliance, and overall human–AI work system performance. Within this perspective, constructs such as clarity, transparency, trust, usefulness, and actionability can be viewed as specific facets of explanation quality and user reliance rather than ad hoc survey choices.

Typical instruments include Likert-scale questionnaires, interviews, and task-based protocols that probe these constructs at the level of individual explanations. General usability and workload scales such as the System Usability Scale (SUS) (Brooke (1996)) remain widely used to assess overall system usability and cognitive load.

Recent financial studies indicate that narrative explanations can increase perceived interpretability and decision confidence compared with purely visual outputs. For instance, Zytek et al. (2024) report higher explanation satisfaction and trust when textual rationales accompany traditional XAI visualisations, while surveys such as Yeo et al. (2023) highlight growing interest in narrative, user-facing explanations in financial applications.

Building on this, Naveed et al. (2022) use focus groups and survey experiments to elicit investors' preferences over different textual explanation styles in robo-advisory settings, highlighting that perceived relevance and quality of explanations are central to trust and acceptance. Complementarily, García-Magariño and Bravo-Agapito (2024) combine interactive visual analytics with explanatory summaries in a stock-trading simulator. In a 22-participant user study they report high Likert ratings for explainability and the probability-distribution subtool, together with an average improvement of around 20% in users' relative profitability over repeated simulations.

2.6.6 Challenges and Future Directions

Despite significant progress, several challenges remain in the standardisation of XAI evaluation. First, the absence of universal benchmarks limits the comparability between studies. Metrics such as fidelity and faithfulness are often defined and computed differently between models and domains, preventing reproducibility and cross-study comparison (Weber et al. (2024); Xiong et al. (2024)).

Second, there is a persistent trade-off between technical rigour and interpretive simplicity. Explanations that are highly faithful to complex models may overwhelm users, whereas aggressively simplified summaries risk misrepresenting the underlying decision process (Hoffman et al. (2018); Rong et al. (2024)).

Third, explanations are prone to over-fitting to data artefacts or specific time periods, a challenge that is amplified in non-stationary financial markets. Stability metrics and regime-conditioned analyses can mitigate this, but cannot fully eliminate sensitivity to changes in market structure (Weber et al. (2024); Xiong et al. (2024)).

Fourth, human evaluations remain inherently subjective, shaped by the expertise of the participants, the structure of the task, and contextual expectations. User studies in XAI repeatedly show that explanations which perform well on proxy tasks do not always translate into better on-task decision quality or appropriately calibrated trust (Doshi-Velez and Kim (2017); Rong et al. (2024)).

Finally, studies note an 'explainability–performance paradox', where imposing transparency constraints or relying on interpretable models can reduce predictive precision or responsiveness to subtle market signals (Ali et al. (2023); Weber et al. (2024); Yeo et al. (2023)). Balancing interpretability and accuracy remains an open research frontier, motivating frameworks that jointly optimise technical fidelity, human-centred quality, and financial validity.

2.7 Conclusion

This chapter established the conceptual foundations. It traced the evolution of algorithmic trading and the growing use of RL in financial decision-making, highlighting how improved adaptability and performance are offset by opacity and the resulting challenges to interpretability, compliance, and trust. Then it reviewed various XAI methods in a financial context, highlighting that technical transparency and explanation quality are now prerequisites for accountable AI-driven finance. Finally, it examined the role of NLP and LLM in extending XAI to communicative, human-centred interpretability by translating technical artefacts into stakeholder-specific narratives.

Building on these foundations, this study adopts the *FinRL* framework (Liu et al. (2022b)) as the experimental backbone and reference implementation for deep RL in quantitative finance[7]. This supports rigour and reproducibility while enabling a model-agnostic framework that links feature attribution and temporal dynamics, stability and regime sensitivity, policy and action attribution, and reward attribution to portfolio performance. It also motivates the use of NLP-based narratives to enhance communicative interpretability.

These elements prepare the ground for the next chapters, which detail the methodology, experimental design, results, and evaluation of both the technical explainability framework (Experiment 1) and the narrative explanation layer (Experiment 2).

[7]Base FinRL implementation available at: `https://github.com/AI4Finance-Foundation/FinRL-Meta/blob/master/examples/FinRL_Ensemble_StockTrading_ICAIF_2020.ipynb`

3 Materials & Methods

This chapter presents the experimental framework used to enhance the transparency and interpretability of Reinforcement Learning (RL)-based trading agents in financial markets.

In line with Objective 1, Experiment 1 develops a model-agnostic explainability framework which tackles feature attribution, stability and regime sensitivity, policy and action attribution, and reward attribution. Starting from pre-processed market data, agents are trained, evaluated on standard risk–return metrics, and analysed through these layers to yield attribution maps, stability profiles, surrogate-policy rules, and reward decompositions. This workflow is summarised in Figure 3.4, with a detailed discussion in Section 3.1 and the corresponding findings and evaluation presented later in Section 4.1.

In line with Objective 2, Experiment 2 uses the resulting numerical and visual artefacts as structured input to LLM-based NLP synthesis, where models such as GPT and Claude generate novice- and expert-oriented explanations that describe agent behaviour, salient drivers, and regime-dependent strategy changes. The architecture for this experiment is described further in Section 3.2, and its results and evaluation are reported in Section 4.2.

Together, the two stages combine XAI techniques with LLM-based narrative generation to address both the technical opacity of RL models and the usability gap in existing explainability methods, moving from internal algorithmic analysis towards explanations that are accessible to non-specialist stakeholders.

3.1 Experiment 1: XAI Framework for Financial RL

3.1.1 FinRL Framework and Trading Environment

Experiment 1 is built on the FinRL framework (Liu et al. (2022b)), which provides a library of benchmark environments, data pipelines, and deep RL agents for quantitative finance[1]. As discussed in Chapter 2, FinRL and its associated benchmark studies, such as Yang et al.

[1]Base *FinRL* implementation available at: `https://github.com/AI4Finance-Foundation/FinRL-Meta/blob/master/examples/FinRL_Ensemble_StockTrading_ICAIF_2020.ipynb`

(2020) and Liu et al. (2022a), have become a reference point for evaluating RL-based trading systems on standardised market datasets. In this study, FinRL serves as the trading environment and implementation scaffold. Financial performance is benchmarked against a DJIA buy-and-hold baseline, which comprises 30 constituent stocks, while explainability quality is assessed using fidelity and stability metrics derived from Xiong et al. (2024).

Adopting FinRL as the experimental backbone serves three purposes. First, it ensures methodological rigour and reproducibility by reusing established data-handling and environment specifications rather than constructing a bespoke simulator. Second, it allows direct comparison with prior work on DJIA daily trading, since the same asset universe and portfolio-level reward structure are retained. Third, it frees the present study to focus on the design and evaluation of a model-agnostic explainability stack, understood here as applying uniformly across the deep RL trading agents with differentiable policies and a shared state–action interface used in this work. This extends the baseline training pipeline with the logging, attribution, and evaluation components needed for the four-layer framework introduced later in this chapter.

3.1.2 Data Selection and Pre-processing

This section describes the data acquisition, feature construction, and normalisation procedures used in Experiment 1. The workflow follows the FinRL stock-trading pipeline[2] (Liu et al. (2022b)), modified from ensemble RL training to provide subsequent explainability analysis. The aim is to obtain a consistent and regime-aware market representation that can be shared by all agents and explainability layers.

3.1.2.1 Data Source and Time Horizon

Historical daily market data were retrieved using the FinRL *YahooDownloader* module, so that the data source and pre-processing remain aligned with the selected FinRL reference study and enable direct comparison of results. The module interfaces with Yahoo Finance[3] through *yfinance*[4] to obtain open, high, low, close, volume (OHLCV) and ticker fields for each asset. Following the baseline configuration in Liu et al. (2022b), the dataset consists of the thirty equities that make up the DJIA, offering a representative and liquid set of large-cap US stocks for portfolio-level trading experiments. The sample spans January 2009 to June 2021, with January 2009 to June 2020 used for model training and

[2]The base code can be found at: `https://github.com/AI4Finance-Foundation/FinRL-Meta/blob/master/examples/FinRL_Ensemble_StockTrading_ICAIF_2020.ipynb`

[3]`https://finance.yahoo.com/`

[4]Documentation regarding *yfinance* can be found at: `https://ranaroussi.github.io/yfinance/`

July 2020 until June 2021 reserved for out of sample evaluation under unseen market conditions. Figure 3.1 summarises the sample period and highlights major global events.

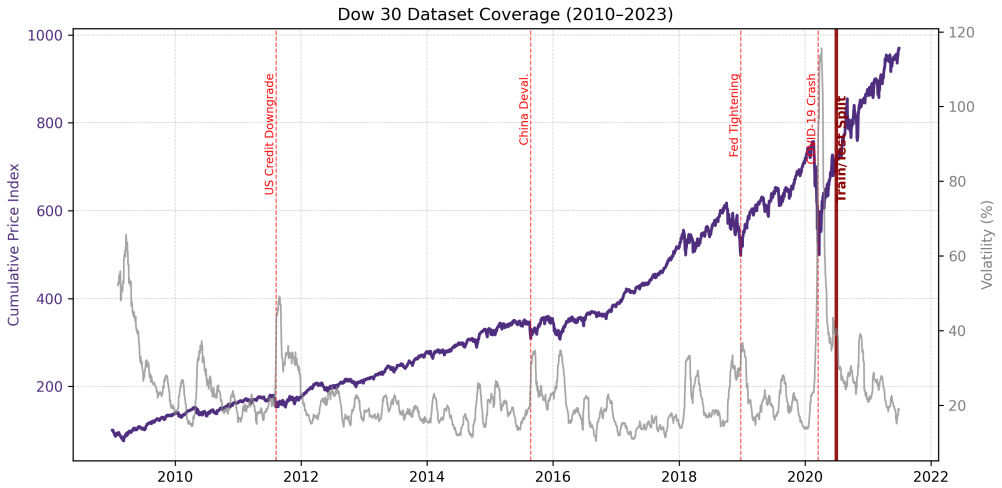

Figure 3.1: DJIA-30 daily sample period used for the trading experiments, retrieved from Yahoo Finance, with major market events and the train–test split indicated.

3.1.2.2 Feature Engineering and Normalisation

In line with Liu et al. (2022b), technical indicators were generated using the *FeatureEngineer* module, which applies standard TA-Lib[5] formulations. For each DJIA constituent, the state includes Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI), Commodity Channel Index (CCI), and Directional Movement Index (DX), 20-day Bollinger upper and lower bands, and a turbulence index, together with the raw OHLCV prices. The complete list of engineered feature set and naming convention are summarised in Table C.1, with a complete list of feature names provided in Appendix C. All continuous variables are Min–Max normalised to $[-1, 1]$ for numerical stability and cross-feature comparability.

Each trading observation is represented as a concatenated state vector comprising normalised market indicators and portfolio variables;

$$S_t = [p_{t,1}, p_{t,2}, \ldots, p_{t,n}, v_{t,1}, v_{t,2}, \ldots, v_{t,n}, b_t]$$

where $p_{t,i}$ denotes the feature set for asset i, $v_{t,i}$ the number of shares held, and b_t the remaining cash balance. This unified state representation is shared by all agents and underpins subsequent feature attribution analyses.

[5]Information about TA-Lib can be found here: `https://ta-lib.org/`

3.1.2.3 Market Regime Segmentation

To support regime-aware explainability, the processed dataset is partitioned into three market phases (bullish, bearish, and sideways) based on realised returns and rolling volatility, as shown in Figure 3.2. This regime labelling is not part of the original FinRL configuration (Liu et al. (2022b)), which treats the DJIA sample as a single continuous period, but is introduced here as an additional post hoc annotation for analysis. The underlying price series, asset universe, and reward structure remain identical to the reference setup, so financial performance and agent behaviour are directly comparable, while the regime labels are used only to stratify attribution and policy results.

Regimes are assigned using a volatility-adjusted return heuristic;

$$\text{Regime}_t = \begin{cases} \text{Bullish}, & \text{if } r_t > \mu_r + \sigma_r \\ \text{Bearish}, & \text{if } r_t < \mu_r - \sigma_r \\ \text{Sideways}, & \text{otherwise}, \end{cases}$$

where r_t is the rolling return, μ_r its mean, and σ_r the corresponding rolling volatility over a 60-day window. The resulting labels are stored as a categorical feature and later used to condition attribution and policy analyses on prevailing market regimes.

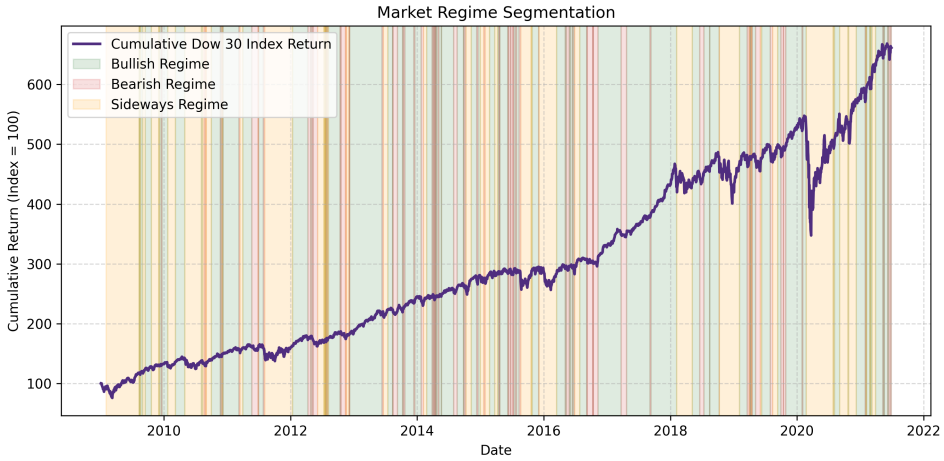

Figure 3.2: Market regime segmentation used for regime-aware attribution and policy analysis (data from Yahoo Finance).

3.1.3 Model Implementation and Training

The model implementation for Experiment 1 builds on Liu et al. (2022b), using the FinRL library with *stable-baselines3* as the RL backend[6]. All agents share the same data, feature set, and training horizon to enable fair comparison.

3.1.3.1 Reinforcement Learning Algorithms

In line with Liu et al. (2022b), the study employs four actor–critic algorithms that together span on-policy and off-policy learning paradigms, with additional algorithmic details provided in Chapter 2. In brief:

- **PPO** uses a clipped surrogate objective that stabilises policy updates in continuous control settings and serves as the primary on-policy benchmark (Pippas et al. (2025); Yuan et al. (2020); Zou et al. (2023a)).
- **A2C** provides a lighter-weight on-policy baseline that uses advantage estimation for variance reduction and sample-efficient learning (Bai et al. (2024); Pippas et al. (2025); Yuan et al. (2020); Zou et al. (2023a)).
- **DDPG** implements deterministic policy gradients for continuous action spaces, combining an actor–critic architecture with off-policy updates (Pippas et al. (2025); Zou et al. (2023a)).
- **TD3** builds on DDPG by introducing twin critic networks and lagged policy updates, which enhance stability and reduce overestimation bias (Bai et al. (2024); Li et al. (2025); Pippas et al. (2025); Yuan et al. (2020); Zou et al. (2023a)).

This enables analysis of how on-policy versus off-policy update dynamics and architectural choices affect both trading performance and the interpretability of learned policies.

3.1.3.2 FinRL Environment Configuration

The trading environment follows the *StockTradingEnv-v2* implementation in FinRL, which formulates portfolio management as a Markov Decision Process. At each trading day t, the agent observes a state vector:

$$S_t = [\text{cash}] + [\text{shares per asset}] + [\text{prices per asset}] + [\text{indicators per asset}],$$

[6]GitHub repository for Stable Baselines3: `https://github.com/DLR-RM/stable-baselines3`

where indicators include MACD, RSI, CCI, DX, and a turbulence index, as defined in Section 3.1.2. The action space is continuous for all algorithms. At time t the agent outputs a vector $a_t \in [-1, 1]^N$ with N equal to the number of traded assets, where component $a_{t,i}$ encodes the signed trade intensity for asset i relative to the per-asset cap $h_{\max}$. Negative values request selling, positive values request buying, and values near zero leave positions largely unchanged. Using the same continuous parameterisation across agents provides a directly comparable basis for the explainability analysis.

3.1.3.3 Reward Function Definition

As illustrated in Figure 3.3, the trading agent observes the current market state, executes a continuous action $a_t \in [-1, 1]^N$, and receives a reward reflecting the change in total portfolio value. For exposition, realised trade directions are sometimes referred to as *Buy*, *Hold*, or *Sell*, but these are interpretative labels derived from the continuous control rather than discrete actions.

Base Reward with Transaction Costs: The base reward follows the standard FinRL formulation, representing the agent's daily return adjusted for transaction costs:

$$R_t \;=\; \frac{V_t - V_{t-1}}{V_{t-1}} \;-\; \lambda_{\text{tc}}\, C_t,$$

where V_t is the total portfolio value at time t, C_t is a dimensionless turnover term defined as the total traded value in period t expressed as a fraction of V_{t-1}, and $\lambda_{\text{tc}} = 0.001$ corresponds to the 0.1% buy/sell transaction fee used in the environment (Appendix C.3), so that $\lambda_{\text{tc}} C_t$ summarises the impact of the FinRL transaction-cost deduction rather than introducing an additional penalty term.

Training-Time Risk Wrapper: During training only, the immediate reward is rescaled by a 30-day rolling volatility term with penalty coefficient 0.05;

$$R_t^{\text{train}} \;=\; \frac{R_t}{1 + 0.05 \cdot \sigma_{t,30}},$$

where $\sigma_{t,30}$ is the standard deviation of the last 30 realised base rewards $R_{t-29:t}$. This wrapper tempers reward spikes and promotes smoother learning while leaving the definition of R_t unchanged for evaluation and attribution.

3.1.3.4 Training Procedure and Hyperparameter Selection

Each RL algorithm was trained independently within the unified environment using the *stable-baselines3* implementations and the Adam optimiser, with a training horizon of 600k steps for all agents. Transaction costs, reward scaling, and the volatility wrapper

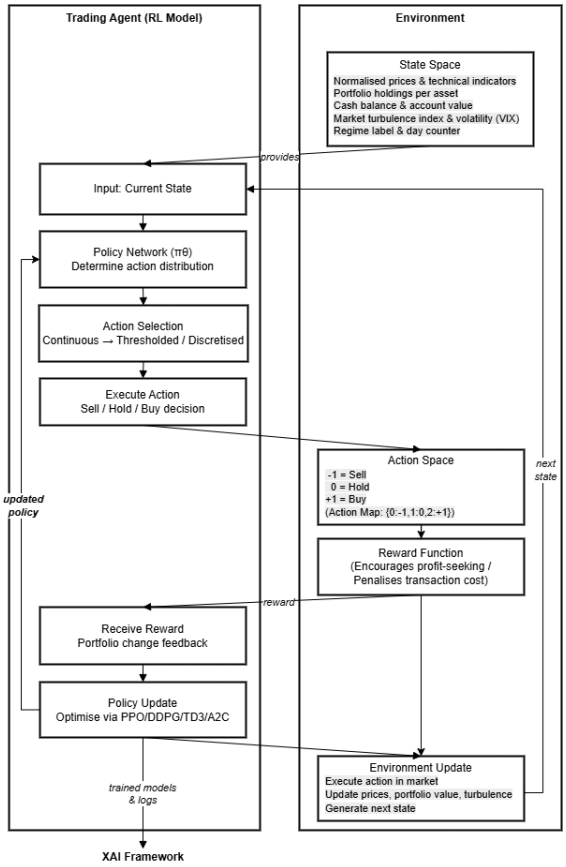

Figure 3.3: Reinforcement learning loop in the *FinRL* trading environment, showing the interaction between agent, state, action, and reward.

were held constant across runs, and evaluation was performed with deterministic policies to facilitate reproducible benchmarking.

The agent hyperparameters were selected to prioritise reproducibility and fair cross-agent comparison rather than maximising returns through extensive tuning. Concretely, we adopt the baseline FinRL training setup and the *stable-baselines3* reference configurations, keeping a shared environment, and evaluation protocol across all agents. This choice reduces degrees of freedom that could otherwise confound later explainability comparisons, since hyperparameter optimisation can materially change both learned behaviours and the resulting attribution patterns. A full list of the environment and model-specific configuration is provided in Appendix C (Tables C.3 and C.4).

Across all algorithms, only standard, stability-oriented hyperparameters were employed. For on-policy methods, Proximal Policy Optimization (PPO) follows the commonly adopted clipped surrogate objective with a clip range of 0.2, combined with a long rollout horizon

($n_{\text{steps}} = 2048$) and multiple optimisation epochs ($n_{\text{epochs}} = 10$). In contrast, Advantage Actor Critic (A2C) uses short rollouts ($n_{\text{steps}} = 5$) together with modest entropy regularisation to mitigate premature policy collapse.

For off-policy methods, both Deep Deterministic Policy Gradient (DDPG) and Twin Delayed DDPG (TD3) use a high discount factor ($\gamma = 0.99$) and a slow target network update rate ($\tau = 0.005$) to stabilise critic learning. Each employs a large replay buffer and moderate Gaussian exploration noise ($\sigma = 0.1$). TD3 further incorporates delayed policy updates (delay $= 2$) and target policy noise clipping (0.5) to reduce overestimation bias.

These configurations reflect standard practice within the selected implementations and were held fixed across agents. This design choice prioritises interpretability and controlled comparison over per-agent hyperparameter optimisation, ensuring that observed differences in behaviour and explainability artefacts arise from algorithmic structure rather than tuning effects.

3.1.4 Evaluation Strategy and Explainability Metrics

The evaluation component of Experiment 1 focuses on the quality of the explanations produced for each RL trading agent. The goal is not only to generate attribution maps and surrogate policies, but to quantify how well these artefacts satisfy standard XAI desiderata such as fidelity, stability, and consistency with realised rewards, as discussed in Section 2.6. Financial performance is summarised separately in Chapter 4 and used only as a plausibility check that the agents implement non-degenerate trading strategies.

Experiment 1 uses four main families of metrics to evaluate explainability. These are applied layer by layer to the outputs described in Section 3.1.5.

3.1.4.1 Masking-based Fidelity (AIM/AUM)

Accuracy on Important Features Masked by Reference Padding (AIM) and Accuracy on Unimportant Features Masked by Reference Padding (AUM) follow the masking-based fidelity criteria introduced in Section 2.6.2. For each explainer, features in the state vector are ranked by their attribution magnitude. A multinomial logistic surrogate is trained to predict discretised *Sell*/*Hold*/*Buy* labels from the full state. AIM measures the surrogate's accuracy when the top-k features are replaced by a neutral baseline, whereas AUM measures accuracy when the bottom-k features are masked. Lower AIM and higher AUM, relative to a random-masking baseline, indicate that highly ranked features are genuinely decision relevant, while low ranked features can be removed with limited impact. These metrics are primarily used in Layer 2.

3.1.4.2 Perturbation Stability (RIS)

The Relative Input Stability (RIS) score quantifies how sensitive an explainer's attribution vector is to small, bounded perturbations of the input. For each decision state, local perturbations are sampled within a regime-specific band and the resulting attributions are compared to the original vector using a normalised distance measure. Lower RIS values correspond to smoother, more stable explanations under small changes in the input, while higher values indicate instability. RIS is estimated on a held-out test window for each agent–explainer pair and stratified by market regime as part of Layer 2.

3.1.4.3 Surrogate Policy Fidelity and Complexity

For the decision tree surrogates used in Layer 3, fidelity is defined as the proportion of time steps where the surrogate reproduces the agent's discretised action. Complexity is measured by tree depth, number of internal nodes and leaves, and the distribution of class labels at the leaves. These quantities instantiate the fidelity and simplicity criteria and make explicit the trade off between behavioural alignment and interpretability.

3.1.4.4 Reward Alignment and Cross-Explainer Concordance

Layer 4 uses rolling ordinary least squares regressions to relate 60 day mean portfolio returns to normalised attribution vectors, aggregated into motif families (momentum, directionality, volatility bands). Evaluation considers the proportion of reward variance explained (R^2) and rank-based concordance of feature and motif importances across explainers and agents. This captures whether the narrative suggested by the attributions is stable across methods and whether it links sensibly to realised outcomes.

In Layer 1, cross-method agreement is measured via rank correlations of global and motif-level importance profiles across SHAP, IG, GradientSHAP and saliency. Together, these metrics capture fidelity, stability, and reward alignment and support comparison of explainability quality across agents and attribution methods.

3.1.5 Explainability Framework

Experiment 1 introduces a structured, four-layer explainability framework for financial RL, combining feature attribution, stability analysis, policy surrogates, and reward attribution, as summarised in Figure 3.4. Prior work in financial XAI and explainable RL indicates that single-method analyses are inadequate for complex sequential decision processes, and instead advocates multi-perspective evaluation spanning feature attribution, robustness,

policy transparency, and links to realised financial outcomes (Izzo (2022); Kumar et al. (2022); Weber et al. (2024); Xiong et al. (2024); Yeo et al. (2023)). Quantitative evaluation of the framework follows the fidelity, stability, and reward-alignment metrics in Section 3.1.4, applied separately at each layer.

First, feature attribution and temporal dynamics are widely used to identify which market indicators drive model decisions and how this influence evolves with changing conditions. Financial XAI studies such as Benhamou et al. (2021), Kumar et al. (2022) and Babaei et al. (2022) show that attribution heatmaps reveal shifts in model attention that correspond to macroeconomic events, volatility shocks, and regime transitions, providing a foundational layer for understanding input–output relationships in trading systems.

Second, explanation stability and regime sensitivity address concerns raised in works such as Xiong et al. (2024), Müller et al. (2022) and Weber et al. (2024), which argue that explanations must be both faithful to the learned policy and robust to small perturbations. In financial settings, where market regimes can shift abruptly, explanation stability is essential to avoid over-fitting artefacts and to ensure interpretability remains consistent within regimes while adapting appropriately across them.

Third, policy and action attribution is motivated by research emphasising the importance of transparent decision rules for sequential models. Surrogate policy extraction, as explored in Attanasio et al. (2020) and supported by action-level explainability approaches in Kumar et al. (2022), provides human-readable approximations of otherwise opaque RL policies. Such surrogates help traders and auditors understand how state features map to discrete trading tendencies over time.

Finally, reward attribution and performance analysis follow the growing trend of linking explanations to economic outcomes. Studies such as Cong et al. (2021) and Guan and Liu (2021) demonstrate that decomposing returns into feature- or motif-level contributions enables assessment of whether an agent's reasoning aligns with financially meaningful drivers. Outcome-grounded interpretability closes the loop between model rationale and realised profitability, an essential requirement for practical deployment.

Guided by these principles, the proposed framework integrates these complementary layers to provide a unified, multi-dimensional interpretability profile for each RL agent.

3.1.5.1 Feature Attribution and Temporal Dynamics (Layer 1)

Layer 1 quantifies the relative contribution of each market feature to the agents' trading decisions over time. Feature attributions are computed using SHAP, Integrated Gradients, GradientSHAP, and Saliency (via Captum), providing a common attribution interface across all four agents. Local surrogate tools such as LIME are not implemented, as their

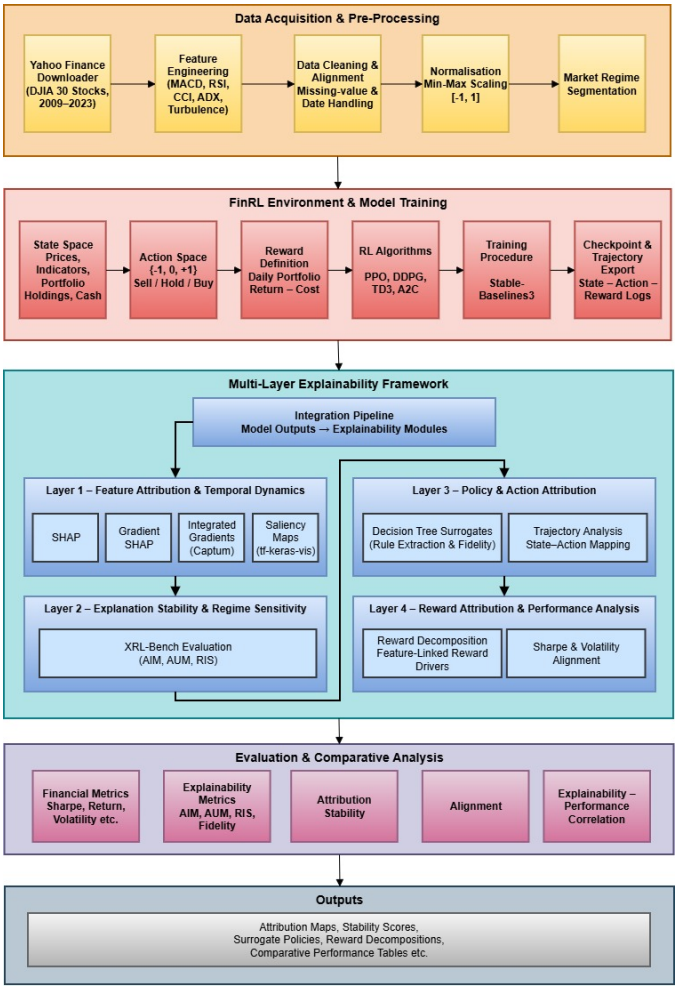

Figure 3.4: Overview of the Experiment 1 pipeline integrating RL and XAI components.

per-sample fitting cost and strong locality make them less suitable for high-dimensional sequential state spaces and repeated rolling-window analysis.

For SHAP, the *shap.KernelExplainer* interface[7] is used to estimate mean absolute Shapley values per feature. Integrated Gradients are implemented via Captum[8]'s *IntegratedGradients* class, following the original formulation of Sundararajan et al. (2017). GradientSHAP extends this approach using stochastic baselines, while Saliency maps capture local sensitivity of outputs to input perturbations.

[7] shap.KernelExplainer documentation: `https://shap.readthedocs.io/en/latest/generated/shap.KernelExplainer.html`

[8] Captum is an open-source interpretability library for PyTorch that provides implementations of attribution methods such as Integrated Gradients: `https://captum.ai/docs/extension/integrated_gradients`

Attributions are computed at checkpoints using 60-day rolling windows, then aggregated by model, checkpoint, and market regime to obtain both global and temporal attribution matrices. The resulting heatmaps illustrate how the influence of key indicators evolves over time and under varying volatility conditions, as shown in Figure 3.5.

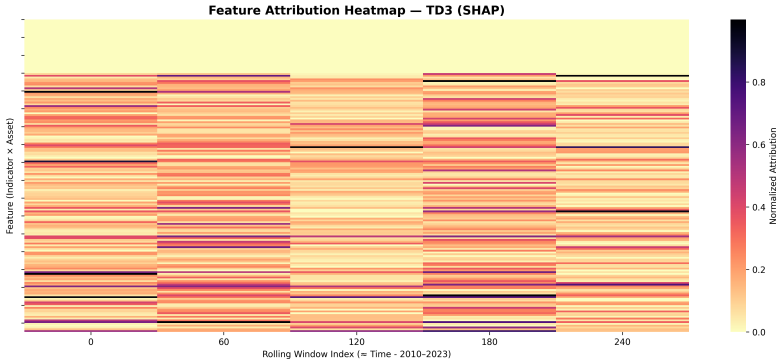

Figure 3.5: Feature attribution heatmap (SHAP $\times$ time) for the TD3 agent across 60-day rolling windows.

3.1.5.2 Explanation Stability and Regime Sensitivity (Layer 2)

Layer 2 evaluates whether the attribution signals from Layer 1 are faithful to the learned decision surface and stable under temporal and input perturbations. Following the masking protocol in Xiong et al. (2024), fidelity is measured by post-hoc accuracy under targeted feature masking, while stability is measured by the RIS score. All metrics are computed both globally and within bull, bear, and sideways regimes, as defined in Section 3.1.2, to expose regime-dependent behaviour.

Masking-based Fidelity (AIM/AUM): Masking-based fidelity uses a policy surrogate and state-wise attribution scores to construct performance curves over different values of k, representing the number of masked features. For AIM, the k most important features, according to the attribution vector, are progressively masked, and the surrogate's predictive accuracy is measured; lower accuracy values indicate that removing these highly ranked features substantially disrupts the surrogate's decisions. Conversely, AUM masks the k least important features and measures the resulting accuracy. Higher values indicate that removing these low-importance features has little effect, thereby suggesting that the attribution method has correctly identified them as uninformative.

For both AIM and AUM, the resulting accuracy curves are summarised by their area under the curve (AUC) across k, and are compared against a random masking baseline (*RAND*) averaged over multiple random feature subsets. Masking is implemented via

zero padding, following the tabular-state masking protocol used in XRL-Bench,[9] so that changes in accuracy can be attributed directly to the removal of information rather than to distributional shifts induced by unrealistic feature replacements.

Policy Surrogate and Labels: For fidelity evaluation, a multinomial logistic surrogate is trained on unmasked states to predict discretised directional labels derived from the agents' continuous controls (with discretisation detailed in Section 3.1.5.3 and results in Section 4.1.4). A chronological split prevents leakage, and train/test accuracies are reported to characterise surrogate reliability. The surrogate is used only for evaluation; the underlying RL policies remain continuous-control.

Stability (RIS): RIS quantifies the relative change in attribution vectors under small input perturbations within a bounded neighbourhood, normalised by the size of the perturbation. Lower RIS values indicate smoother, more stable explanations. RIS is estimated on a held-out test window for each agent–explainer combination and the resulting median scores are reported, complemented by qualitative analysis of how stability varies across pre-defined market regimes.

3.1.5.3 Policy and Action Attribution (Layer 3)

Layer 3 shifts from feature-level importance to behavioural interpretation by constructing transparent surrogates for each agent's decision policy. In this framework, the term *policy surrogate* refers to two complementary models. The first is multinomial logistic regression used only for AIM/AUM masking-based fidelity in Layer 2. The second is a depth-limited decision tree in Layer 3 that provides human-readable symbolic rules from the same discretised Buy/Hold/Sell labels.

For each algorithm, a decision tree is trained to map market states to directional action labels derived from the continuous actions. Let $a_t \in [-1, 1]^N$ denote the action vector at time step t and define the mean control

$$\bar{a}_t \;=\; \frac{1}{N} \sum_{i=1}^{N} a_{t,i}.$$

Directional labels are obtained by rounding $\bar{a}_t$ to the nearest element of $\{-1, 0, +1\}$, corresponding to Sell, Hold, and Buy. This discretisation is used only for surrogate learning and reporting; the underlying trading policies remain continuous.

[9] Zero padding and accuracy computation follow the tabular masking protocol in Xiong et al. (2024)

Surrogate fidelity is measured as the proportion of time steps for which the tree reproduces the agent's directional decision. The surrogates are visualised as decision trees using `matplotlib.plot_tree`[10], highlighting indicator thresholds that govern key decisions. Trajectory plots further compare surrogate and original actions over time, linking interpretable rules to observed portfolio performance.

3.1.5.4 Reward Attribution and Performance Analysis (Layer 4)

Layer 4 links interpretability to financial relevance by relating Layer 1 attributions to realised rewards. Daily portfolio rewards R_t (Section 3.1.3.3) are aggregated into overlapping 60-day windows, and window-level rewards are modelled as linear combinations of normalised attribution scores:

$$R_k^{(60)} = \sum_{i=1}^{n} \beta_i \, \tilde{a}_{i,k} + \varepsilon_k,$$

where $R_k^{(60)}$ is the mean daily reward in window k, $\tilde{a}_{i,k}$ is the min–max normalised attribution for feature i in window k, β_i is the estimated contribution weight, and ε_k is the residual term. Portfolio-level quantities such as cash, account value, and turbulence are excluded from the predictor set and used only as contextual variables. Coefficients are estimated via ordinary least squares, and diagnostics are used to confirm that the model captures the dominant linear structure.

For interpretation in Chapter 4, feature-level coefficients are grouped into motif families, namely momentum, directionality, and volatility bands, and the most influential motifs per window are visualised in the Layer 4 contribution panels.

Together, these four layers provide a unified interpretability profile for each RL agent, progressing from local feature attributions through stability and surrogate policy rules to reward alignment. The resulting structured outputs form the basis for the LLM-based narrative explanations and human-centred evaluation in Experiment 2.

3.2 Experiment 2: LLM-Based Synthesis and User Evaluation

Building upon the quantitative transparency established in Experiment 1, this phase extends the framework towards natural language-based interpretability. While the first experiment produced numerical and visual artefacts that quantify each agent's behaviour

[10] Matplotlib documentation: `https://matplotlib.org/`

through feature attributions, stability indices, policy surrogates, and reward decompositions, these outputs remain technical and difficult for non-expert users to interpret.

Experiment 2 addresses this gap by using LLMs to convert structured XAI artefacts into coherent, human-readable narratives. It tests whether these explanations preserve factual fidelity while improving clarity, accessibility, and trust. The experiment evaluates both their linguistic and factual quality across providers and prompt configurations, as well as participants' perceptions of clarity, trustworthiness, and cognitive effort.

Whereas prior work on LLM-enhanced explainability in finance often generates explanations directly from model predictions or unstructured text such as news and analyst reports (Arsenault et al. (2024); Carta et al. (2022); Singh et al. (2024); Weber et al. (2024); Zytek et al. (2024)), Experiment 2 starts from the structured XAI artefacts produced in Experiment 1, including rolling feature attributions, regime labels, surrogate policy traces, and reward attribution summaries. LLMs are tasked with verbalising these artefacts into audience-targeted narratives for novice and expert traders, and the resulting explanations are evaluated using both automated text-based metrics and a human-centred user study. Natural language synthesis is therefore treated as an integral stage of the explainability pipeline rather than a cosmetic add-on.

3.2.1 LLM-Based Synthesis Framework and Pipeline Architecture

The second experimental phase introduces a language-based synthesis framework that transforms structured XAI artefacts from Experiment 1 into natural language explanations. Acting as a translation layer between machine-level interpretability and human understanding, the framework operates as an end-to-end pipeline that includes data preparation, prompt generation, LLM inference, and validation, as shown in Figure 3.6.

An input corpus is first created by integrating the interpretability outputs from Experiment 1. These include SHAP and IG attributions, market regime labels, surrogate policy rules, and reward decomposition metrics. Each record represents a full decision context, containing state features, executed action, realised reward, and summary indicators from all four layers. To control factual scope, a dedicated *allowed_features* field lists the most influential indicators for that state and is later used as a whitelist for hallucination detection.

This whitelist is intentionally defined on a per-decision basis rather than as a global vocabulary. It constrains the LLM to discuss only indicators and drivers that are explicitly present in the local decision context, supporting two complementary controls. First, it enables a prompt-level faithfulness guardrail that prohibits external knowledge and restricts

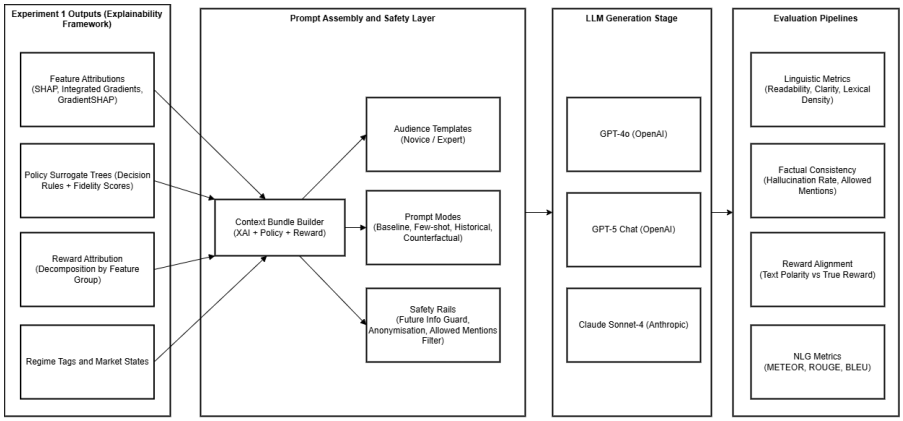

Figure 3.6: End-to-end pipeline of the LLM-based synthesis framework, showing how XAI artefacts are converted into natural-language explanations via prompt assembly, model inference, and evaluation.

mentions to the listed 'Important Features' and 'Reward Drivers'. Second, it supports a post-hoc validation step that flags any out-of-scope mentions as scope violations.

The approach is deliberately conservative. If the whitelist omits a genuinely relevant secondary indicator, the model is prevented from mentioning it even when it would be a reasonable supporting detail. Conversely, adherence to the whitelist does not guarantee that the narrative is correct about magnitudes, sign, or causality. The whitelist should therefore be interpreted as a factual scope control rather than a complete verifier of explanation correctness.

A stratified sampling strategy yields 144 decision states that are balanced across agents and market regimes. For each state, prompts are generated under five configurations, namely a base zero shot condition, few shot prompting, historical prompting, counterfactual prompting, and a synthesis mode that combines few shot exemplars with historical market context. All configurations use the same factual inputs but differ in how much exemplification and temporal information they provide.

Inference is executed via standardised APIs for OpenAI (GPT-4o, GPT-5) and Anthropic (Claude 4 Sonnet). No model-specific fine-tuning is performed; models are used in their provider-default hosted configurations, with behaviour controlled purely through prompt design. All runs use deterministic decoding (temperature set to zero) and a maximum output length of 400 tokens. Outputs are serialised as *.jsonl* files with metadata for model, provider, audience, and basic validation statistics.

Using provider-hosted models without any fine-tuning is a deliberate design choice that improves comparability across providers and avoids introducing an additional training corpus whose provenance, bias, and leakage risk would itself require governance.

However, it is also a limitation because fine-tuning could plausibly improve format adherence and domain calibration. In particular, supervised fine-tuning or preference tuning on a small set of high-quality, domain-specific explanation exemplars could reduce stylistic variance, improve consistent use of the supplied artefacts, and potentially lower hallucination by reinforcing the desired "grounded summarisation" behaviour. The constraint in this study is that there is no human gold standard corpus of per-decision explanations for this task, so any tuning set would likely be synthetic or weakly-labelled, which risks overfitting to template artefacts and inflating overlap-based metrics without guaranteeing improved faithfulness. For these reasons, control is imposed at inference time via deterministic decoding and prompt constraints, and fine-tuning is left as future work alongside stronger grounding mechanisms.

Each generated explanation undergoes automated validation to ensure factual grounding. Detected mentions of indicators and regimes are compared against the record-specific *allowed_features* list, with unsupported mentions flagged as hallucinations. Hallucination rate, allowed-mention coverage, readability scores, and token lengths are stored alongside the explanations for subsequent analysis. Overall, the synthesis framework enforces alignment between the structured numerical representations from Experiment 1 and their linguistic counterparts, combining deterministic inference with rule-based validation to produce verifiable trading narratives.

3.2.2 Prompt Assembly and Structure

For each corpus record, a context-aware prompt is assembled using audience- and task-specific templates. These templates encode stylistic and epistemic rules, including explicit statement of the agent's action and reward, reference to the prevailing market regime, grounding in observed indicators, and a prohibition on forward-looking statements. This keeps the generated narratives faithful to the experimental data and consistent with the anti-leakage principles formalised in the faithfulness guardrail in Appendix E.2.

A typical corpus record is shown in Appendix E.4, which illustrates how quantitative artefacts from Layers 1 to 4 are interpolated into a structured context block and paired with natural-language task instructions. Appendix E.3 documents the full JSON-like record schema used for prompt assembly during generation.

3.2.2.1 Template design and provenance

In this study, a *template* is a parameterised pair of system and user messages with fixed rules and variable placeholders that are filled from a single corpus record. Templates were authored as part of the proposed pipeline to operationalise two requirements motivated

in Chapter 2, being audience-appropriate communication, and faithfulness to the quantitative XAI substrate. Concretely, each instantiated prompt contains:

- **A structured context block** populated from the corpus record, including the agent identifier, regime label, executed action, realised reward, and layer summaries (feature attribution, stability scores, surrogate fidelity and rules, and reward drivers).
- **A constrained feature vocabulary**, implemented via *allowed_features*, which restricts the narrative to decision-relevant indicators and supports the hallucination checks in Section 3.2.3.
- **Task instructions** that specify the explanatory goal (for example, action rationale, regime influence, and reward consistency) and an explicit prohibition on forward-looking or external statements to prevent leakage beyond the supplied decision context.
- **Audience rules** encoded in the system message. Novice templates prioritise short sentences and minimal jargon, whereas expert templates permit denser quantitative phrasing while still avoiding model-internal terminology.

Across the five prompting modes summarised in Table 3.1, the same core template is modified only by adding or removing bounded context. The *few-shot* mode prepends exemplars with the same schema; the *historical* mode injects a short curated context field; the *counterfactual* mode restricts generation to a single "what if" based on one listed feature; and the *synthesis* mode asks for a cross-agent comparison grounded in Layers 1 to 4. The full instantiated templates and the complete corpus schema are provided for reproducibility in Appendix E.3 and Appendix E.4, while the present section records the minimum structural details required to interpret the results and validations reported in Section 4.2.

Table 3.1: Summary of prompting modes used in Experiment 2 and their design intent.

Prompting mode	Added fields or context	Primary objective
Zero-shot (base)	None. Uses only the per-decision context block (Layers A–D) and audience instructions.	Establish a baseline for clarity and factual adherence under minimal guidance.

(continued…)

Few-shot	Prepends a small number of worked exemplars instantiated from the same schema and constraints.	Improve output structure, coverage of required elements, and consistent grounding in supplied artefacts.
Historical-context	Injects a short, curated *historical_context* field describing prior market conditions relevant to the decision.	Encourage temporally coherent explanations while remaining anchored to the provided decision context.
Counterfactual	Restricts the task to one realistic increase or decrease in a single listed feature (from *allowed_features* or reward drivers).	Provide a bounded "what-if" explanation that remains tied to observed indicators.
Synthesis (cross-model)	Aggregates layer summaries across multiple agents, with an explicit comparison task.	Summarise and compare interpretability across agents and identify the most transparent option for a given audience.

3.2.2.2 Audience Roles and Prompt Variants

As detailed in Appendix E.3, templates define distinct communicative roles via their '*system*' instructions. Novice-oriented prompts instruct the model to act as an explanatory assistant for non-technical investors, prioritising simple sentence structure, explicit causal connectors, and avoidance of specialist jargon. Expert-oriented prompts position the model as an analytical writer addressing finance professionals, with more compact phrasing, explicit reference to framework metrics, and denser quantitative detail, while still avoiding unnecessary ML-specific terminology.

The same underlying templates are adapted to implement the different prompting modes. Counterfactual prompts invite the model to reason about a single realistic alteration in one of the listed indicators without introducing unseen features or future information. Historical prompts inject a short summary of preceding market conditions to encourage temporally coherent reasoning, and the synthesis configuration explicitly encourages the model to draw on evidence from all four layers of the explainability framework. Across all modes, the prompts enforce grounding in the provided artefacts and adherence to the faithfulness guardrail, as detailed in Appendix E.2.

3.2.2.3 Base Terms and Factuality Diagnostics

To support hallucination detection, a domain-specific lexicon of base terms is defined and integrated into the validation pipeline. The list, which can be found in Appendix E, Table E.1, includes key financial indicators, RL descriptors, and descriptive signal terms, supplemented with specialised terminology from Liu et al. (2022a). This lexicon prevents domain-relevant phrases from being incorrectly flagged as hallucinated while maintaining sensitivity to genuinely spurious or unverifiable statements.

3.2.3 Evaluation Metrics and Validation Framework

3.2.3.1 Reference Construction

Because no human-written references exist for each decision instance, a corpus of pseudo-references is automatically generated from the factual XAI corpus; they are termed 'pseudo' because they function as reference texts for overlap-based evaluation, but are machine-constructed summaries rather than human gold-standard explanations. Each pseudo-reference encodes the prevailing market regime, the agent's executed action, the most influential features for that decision, and the sign of the resulting reward. These references provide a consistent factual baseline against which candidate explanations can be compared.

3.2.3.2 Lexical and Semantic Similarity

The first group of metrics quantifies textual overlap and semantic correspondence between generated explanations and their pseudo-references. BLEU is used for n-gram precision, ROUGE-1 and ROUGE-L for unigram recall and longest common subsequence overlap, and METEOR for synonym- and stem-aware similarity. Together, these metrics assess how closely the generated narratives reproduce the structure and content of the reference descriptions without requiring exact lexical identity. In this study, the focus remains on n-gram based metrics (BLEU, ROUGE, METEOR) for computational efficiency and comparability with prior work.

3.2.3.3 Readability and Linguistic Clarity

Readability is assessed using Flesch Reading Ease and the Gunning Fog index as approximate indicators of how accessible each explanation is to different reader groups. A derived clarity score combines reading difficulty with simple length heuristics to penalise

overly short or excessively verbose texts, enabling comparison of how effectively each model adapts its linguistic style to novice versus expert prompts.

3.2.3.4 Factual Grounding and Hallucination Rate

In this study, the term hallucination is used in an operational sense to measure whether an explanation introduces out-of-scope decision factors relative to the structured XAI record. Each generated explanation is parsed for mentions of indicators, regimes, and domain signals, and these mentions are matched against the record-specific *allowed_features* list (and the explicitly provided reward-driver fields). Mentions that fall outside this per-record whitelist are treated as hallucinated, and the hallucination rate is computed as the proportion of detected domain mentions that are unsupported by the allowed vocabulary.

Concretely, for each generated explanation we extract a set of *feature-like mentions* (tokens that resemble indicators or feature names after normalisation and stopword filtering). Each mention is then matched against the record-specific whitelist A, constructed from the decision's `allowed_features` (and the explicitly provided reward-driver fields). We count $n_{\text{forbidden}}$ as the number of extracted mentions that do not match the whitelist, and n_{total} as the total number of extracted feature-like mentions. The hallucination rate is then computed as:

$$h \;=\; \frac{n_{\text{forbidden}}}{\max(1,\, n_{\text{total}})}\,. \tag{3.1}$$

A hallucination rate of 0.1 therefore means that 10% of the detected feature-like mentions in the explanation were out of scope relative to the decision record (for example, 2 out of 20 extracted mentions were not contained in the whitelist). If no feature-like mentions are detected, h is defined as 0 by convention.

This metric therefore captures feature invention and scope violations, rather than guaranteeing overall factual correctness. It does not directly detect errors such as misstated numerical values, incorrect sign or directionality claims, or causal assertions that reuse allowed terms but are not warranted by the supplied evidence. To reduce false positives for generic finance language, validation is paired with the domain lexicon of base terms described in Section 3.2.2.3 and Appendix E.1.

3.2.3.5 Aggregate Evaluation and Cross-Model Comparison

Metric values are aggregated across samples to obtain mean scores by provider, model, prompting configuration, and audience. The resulting tables support direct comparison of fluency, faithfulness, and clarity between models (GPT-4o, GPT-5, and Claude 4 Sonnet) and between novice- and expert-oriented prompts. This quantitative framework provides

a model-agnostic basis for assessing linguistic quality and factual reliability, complementing the technical explainability metrics from Experiment 1 and linking algorithmic transparency to communicative clarity.

3.2.4 Human-Centred Evaluation

While the automated evaluation in Section 3.2.3 provides objective linguistic and factual metrics, it does not capture how users perceive the explanations' interpretability, trustworthiness, or usefulness. To complement these results, a human-centred study was conducted comparing participants' perceptions of LLM-generated narratives with the traditional XAI visual artefacts from Experiment 1. Following the taxonomy of Doshi-Velez and Kim (2017), it is framed as a human-grounded experiment with novice participants performing simplified explanation-card tasks that preserve the core interpretability questions of the trading setting, with full questionnaire content reported in Appendix F.

3.2.4.1 Participants and Study Design

Twelve novice participants, a sample size consistent with prior human-centred financial XAI studies (García-Magariño and Bravo-Agapito (2024); Naveed et al. (2022); Zytek et al. (2024)), were recruited via convenience sampling from local academic and professional networks. Participants were predominantly students and early-career professionals with basic familiarity with financial markets. Inclusion criteria were: age ≥ 18, English proficiency, and demonstrable involvement in finance, such as having studied economics or finance, working in an economic or financial role, or prior experience trading in the stock market. Participation was anonymous and uncompensated.

Seven explanation cards were constructed from Experiment 1 artefacts, covering (i) rolling SHAP feature contributions for PPO, (ii) a TD3 policy tree excerpt, (iii) TD3 reward attribution via Integrated Gradients, (iv) DDPG attribution stability, (v) DDPG indicator–reward influence, (vi) Dow Jones market regimes (2009–2021), and (vii) a cross-model feature-importance comparison.

For each card, participants first viewed the traditional visual explanation (TE) and then the matched natural-language explanation (NLE) derived from the LLM, resulting in a paired TE to NLE presentation per card.

3.2.4.2 Measures and Procedure

After each card, participants rated five perception dimensions on 5-point Likert scales (1–5): perceived clarity, trust, transparency, usefulness, and actionability. These dimen-

sions were adapted from established interpretability and XAI evaluation frameworks (Doshi-Velez and Kim (2017); Hoffman et al. (2018); Lipton (2017)) and operationalised via item wordings listed in Appendix F. Clarity and transparency serve as proxies for explanation quality and user understanding, while trust, usefulness, and actionability capture satisfaction and appropriate reliance on the RL agent.

Participants also answered preference questions comparing the visual and narrative formats for each card (*Visual*, *Narrative*, *Both*, *or Neither*), as well as a brief set of post-study questions on overall format preferences and perceived strengths and weaknesses of each explanation type. Open text fields captured short qualitative comments, which were later used to contextualise the quantitative ratings.

For each outcome, participant-level scores were obtained by averaging relevant Likert items, and these were summarised across participants and cards to obtain descriptive statistics. Categorical preferences were summarised as counts and proportions. The resulting descriptive findings are reported in Section 4.2.5, where they are triangulated with automated readability and factuality metrics from Experiment 2.

3.3 Conclusion

This chapter has outlined the experimental design for the two core components of the study. Experiment 1 develops a model-agnostic explainability framework for RL-based trading agents, whilst Experiment 2 layers LLM-based narrative synthesis and a human-centred user study on top of these technical artefacts. Together they move from technical transparency to communicative interpretability, with implementation and software-stack details documented in Appendix A.1. Chapter 4 presents the empirical results of both experiments and evaluates the proposed framework.

4 Results & Discussion

This chapter presents the empirical outcomes of the two-stage experimental framework introduced in Chapter 3. Results for Experiment 1 are reported in Section 4.1, covering financial and trading performance of the four RL agents together with the four explainability layers. Experiment 2 is presented in Section 4.2, where LLM-generated natural-language explanations are evaluated using automated linguistic metrics and a human-centred user study. Section 4.3 then synthesises findings across both experiments and relates them back to the research questions and objectives.3

4.1 Experiment 1 Results & Evaluation

This section presents the empirical findings of the first experimental stage, which evaluates the financial performance and explainability properties of the RL-based trading agents using the four-layer framework introduced in Chapter 3. These layers are:

- **Layer 1:** Feature Attribution and Temporal Dynamics;
- **Layer 2:** Explanation Stability and Regime Sensitivity;
- **Layer 3:** Policy and Action Attribution;
- **Layer 4:** Reward Attribution and Performance Analysis.

4.1.1 Baseline Financial Performance

All four RL agents achieve competitive financial performance on the DJIA test window, with three of the four models outperforming the buy-and-hold benchmark on risk-adjusted metrics (Figure 4.1). Table 4.1 compares the reproduced results with the FinRL benchmarks reported in Liu et al. (2022b). Additional information on the financial performance of the models can be found in Appendix C.5.

Table 4.1: Baseline financial performance of RL agents as reported in Liu et al. (2022b) and reproduced in this study.

	A2C		PPO		DDPG		TD3	
Metric	**FinRL**	**This study**	**FinRL**	**This study**	**FinRL**	**This study**	**FinRL**	**This study**
Initial value	$1.00M	$1.00M	$1.00M	$1.00M	$1.00M	$1.00M	$1.00M	$1.00M
Final value	$1.46M	$1.46M	$1.42M	$1.38M	$1.40M	$1.32M	$1.39M	$1.38M
Annualised return (%)	46.65	46.65	41.90	38.66	40.34	32.86	39.38	38.09
Annualised Std (%)	17.86	17.87	16.33	16.12	17.28	16.55	15.08	14.82
Sharpe ratio	2.24	2.23	2.23	2.11	2.05	1.80	2.28	2.25
Max drawdown (%)	-7.59	-7.60	-9.41	-22.67	-8.10	-7.98	-8.92	-8.76

Figure 4.1: Cumulative returns on the test window (July 2020 to June 2021) for A2C, PPO, DDPG, TD3, and the DJIA buy-and-hold benchmark.

From an evaluation standpoint, this establishes that the explainability analysis is applied to agents that are neither degenerate nor poorly performing. The reproduced policies attain annualised Sharpe ratios between roughly 2.0 and 2.3, which is comparable to deep RL trading systems on major equity markets. For example, Yang et al. (2020) report Sharpe ratios around 1.1 for individual actor–critic agents and 1.3 for an ensemble strategy on the DJIA, Théate and Ernst (2021) obtain Sharpe ratios up to approximately 1.5 on large-cap equities, and Wu et al. (2020) show that GDQN and GDPG strategies yield consistently positive risk-adjusted returns across U.S., U.K., and Chinese stocks. Against this backdrop, the attribution and stability metrics can be interpreted as characterising behaviour that is competitive with existing RL trading systems. The analyses use the final checkpoint for each agent, with intermediate checkpoints consulted for stability checks but omitted from the tables and figures for brevity.

4.1.2 Feature Attribution and Temporal Dynamics (Layer 1)

Layer 1 of the framework investigates the underlying drivers of each agent's trading behaviour by quantifying feature influence on policy actions using four attribution techniques (SHAP, Integrated Gradients, GradientSHAP, and Saliency Maps).

Feature names follow the pattern *TICKER_INDICATOR*, where the ticker denotes a Dow 30 constituent and the indicator suffix belongs to one of four families; *boll_lb*/*boll_ub* (20-day lower and upper Bollinger band), *DX_30* (30-day Directional Movement Index), *RSI_30* (30-day Relative Strength Index) and *CCI_30* (30-day Commodity Channel Index). Appendix C contains a list of all tickers and features.

Figure 4.2 reports the global SHAP-based feature importances per agent. Across all four agents, volatility-bounded features from Bollinger bands (*boll_lb* and *boll_ub*) appear consistently among the highest-impact inputs, indicating that distance to price envelopes is a core driver of risk-on versus risk-off exposure. Directional movement indicators (*DX*) are especially prominent for A2C and also feature for PPO and TD3, while relative-strength terms (*RSI*) enter the top attributions primarily for DDPG and TD3. Company-specific Bollinger metrics (such as *BA_boll_lb*) reinforce the picture that volatility-derived boundaries form a dominant input family shaping policy formation.

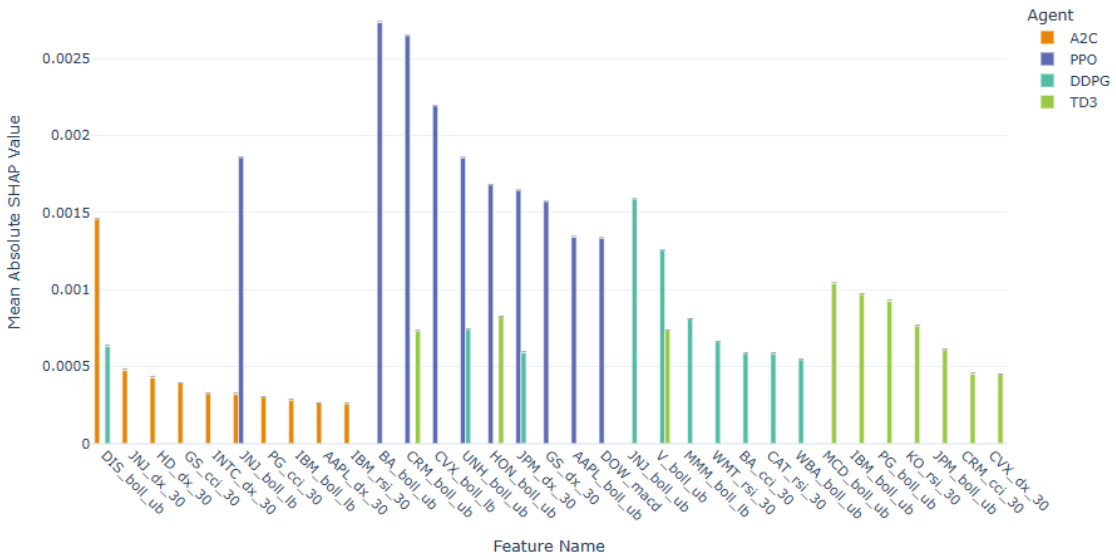

Figure 4.2: Top ten influential features per model (global SHAP values).

To examine temporal dynamics, rolling-window Integrated Gradients were applied to measure the persistence of feature influence across market phases. Figure 4.3 visualises the evolution of mean absolute attributions in 60-day windows. The heatmaps show that upper Bollinger band features for several Dow 30 constituents become increasingly

prominent in later intervals, most notably for *JPM*, *IBM*, *MCD* and *HON* in the TD3 agent, with related Bollinger terms also intensifying for DDPG. Earlier windows place relatively more weight on directional and momentum signals, while later windows shift emphasis towards volatility-sensitive indicators, reflecting regime-dependent changes in explanatory focus rather than static feature reliance.

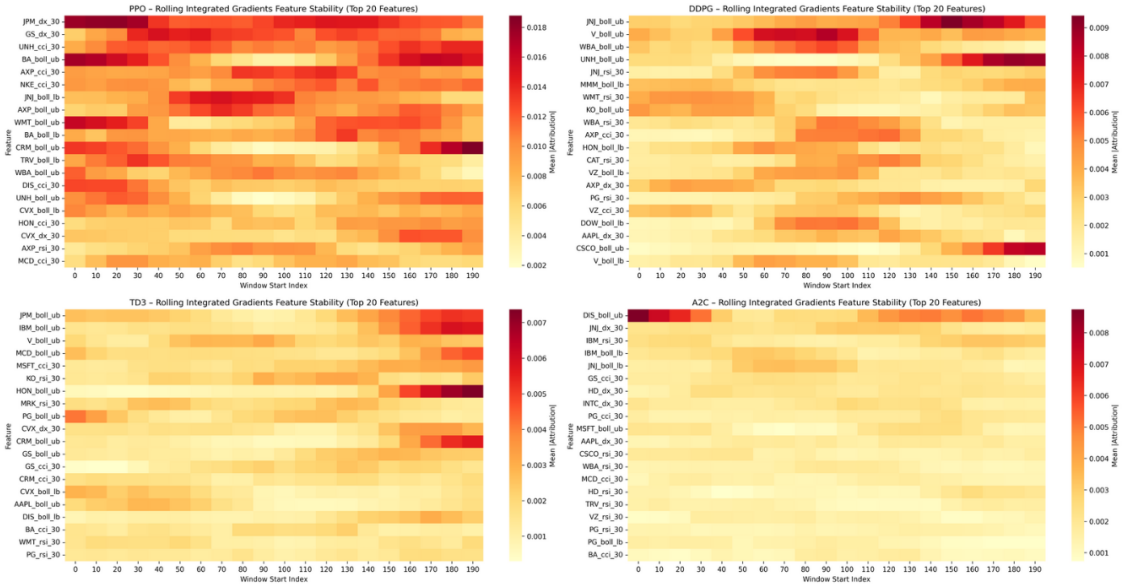

Figure 4.3: Rolling Integrated Gradients feature stability (top twenty features; window size = 60).

4.1.2.1 Representative Periods

To make the regime dependence concrete, Table 4.2 summarises three non-overlapping 60-day windows from the test set together with the dominant drivers per agent and the change relative to the preceding window.

4.1.2.2 Regime Sensitivity

Attribution magnitudes were aggregated by regime tag (*bull*, *bear*, *sideways*). Table 4.3 reports, for each agent, the top five features per regime, ranked by regime-specific mean absolute attribution (the global mean is computed over the full test window). The *bear* regime was sparse between the July 2020 to June 2021 test window, and hence those entries are indicative.

Table 4.2: Representative 60-day windows and dominant drivers (mean absolute attribution, top three per agent).

Window	PPO	DDPG	TD3	A2C
Aug-Sep 2020	JPM_DX_30, GS_DX_30, UNH_CCI_30	JNJ_boll_ub, V_boll_ub, WBA_boll_ub	JPM_boll_ub, IBM_boll_ub, V_boll_ub	DIS_boll_ub, JNJ_DX_30, IBM_RSI_30
Nov-Dec 2020	BA_boll_lb, CRM_boll_lb, IBM_boll_lb	AXP_DX_30, PG_RSI_30, VZ_CCI_30	MCD_boll_ub, MSFT_CCI_30, KO_RSI_30	HD_DX_30, INTC_CCI_30, MRK_RSI_30
Apr-May 2021	CVX_boll_lb, HON_CCI_30, AXP_RSI_30	JNJ_boll_ub, IBM_RSI_30, UNH_boll_lb	HON_boll_ub, PG_boll_ub, CRM_boll_ub	CSCO_RSI_30, WBA_RSI_30, MCD_CCI_30

Table 4.3: Top five features by regime (ranked by regime-specific mean absolute attribution).

Agent	Regime	Top features
PPO	Bull	BA_boll_lb, CRM_boll_lb, IBM_boll_lb, HON_CCI, AXP_RSI
PPO	Sideways	JPM_DX, GS_DX, UNH_CCI, CVX_boll_lb, AXP_RSI
PPO	Bear	JPM_boll_ub, IBM_boll_ub, MCD_boll_ub, HON_boll_ub, WMT_boll_ub
DDPG	Bull	AXP_DX, PG_RSI, VZ_CCI, DOW_boll_ub, AAPL_DX
DDPG	Sideways	JNJ_boll_ub, V_boll_ub, WBA_boll_ub, UNH_boll_lb, IBM_RSI
DDPG	Bear	JNJ_boll_ub, V_boll_ub, IBM_RSI, VZ_boll_lb, CAT_RSI
TD3	Bull	JPM_boll_ub, IBM_boll_ub, V_boll_ub, MCD_boll_ub, MSFT_CCI
TD3	Sideways	KO_RSI, HON_boll_ub, MRK_RSI, PG_boll_ub, GS_CCI
TD3	Bear	HON_boll_ub, PG_boll_ub, CRM_boll_ub, CVX_boll_ub, AAPL_DX
A2C	Bull	DIS_boll_ub, JNJ_DX, IBM_RSI, HD_DX, INTC_CCI
A2C	Sideways	MRK_RSI, MSFT_boll_ub, AAPL_DX, CSCO_RSI, WBA_RSI
A2C	Bear	DIS_boll_ub, JNJ_DX, IBM_RSI, HD_DX, TRV_RSI

Notes. Feature names follow the pattern *TICKER_INDICATOR*. The complete list of engineered feature set and naming conventions are summarised in Table C.1, while Dow 30 ticker information can be found in Table C.2.

4.1.2.3 Directionality

Global SHAP bars quantify salience but not sign. Direction was therefore inferred from the signed 60-day reward-contribution series per feature, with $P(>0)$ denoting the share of windows with positive mean contribution. Confidence intervals were obtained by bootstrap over windows (1,000 resamples). For PPO, *BA_boll_lb* was consistently positive and *CVX_boll_lb_ub* negative (both $P(>0) = 100\%$). For A2C, *JNJ_boll_lb* and *INTC_DX_30* were positive and *MSFT_boll_ub* negative. For TD3, *JPM_boll_ub* was positive, with *MRK_CCI_30* and *DIS_boll_lb* negative. Signs were stable with narrow intervals,

indicating consistent effects across windows.

4.1.2.4 State-Level Saliency Patterns

Saliency maps were used to examine gradient based sensitivity at both global and state level. Global Saliency scores were obtained by averaging the absolute gradient of the action logits with respect to the normalised state vector over the test window. Figure 4.4 plots the ten most salient features per agent. Across agents, the largest gradients fall on the same volatility and momentum families highlighted by global SHAP and Integrated Gradients, including Bollinger bands, *MACD* and *DX*, *RSI*, and *CCI* terms. A2C and TD3 are particularly sensitive to upper Bollinger bands for large constituents, while PPO and DDPG place more weight on *MACD* and directional movement indicators.

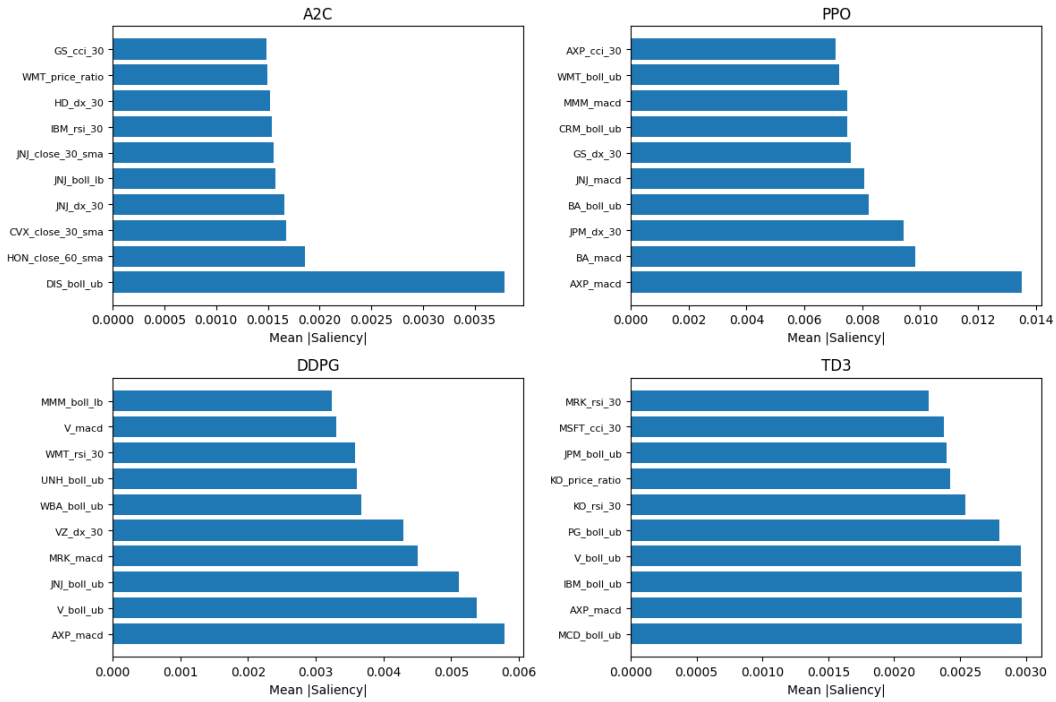

Figure 4.4: Top ten features by global Saliency for each agent (mean absolute gradient with respect to the state).

Saliency is moderately concentrated. The ten most salient features account for roughly 8.3–9.7% of the total gradient mass per agent, and the top twenty for about 14.1–16.6%. On average, the top-ten features are around 2.5 times more salient than the mean feature, indicating that each policy's instantaneous sensitivity is focused on a relatively small subset of technical indicators rather than being diffuse across the full state vector.

At the state level, Saliency maps were inspected around days with large position changes and drawdowns.[1] These local maps corroborated the global rankings. Short bursts of gradient mass occur on the same volatility and momentum indicators, particularly when prices approach Bollinger band extremes or directional movement signals spike. This locality makes Saliency informative for diagnosing specific trading episodes, such as the build up to PPO's late 2020 drawdown, whereas Integrated Gradients and GradientSHAP provide a smoother view of how influence accumulates over longer trajectories.

4.1.2.5 Cross-Method Agreement

To assess robustness across attribution methods, Kendall's τ was computed between the top-20 rankings from SHAP, Integrated Gradients, GradientSHAP and Saliency Maps for each agent using the intersection of named features. A value is reported only when at least five features overlap. Kendall's τ is used because the comparison concerns agreement in the relative ordering of overlapping top-20 feature rankings rather than linear association of attribution magnitudes; τ measures pairwise concordance and is well suited to short ranked lists with potential ties.

Table 4.4: Rank agreement (Kendall's τ) between attribution methods on top-20 features.

Agent	SHAP vs IG	SHAP vs GradSHAP	IG vs Saliency	Mean
A2C	0.62	0.36	0.82	0.60
PPO	0.41	0.31	0.73	0.48
DDPG	0.49	0.09	0.72	0.43
TD3	0.63	0.71	0.56	0.63

As summarised in Table 4.4, rank agreement between SHAP, Integrated Gradients, GradientSHAP, and Saliency is generally moderate to strong on the overlapping top features, with Kendall's τ typically lying between about 0.3 and 0.8. The main exception is the DDPG SHAP–GradientSHAP comparison ($\tau = 0.09$), which indicates that these two gradient based explainers emphasise different subsets and orderings within the overlapping top ranked features for this agent.

Cross method comparisons on the full importance profiles (Figures 4.5 and 4.6) show that SHAP, Integrated Gradients, and GradientSHAP exhibit very high agreement on global feature importance (correlations $r \approx 0.92$–0.98), while Saliency is more weakly correlated and more locally variable. This pattern is consistent with findings from XRL-Bench, where

[1] Saliency was computed as the gradient of the action logits with respect to the normalised state vector, using the same zero baseline as for Integrated Gradients.

SHAP based and gradient based explainers achieve markedly higher fidelity (AIM, AUM) and stability (RIS) than perturbation based saliency approaches such as SARFA, PS, and TabularLIME (Xiong et al. (2024)). In a different financial context, Müller et al. (2022) evaluate several SHAP variants using a triad of fidelity, stability, and robustness metrics and report that RESHAPE and LossSHAP produce more faithful, stable, and robust attribute rankings than alternative methods.

Financial XAI surveys reach similar high-level conclusions about practice. Feature-relevance methods, particularly SHAP and its variants, are by far the most widely adopted tools for identifying and ranking key drivers of model predictions in credit scoring, portfolio optimisation, anomaly detection, and related tasks (Benhamou et al. (2021); Dikmen and Burns (2022); Weber et al. (2024); Yeo et al. (2023); Černevičienė and Kabašinskas (2024)). Accordingly, this study treats SHAP-style attributions as the primary mechanism for robust global ranking of drivers, with gradient-based methods providing complementary, more temporally local detail along agent trajectories.

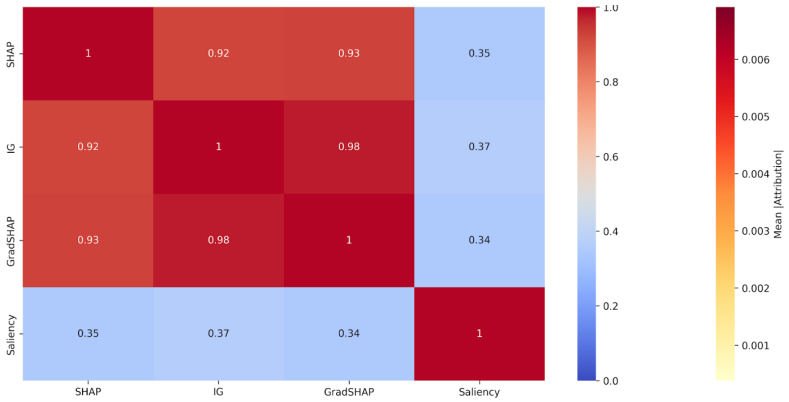

Figure 4.5: Correlation of feature-importance profiles across explainability methods.

Figure 4.6: Cross-method consensus on top features (normalised importance across SHAP, Integrated Gradients, GradientSHAP, and Saliency).

Collectively, the findings of this layer show that feature attribution exposes which technical indicators shape policy actions and how their relevance evolves with changing regimes. For PPO, the high salience of *BA_boll_lb* and related volatility features during late 2020 coincides with its deeper drawdown in that interval, while A2C's steadier reliance on *DX*/*RSI* aligns with its higher Sharpe and milder drawdown. Cross-method agreement and the external benchmarks above indicate that these patterns reflect stable, non-degenerate attributions rather than artefacts of any particular explainer.

4.1.3 Explanation Stability and Regime Sensitivity (Layer 2)

Layer 2 evaluates whether the attributions from Layer 1 are both decision-relevant and robust. Fidelity is assessed using masking-based post-hoc accuracy curves (AIM for masking the most important features, AUM for masking the least important), and stability is assessed with the perturbation-based RIS score, following Xiong et al. (2024).

Table 4.5 summarises the AUC-over-k accuracies for AIM, AUM, and the RAND reference on a held-out test window. For PPO, the expected separation AIM $<$ RAND $<$ AUM is clearly visible, indicating that masking high-SHAP features degrades surrogate agreement while masking low-SHAP features largely preserves it. A2C shows weaker separation, consistent with flatter importance spectra or a more modest surrogate fit, while DDPG and TD3 show intermediate patterns with AUM exceeding RAND but AIM not always below RAND.

Masking curves and the derived AIM/AUM scores, as shown in Table 4.5, indicate that, for PPO in particular, masking top-ranked features substantially degrades surro-

Table 4.5: Masking fidelity summary (AUC of surrogate test accuracy across masking fraction k).

Agent	Unmasked acc.	AIM AUC (↓)	AUM AUC (↑)	RAND AUC (ref.)
A2C	0.444	0.449	0.458	0.455
PPO	0.540	0.429	0.541	0.504
DDPG	0.508	0.515	0.532	0.496
TD3	0.476	0.518	0.513	0.461

Notes. Arrows indicate the preferred direction (↓ lower is better, ↑ higher is better). AIM↓ is better because masking the explainer-ranked most important features should reduce accuracy faster; AUM↑ is better because masking the least important features should preserve accuracy. RAND is a random-masking reference.

gate accuracy (low AIM). By contrast, masking low-ranked features leaves surrogate agreement largely intact (with higher AUM than RAND, with the expected separation AIM $<$ RAND $<$ AUM). This pattern conforms to the AIM/AUM semantics in Xiong et al. (2024) for tabular domains, and can be viewed as a functionally grounded proxy for the fidelity-oriented notions of explanation quality discussed in Doshi-Velez and Kim (2017); Lipton (2017). For the other agents, the separation between AIM and AUM is more modest, with A2C showing particularly flat importance spectra and DDPG and TD3 exhibiting intermediate patterns. This reflects both flatter global attributions and the additional noise introduced by financial time series (Lo and MacKinlay (1999); Tsay (2010)), but the ordering remains directionally consistent.

4.1.3.1 Stability under Perturbations (RIS)

RIS measures how much an explainer's attributions change under small, bounded perturbations to the input state, normalised by the input change. Lower RIS indicates smoother, more stable explanations. The XRL-Bench procedure is followed (Xiong et al. (2024)); bounded i.i.d. perturbations within a calibrated neighbourhood that preserves indicator ordering, evaluation on the same held-out window, and summary by the median and upper quantiles over time. Agent-explainer RIS summaries are consolidated in Table 4.6.

The RIS values in Table 4.6 show that SHAP, Integrated Gradients, and GradientSHAP maintain comparable levels of robustness under small perturbations, with slightly elevated instability around regime boundaries. Saliency achieves the lowest RIS (smoothest attributions) but does not exhibit the strongest masking fidelity in this setting. This illustrates that stability and fidelity are distinct dimensions of explanation quality, consistent with the separation of fidelity and stability metrics in Xiong et al. (2024) and with broader interpretability taxonomies that treat multiple evaluation criteria separately (Doshi-Velez

Table 4.6: Layer 2 perturbation stability (RIS) by agent and explainer on the held-out window. Lower RIS indicates smoother, more stable explanations. Values shown are medians over time.

Model	SHAP	GradSHAP	IG	Saliency
A2C	1.325	1.066	1.066	0.565
DDPG	1.025	0.931	0.868	0.484
PPO	1.059	1.026	1.044	0.563
TD3	1.099	0.946	0.914	0.496

and Kim (2017); Lipton (2017)). SHAP offers the most favourable balance between decision fidelity and stability, Integrated Gradients and GradientSHAP deliver similar but slightly noisier patterns, and Saliency is best interpreted as a supplementary, locally sensitive view.

On-policy agents (PPO, A2C) and off-policy agents (DDPG, TD3) show broadly similar RIS magnitudes, with variation across explainers larger than the differences between algorithm classes. RIS tends to increase modestly around regime transitions and remains comparatively lower within extended bull or sideways phases, which matches the intuition that attribution stability is challenged most during volatile shifts.

It is important to emphasise that AIM, AUM, and RIS provide relative checks on fidelity and stability under the chosen background distribution rather than formal guarantees of causal influence, and are therefore interpreted as comparative diagnostics across methods and agents rather than absolute measures of explanation quality.

Relative to existing explainable RL work in trading (Kumar et al. (2022)), the proposed framework extends evaluation from single-method attribution to cross-method consensus, and from predominantly visual inspection to explicit stability metrics (AIM/AUM/RIS). In contrast to Kumar et al. (2022), which focuses primarily on SHAP explanations for a single RL agent, this evaluation demonstrates that a multi-agent, multi-method attribution layer can still maintain strong agreement on salient indicators in a noisy financial environment, thereby strengthening claims of robustness.

4.1.3.2 Ordering

Overall, the results establish the empirical ordering

$$\text{SHAP} \;>\; \text{GradSHAP} \approx \text{IG} \;>\; \text{Saliency},$$

for the fidelity–stability trade-off within this financial setting, with AIM/AUM and RIS jointly indicating that SHAP-based attributions provide the most favourable balance between decision relevance and robustness.

4.1.4 Policy and Action Attribution (Layer 3)

Layer 3 moves from feature-level importance to behavioural interpretation by fitting decision-tree surrogates to each agent's state–action trajectories. Continuous controls are post-hoc discretized to three directional labels (*Sell*/*Hold*/*Buy*) and used to train depth-limited trees that approximate the agents' global decision structure. The goal is interpretability rather than optimisation, providing compact rule-like policies that can be related back to the feature-level evidence from Layers 1 and 2.

Table 4.7 summarises surrogate complexity per agent. All trees remain shallow with a modest number of leaves, supporting human auditability. Across agents, splits recurrently involve momentum (*RSI*, *MACD*), trend (*CCI*) and volatility (*turbulence*, *DX*) features, albeit with different hierarchical emphasis. On-policy agents (A2C, PPO) tend to place volatility gates near the root, while off-policy agents (DDPG, TD3) display slightly deeper hierarchies with additional momentum and trend checks.

Table 4.7: Surrogate complexity and leaf-level class balance. Lower depth and fewer leaves indicate simpler symbolic policies.

Agent	Max depth	Nodes	Leaves	Major leaf class
A2C	4	40	14	Hold
PPO	4	34	12	Buy
DDPG	4	37	13	Sell
TD3	5	43	15	Buy

From an evaluation standpoint, the surrogate-policy trees achieve moderate fidelity to the underlying agents. Consistent with the fidelity concepts in Section 2.6.2, fidelity is defined as the proportion of agent actions reproduced by the surrogate on the evaluation window. Global fidelity scores, as shown in Table 4.8, lie between 0.486 and 0.582 across agents, substantially above implicit majority-class baselines derived from the buy/hold/sell distributions. In the broader literature on policy distillation and interpretable policies, surrogate models are typically designed to preserve most of the original agent's performance while remaining transparent (Dispoto et al. (2025); Li et al. (2025)). In contrast, the present decision trees are deliberately shallow and operate on a three-way discretisation of continuous controls. Against this backdrop, fidelities around 0.5–0.6 still indicate that a non-trivial portion of the decision surface is compressible into human-readable rules, even in a noisy, multi-asset trading setting.

Regime-specific evaluation (Table 4.9) shows a consistent pattern. On-policy agents (A2C, PPO) are approximated more closely in bullish segments, whereas off-policy agents (DDPG, TD3) attain higher fidelity in sideways conditions. More broadly, these results

Table 4.8: Decision–tree surrogate fidelity and feature importances across RL agents (evaluation window July 2020–June 2021).

Model	Fidelity	Samples	Buy	Hold	Sell	MACD	RSI_{30}	CCI_{30}	DX_{30}	Turbulence
A2C	0.486	251	0.339	0.331	0.331	0.122	0.155	0.000	0.484	0.240
PPO	0.562	251	0.355	0.307	0.339	0.143	0.027	0.111	0.120	0.599
DDPG	0.582	251	0.323	0.331	0.347	0.184	0.275	0.151	0.078	0.312
TD3	0.570	251	0.375	0.171	0.454	0.110	0.299	0.374	0.086	0.131

echo evidence from RL-based trading studies that agent performance is strongly regime-dependent. Yang et al. (2020), for example, report that A2C performs better in bearish markets, while PPO and DDPG are preferred in bullish regimes. Choudhary et al. (2025) show that their risk-adjusted PPO-based RA-DRL strategy maintains profitability across sequences of bullish and flat market phases. Although these works do not report surrogate fidelity by regime, they motivate the need to evaluate policies with respect to market phase. The regime-wise surrogate analysis presented here therefore extends existing work by testing whether interpretability proxies remain stable across distinct market conditions, addressing calls in recent XAI reviews for more context-aware evaluation beyond aggregate metrics (Weber et al. (2024); Černevičienė and Kabašinskas (2024)).

Table 4.9: Regime-specific surrogate fidelity across trading agents.

Model	Bullish	Sideways	Samples (total)
A2C	0.684	0.517	240
PPO	0.621	0.490	240
DDPG	0.600	0.662	240
TD3	0.600	0.559	240

Methodologically, constraining each agent to a single shallow surrogate aligns with the argument of Doshi-Velez and Kim (2017) that interpretability often requires favouring simpler, lower-capacity models over maximising predictive accuracy. Li et al. (2025) take a complementary approach by constructing interpretable policies that are designed to preserve most of the original agent's performance. Because continuous portfolio controls are discretised into a single Buy/Hold/Sell label based on the mean action across assets, these surrogates are best interpreted as capturing directional risk shifts at portfolio level rather than stock-specific allocation decisions.

4.1.4.1 Action Tendencies

Because the surrogates are trained on discretised versions of the agents' continuous controls, any imbalance in Buy/Hold/Sell usage is reflected in the learned rules. Table 4.10 re-

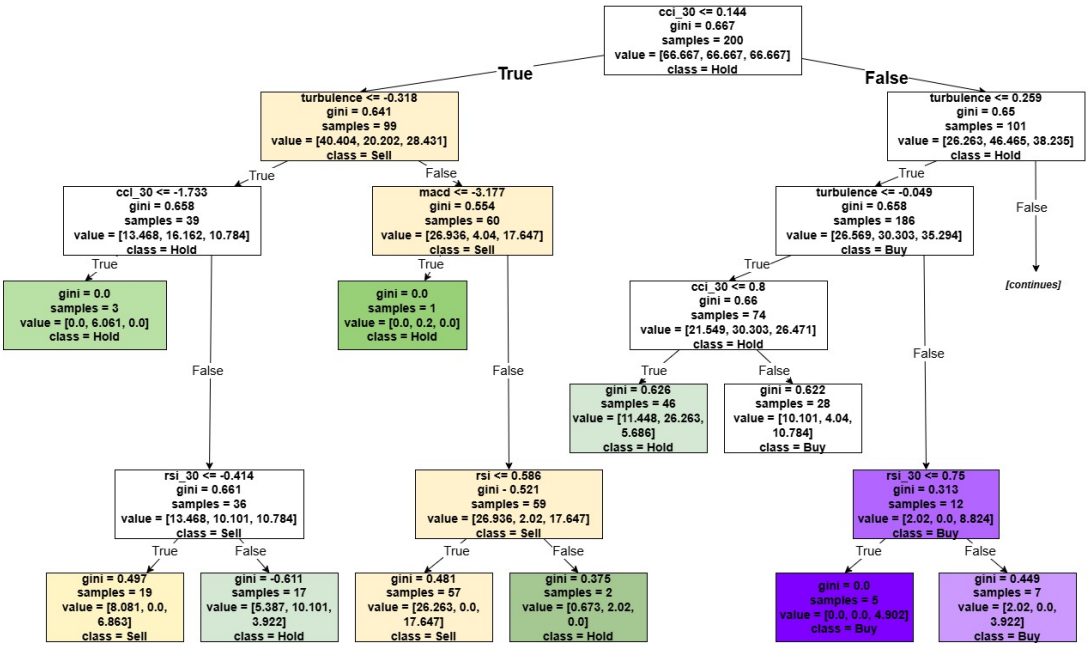

Figure 4.7: An example of a surrogate policy tree, particularly for the A2C agent.

ports the directional mix in the training trajectories, which helps explain why Hold leaves are more common for conservative agents. Leaf class frequencies describe the surrogate's terminal predictions, whereas the action proportions capture the agent's empirical behaviour after discretisation. The two distributions therefore need not coincide.

Table 4.10: Action distribution in the trajectories used to fit surrogate trees (post-hoc directional labels from continuous controls).

Agent	Buy	Hold	Sell
A2C	0.339	0.331	0.331
PPO	0.355	0.307	0.339
DDPG	0.323	0.331	0.347
TD3	0.375	0.171	0.454

4.1.4.2 Qualitative Regime Awareness

Reading the trees by regime yields consistent patterns. In trending periods, on-policy agents (A2C, PPO) often place a volatility gate near the root with momentum thresholds downstream, enabling decisive Buy states when turbulence is contained and RSI recovers. In sideways conditions, the off-policy agents (DDPG, TD3) exhibit deeper hierarchies with

additional momentum and trend checks, producing more selective Buy/Sell switching and fewer ambiguous Hold leaves. These qualitative patterns mirror the regime-wise fidelity differences in Table 4.9, where on-policy agents are approximated more closely in bullish windows and off-policy agents more closely in sideways markets.

4.1.4.3 Cross-layer Alignment

The dominant split variables in the trees coincide with the high-ranking, stable features identified in Layers 1 and 2, indicating that the symbolic policy rules are consistent with the underlying attribution patterns rather than artefacts of the surrogate fitting. Momentum and volatility factors appear repeatedly, and their thresholds align with the qualitative narratives in earlier sections. Combined with the global and regime-wise fidelity scores above, this suggests that Layer 3 captures a non-trivial and economically meaningful portion of each agent's policy in a form that can be inspected and communicated to end-users.

4.1.5 Reward Attribution and Performance Analysis (Layer 4)

Layer 4 completes the model-agnostic explainability framework by linking feature attributions to realised portfolio performance. Using the rolling-window decomposition described in Section 3.1.5.4, Layer 1 attributions are aggregated over overlapping 60-day mean-return segments to yield motif-level contribution profiles. The linear specification is adopted as a simple, descriptive decomposition of 60-day mean returns in terms of attribution-based motifs, rather than as a fully specified causal model of asset or portfolio returns.

Across agents and explainers, the decomposition shows that a relatively small subset of indicators accounts for most of the realised rewards. Momentum-based features (for example *MACD*, *CCI*, *RSI*) display sustained positive contributions in upward-trending periods, whereas volatility bands (*boll_lb*, *boll_ub*) frequently switch sign around regime boundaries, reflecting defensive rebalancing when prices approach envelope extremes. Directionality and trend motifs (*DX* and related slopes) provide more stable, moderate contributions. The resulting motif-level structure is summarised in Table 4.11, with per-window panels of the A2C model provided as an example in Figures 4.8–4.10.

Table 4.11: Summary of dominant Layer 4 contributors by agent. Rankings reflect the three most influential motif families by average absolute contribution over the test window.

Agent	Top contributor motif	Prevailing sign	Strongest in
A2C	Volatility bands	Positive	Sideways
	Directionality / trend	Positive	Bull
	Momentum	Negative	Sideways
DDPG	Volatility bands	Negative	Sideways
	Momentum	Positive	Bull
	Directionality / trend	Positive	Bull
PPO	Volatility bands	Positive	Bull
	Directionality / trend	Positive	Bear
	Momentum	Negative	Bear
TD3	Volatility bands	Positive	Bear
	Momentum	Negative	Bull
	Directionality / trend	Positive	Sideways

Notes. Motifs group related indicators (for example *RSI*, *CCI*, and *MACD* as Momentum; *DX* and moving-average slope as Directionality / trend; Bollinger lower and upper bands as Volatility bands). "Prevailing sign" is the mean contribution sign after aggregation across windows, and "Strongest in" is the regime where the motif has the largest absolute contribution.

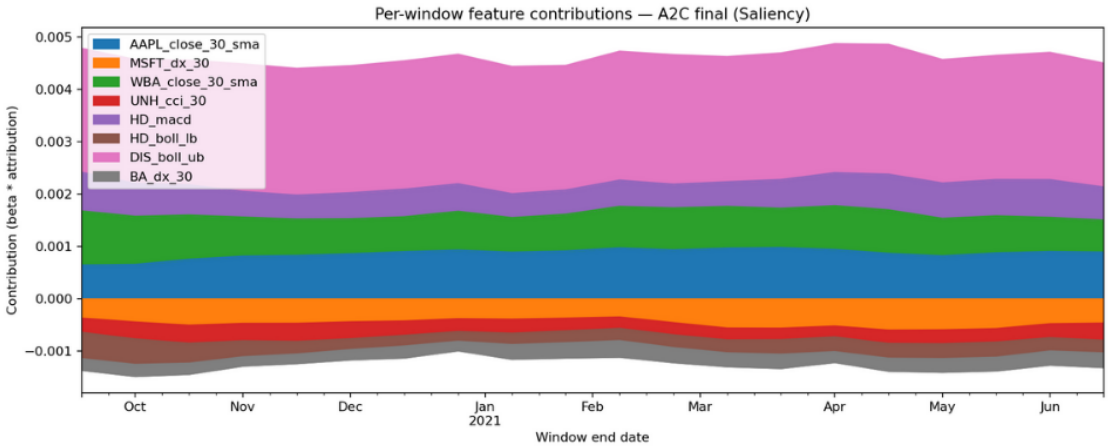

Figure 4.8: An example of per-window feature contributions for the A2C agent (Saliency, 60-day rolling window).

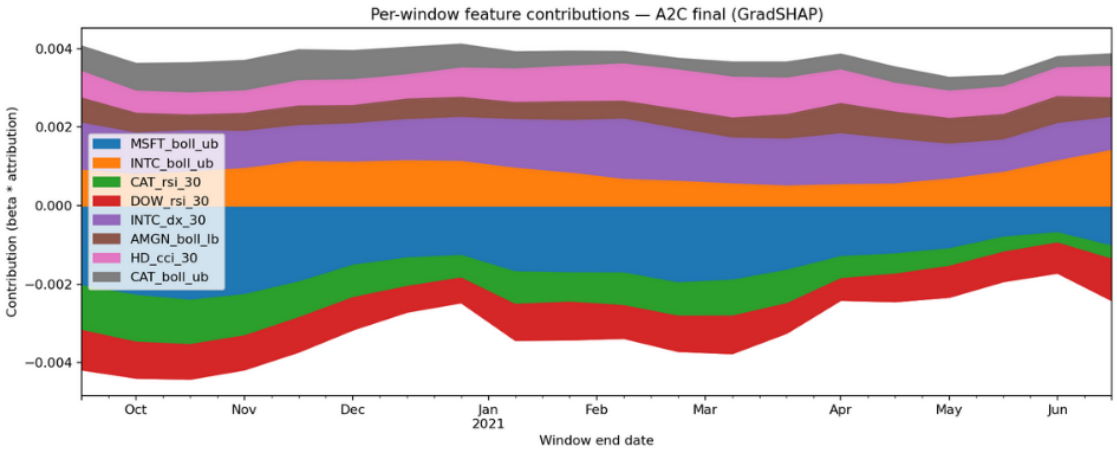

Figure 4.9: An example of per-window feature contributions for the A2C agent (GradientSHAP, 60-day rolling window).

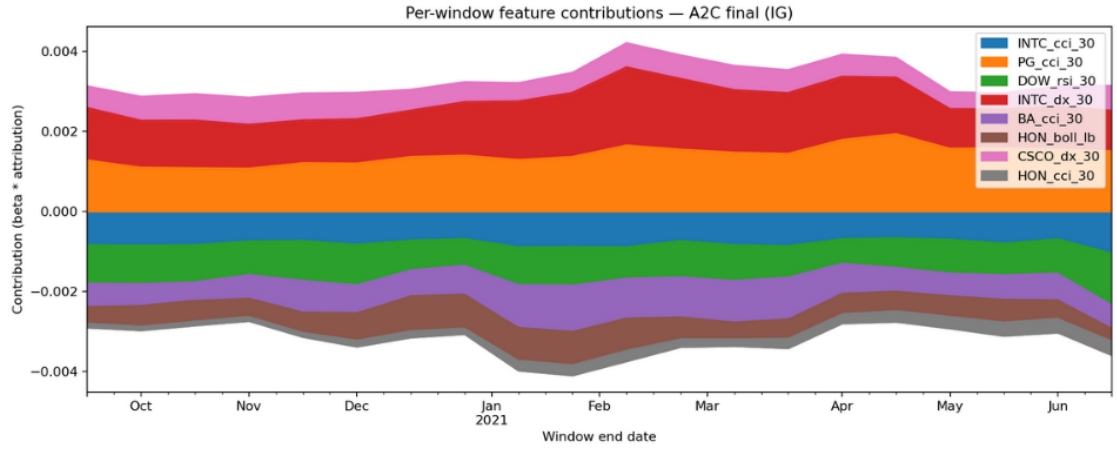

Figure 4.10: An example of per-window feature contributions for the A2C agent (Integrated Gradients, 60-day rolling window).

4.1.5.1 Cross-Explainer Agreement

Qualitatively, the ranking of dominant motifs is consistent across Saliency, GradientSHAP and Integrated Gradients. Momentum and directionality motifs co-lead in uptrends, while volatility bands expand around regime boundaries. Disagreements, where present, are local and short-lived, typically arising at turning points where sign inversions are expected. This broad agreement mirrors the stability signals in Layer 2 and supports the internal coherence of the framework.

Spearman correlations in Table 4.12 quantify this concordance. Integrated Gradients, GradientSHAP, and SHAP exhibit strong agreement at feature level ($\rho_{\text{feat}} \approx 0.86$–0.92)

across agents, with similarly high agreement at motif level. Saliency diverges at feature level but aligns more closely once contributions are aggregated into momentum, directionality, and volatility motifs. A similar pattern appears in XAI studies that use feature-attribution methods to link inputs to trading performance or returns (Cong et al. (2021); Kumar et al. (2022)), and surveys of XAI for financial time series report that perturbation-based and gradient-based methods often yield broadly consistent importance structures while highlighting complementary aspects of the signal (Arsenault et al. (2024)).

Table 4.12: Cross-explainer concordance (Spearman's ρ of $|\hat{\beta}|$).

	A2C		DDPG		PPO		TD3	
Pair	ρ_{feat}	ρ_{motif}	ρ_{feat}	ρ_{motif}	ρ_{feat}	ρ_{motif}	ρ_{feat}	ρ_{motif}
GradSHAP vs IG	0.903	0.895	0.895	1.000	0.897	0.895	0.919	1.000
GradSHAP vs SHAP	0.860	0.895	0.864	1.000	0.862	0.895	0.879	0.895
GradSHAP vs Saliency	0.070	0.205	0.113	0.667	0.025	0.205	-0.070	0.410
IG vs SHAP	0.873	1.000	0.864	1.000	0.883	1.000	0.886	0.895
IG vs Saliency	0.081	0.410	0.148	0.667	0.010	0.410	-0.060	0.410
SHAP vs Saliency	0.014	0.410	0.049	0.667	-0.046	0.410	-0.121	0.205

Note. Motif-level ρ can take on boundary values (for example 1.000) when the number of motif bins is small and many ranks are tied. These entries should be read as indicating very high agreement rather than perfect ordinal alignment.

4.1.5.2 Cross-Layer Alignment

The reward decomposition is consistent with the patterns observed in the earlier layers. The motifs that carry the largest and most persistent contributions in Layer 4 are the same momentum and directional-trend families that emerge as salient in Layer 1 and stable across regimes in Layer 2. In addition, the windows where contributions change sign coincide with phases in which the surrogate policies in Layer 3 shift towards more cautious actions, particularly when volatility-band indicators dominate.

Relative to financial XAI frameworks that focus on feature-importance and local explanation tools in forecasting and risk-management applications (Bussmann et al. (2020); Carta et al. (2022)), the present evaluation adds two elements. First, it explicitly checks whether the same indicator families drive reward attributions across multiple explainers and agents, providing an additional safeguard against method-specific artefacts. Secondly, working at both feature and motif levels mirrors the trend identified in recent surveys of XAI in finance to present explanations in terms of higher-level financial factors or feature categories rather than isolated signals (Arsenault et al. (2024); Černevičienė and Kabašinskas (2024)).

Taken together, the per-window reward panels and cross-explainer concordance results support the claim that the Layer 4 narratives about which indicators help or hurt performance are not artefacts of any particular explainer but reflect a stable, cross-method signal that is consistent with the earlier attribution and policy layers.

4.1.6 Discussion of Experiment 1 Findings

Experiment 1 set out to characterise the trading behaviour of the four RL agents in a realistic DJIA setting and to evaluate whether a structured, four-layer explainability stack can deliver a coherent account of their decisions, using a buy-and-hold benchmark as a financial reference point. Building on the design principles outlined in Section 3.1.5, the discussion below revisits each layer in turn and considers how the empirical findings support the layered approach, moving from financial plausibility through feature-level drivers and stability to policy structure and reward alignment.

4.1.6.1 Financial Behaviour as Context

The first requirement for any interpretability exercise is that the underlying agents exhibit financially credible behaviour. On the held-out window (July 2020 to June 2021), three of the four agents (A2C, PPO, TD3) exceeded the DJIA buy-and-hold benchmark on risk-normalised measures, while DDPG remained competitive in terms of cumulative return despite slightly lower Sharpe and Sortino ratios. TD3 combined strong risk-adjusted performance with volatility close to the benchmark, whereas A2C achieved the highest Calmar ratio through a favourable return–drawdown balance. By contrast, PPO delivered competitive returns at the cost of a deeper drawdown during late 2020.

This heterogeneity in performance profiles provides the starting point for the framework. The agents are sufficiently effective and diverse that it is meaningful to ask how they achieve these outcomes, and any explainability pipeline must account for both comparatively conservative and more aggressive trading styles rather than a single canonical policy.

4.1.6.2 Layer 1: Feature Drivers as a Common Vocabulary

Layer 1 addresses the foundational question of which market signals each agent treats as informative. Without this base, subsequent analyses of stability, policy logic, or rewards would lack an economically interpretable vocabulary. The attribution results show that a small, coherent set of indicators dominates across agents. Bollinger bands (*boll_lb*, *boll_ub*) organise risk-on versus risk-off exposure, while directional movement and mo-

mentum indicators (*DX*, *RSI*, *CCI*) modulate entry and exit tendencies and adjust positioning around regime boundaries. Rolling analyses further indicate that volatility-sensitive features strengthen around turning points, whereas momentum signals carry more weight in sustained trends.

These patterns align closely with standard technical narratives of trend-following and volatility management. They justify placing feature attribution and temporal dynamics at the base of the framework. Layer 1 converts opaque neural policies into a set of recurring indicator motifs that can be recognised and critiqued by domain experts, and that provide the shared language used in the higher layers.

4.1.6.3 Layer 2: Stability and Faithfulness as Validity Checks

However, feature importance profiles on their own could still be misleading if they were unstable or artefacts of a particular attribution method. Layer 2 was therefore designed as a validity check on Layer 1, focusing on whether the identified drivers are both causally influential for the agents' decisions and robust over time and regimes.

Masking-based fidelity metrics (AIM/AUM) show that, for A2C and PPO, masking features ranked as important by SHAP degrades performance more than random masking, while masking features deemed unimportant leaves performance relatively intact. DDPG and TD3 exhibit weaker but directionally consistent patterns, and Integrated Gradients and GradientSHAP broadly agree with SHAP on which features matter most. At the same time, RIS indicates that the most informative explainers are not arbitrarily volatile; on-policy agents display tighter stability distributions than off-policy agents, and Saliency, while very smooth, is revealed by the masking tests to be comparatively less discriminative.

Taken together, these results show why a dedicated stability and fidelity layer is necessary. Layer 2 does not introduce new drivers but tests whether the Layer 1 signals survive basic perturbation and method-robustness checks. In doing so, it supports the interpretation that volatility bands and momentum indicators are genuine explanatory signals rather than fragile artefacts and clarifies which attribution methods are most reliable in this setting.

4.1.6.4 Layer 3: Policy Structure for Auditability

Layer 3 responds to a different requirement. Stakeholders often need to see trading logic in a form that resembles human rules rather than heatmaps or importance scores. The aim is therefore to expose global policy structure without constraining or retraining the underlying agents. Decision-tree surrogates offer a compromise between fidelity and

transparency, mapping high-dimensional states to discretised Buy/Hold/Sell labels while retaining a compact rule set.

In Experiment 1, the learned surrogates remained shallow and achieved non-trivial fidelity (≈ 0.49–0.58) relative to majority baselines, indicating that they capture systematic aspects of the agents' behaviour while remaining readable. Crucially, their dominant split variables mirror the motifs highlighted in Layers 1 and 2, with thresholds on Bollinger bands, directional movement, and momentum indicators governing many of the key branches. Regime-wise fidelity patterns further indicate that on-policy agents are approximated more closely in trending conditions, whereas off-policy agents are better captured in sideways periods, reflecting their differing learning dynamics.

This layer therefore supports the claim that the framework can elevate feature-level insights into symbolic policy views suitable for audit and communication. The surrogates are not intended as replacements for the original policies, but as an additional lens that remains aligned with the driver structure established in the earlier layers.

4.1.6.5 Layer 4: Reward Alignment and Cross-Method Coherence

Finally, Layer 4 is motivated by the need to link explanations back to economically meaningful outcomes. Even stable, interpretable policies are of limited value if the highlighted mechanisms cannot be related to realised returns and risk. The rolling reward decompositions address this by regressing 60-day mean returns on normalised feature-level attributions and aggregating coefficients into motif families such as momentum, directionality, and volatility bands.

The resulting profiles show that a relatively small subset of motifs accounts for most of the variation in realised rewards and that their contribution signs behave in financially plausible ways. Momentum contributes positively in sustained uptrends, volatility bands switch sign around regime boundaries as agents de-risk near envelope extremes, and directionality terms provide more moderate but persistent contributions across conditions. Cross-explainer rank concordance further demonstrates that Integrated Gradients, GradientSHAP, and SHAP agree on these reward-relevant motifs at both feature and motif level, while aggregation mitigates some of the discrepancies introduced by Saliency.

Layer 4 thus completes the framework by tying the common driver set to performance and by checking that reward narratives are not artefacts of a single attribution technique. It provides the bridge from internal model reasoning to portfolio-level outcomes and supports the use of motif-level summaries as a stable unit of explanation.

4.1.6.6 Comparative Value of the Four-Layer Framework

Taken together, the four layers provide a more complete and reliable characterisation of the agents than any single explainability technique. A typical RL-for-trading study might report financial performance and one attribution view (for example a SHAP summary plot or a saliency heatmap), which reveals influential inputs but leaves open three questions; whether those attributions are faithful to the learned policy, whether they translate into interpretable decision rules, and whether they align with realised financial outcomes (Izzo (2022); Kumar et al. (2022); Weber et al. (2024); Yeo et al. (2023)).

The present framework addresses these gaps sequentially. Layer 1 shows that a compact and intuitive set of indicators (volatility bands plus directional and momentum oscillators) dominates behaviour across agents, already improving on purely visual inspection. Layer 2 then demonstrates that these drivers are not superficial; masking-based fidelity and RIS stability scores indicate that removing high-ranked features genuinely disrupts surrogate decisions, and that the most informative explainers remain reasonably stable within regimes. Without this layer, it would be difficult to distinguish robust explanatory signals from artefacts of a particular method or sampling window.

Layer 3 adds a policy-level perspective that is largely absent from single-layer analyses. The shallow decision-tree surrogates recover rule-like structures that mirror the motifs identified in Layers 1 and 2, and achieve non-trivial fidelity across regimes. This makes it possible to express parts of the agents' behaviour as transparent conditionals (for example tightening positions when volatility bands are breached) rather than only as importance scores, which is closer to how traders and auditors reason about strategies. Finally, Layer 4 links these motifs to realised rewards, showing that the same volatility and momentum structures that drive attributions also account for a substantial share of return variation, and that their contribution signs behave in financially plausible ways.

Across the four layers, the evaluation confirms that the explainability framework is both behaviourally credible and methodologically robust. Financial baselines are aligned with the FinRL replication and comparable RL trading studies, which ensures that the explanations are grounded in realistic agent behaviour rather than artefacts of an unrealistic environment. Layers 1 and 2 indicate strong cross-method agreement on salient drivers and stable importance structures under perturbation, consistent with XRL-Bench style criteria for fidelity and robustness. Layer 3 decision-tree surrogates achieve moderate fidelity with low complexity, and their regime-dependent performance matches known strengths of on-policy and off-policy agents in trending versus sideways markets. Layer 4 reward decompositions exhibit high cross-explainer rank concordance at both feature and motif levels, suggesting that global reward narratives remain consistent across attribution

methods.

Framed in this way, the value of the framework is not only that it offers four views instead of one, but that each additional layer constrains and cross-checks the previous ones. Feature-level importance feeds into stability tests, which in turn feed into surrogate policy rules and reward attribution; inconsistencies would be visible as breakdowns in this chain. In Experiment 1, the empirical results and evaluation evidence support the design rationale introduced in Section 3.1.5. The same small set of indicator motifs recurs across layers and agents, passes basic robustness checks, can be summarised as readable decision rules, and is tied to economically meaningful performance differences.

This layered structure is also what enables the second stage of the thesis. Because the framework produces consistent, motif-level descriptors of behaviour, the outputs of Experiment 1 can be used as a structured substrate for LLM-based narratives in Experiment 2. In that sense, the four-layer explainability stack is not only more informative than single-method XAI for Experiment 1, but also provides the necessary scaffolding for the communicative, human-centred explanations evaluated later in the study.

4.2 Experiment 2 Results & Evaluation

This section reports outcomes from the second experimental stage, which investigates whether LLM-generated explanations can improve the accessibility and perceived clarity of RL trading decisions derived from Experiment 1. Visual and quantitative artefacts such as global and rolling attributions, policy surrogates, reward decompositions, and regime tags were translated into audience-tailored natural-language explanations.

4.2.1 Inputs, Prompting, and Output Protocol

Novice and expert explanations were generated from the structured XAI corpus described in Section 3.2, using four prompting modes (baseline, few-shot, historical, counterfactual) across three providers (Claude Sonnet 4, GPT-4o, GPT-5 Chat). The prompts incorporated the local decision context (state features, regime, action, reward and layer output) and applied the future-information constraint and leakage controls specified in Section 3.2.1.

To mitigate temporal leakage and memorisation issues highlighted in recent financial LLM work (Kang and Liu (2023); Lopez-Lira et al. (2025)), the three-stage future-information constraint and post-hoc filtering protocol in Section 3.2.2 are applied when assembling prompts. This enforces strict temporal cut-offs and removes leaked future indicators, so that explanation quality is assessed only on information available at decision time, in line with recommendations for post-cutoff evaluation and tightly controlled

query templates in financial LLM studies. As a result, the quantitative results reported below can be interpreted as a conservative estimate of the quality of the explanation under realistic deployment constraints.

4.2.2 Overview of Experiment 2 Outputs

The consolidated Experiment 2 corpus aggregated, for each local trading context, the observed state features, regime label, executed action, realised reward, and the structured explainability outputs from Experiment 1. In total, 30,276 distinct decision states were available, with regime frequencies reported in Table 4.13. The distribution is dominated by bull conditions, with comparatively fewer bear instances.

Table 4.13: Decision states by regime in the Experiment 2 corpus.

Regime	**No. of decision states**
Bear	120
Bull	20,524
Sideways	9,632
Total	**30,276**

As discussed in Section 3.2.1, for the main generation runs, a stratified sample of 48 decision states per regime was drawn, yielding 144 local trading contexts. These contexts were reused across all three providers and paired with both audience types and all prompting modes, so that every provider contributed explanations under the full set of audience–prompt combinations.

Automated evaluation covered three groups of metrics. First, lexical and semantic overlap with pseudo-references was quantified using BLEU, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR. Second, readability and linguistic clarity were assessed via Flesch Reading Ease, Gunning Fog, a derived clarity score, and lexical density. Third, factual grounding and alignment with the numerical substrate were captured through a hallucination rate, an allowed-mentions rate, and a reward-alignment score relating explanation tone to realised return direction.

4.2.3 Quantitative Quality and Factuality

As outlined in Section 3.2, Experiment 2 treats prompt design as a controlled ablation over three factors that are expected to shape explanation quality; audience framing (novice, expert), prompting strategy (baseline, few-shot, historical, counterfactual, and a synthesis few-shot+historical condition), and LLM provider. Audience framing operationalises the

human-centred design requirement that explanations be tailored to users with different levels of financial and technical expertise, while the prompting modes vary the amount of exemplification and temporal context provided in the templates. The provider dimension tests whether these patterns are robust across model families when the same structured XAI substrate and guardrails are used.

The underlying factual substrate is fixed by the corpus of decision states and XAI artefacts described in Section 3.2.1, with only the language realisation changing across conditions. Clarity, hallucination, and reward-alignment scores follow the automated evaluation metrics defined in Section 3.2.3 and are summarised here at a descriptive level.

Figure 4.11 summarises mean clarity across the audience–prompting grid. Novice framing increases readability and clarity across all providers, as expected. Few-shot prompting yields the highest pooled clarity in both audiences and regularises style, while historical prompting reduces clarity but produces more temporally anchored narratives. Counterfactual prompting introduces contrastive cues and slightly lowers clarity on average, although reward alignment typically remains high. The synthesis configuration sits close to historical in clarity, trading some simplicity for richer temporal context.

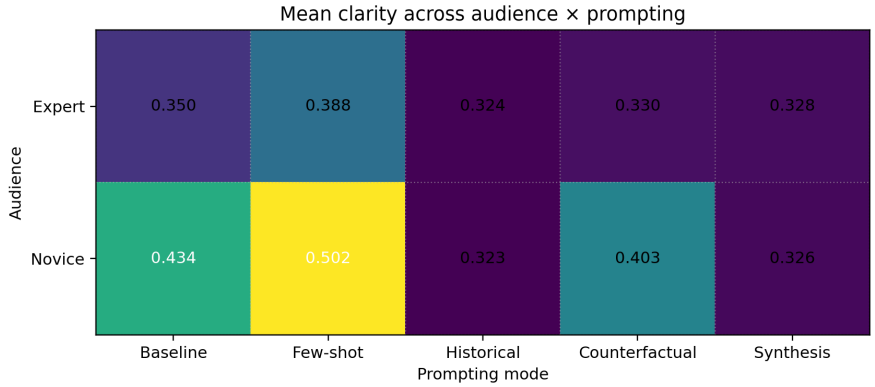

Figure 4.11: Mean clarity across the audience × prompting grid.

The relation between hallucination and reward alignment is illustrated in Figure 4.12. Lower hallucination rates tend to coincide with higher alignment, but several cells attain near-perfect alignment despite moderate hallucination, indicating that lexical grounding and return-consistency capture complementary dimensions of explanation quality rather than a single continuum.

These descriptive tendencies are reflected in Table 4.14 and the error profiles in Figure 4.13, which together provide a more granular view of how audience framing, prompting strategy, and provider affect explanation quality and factual grounding.

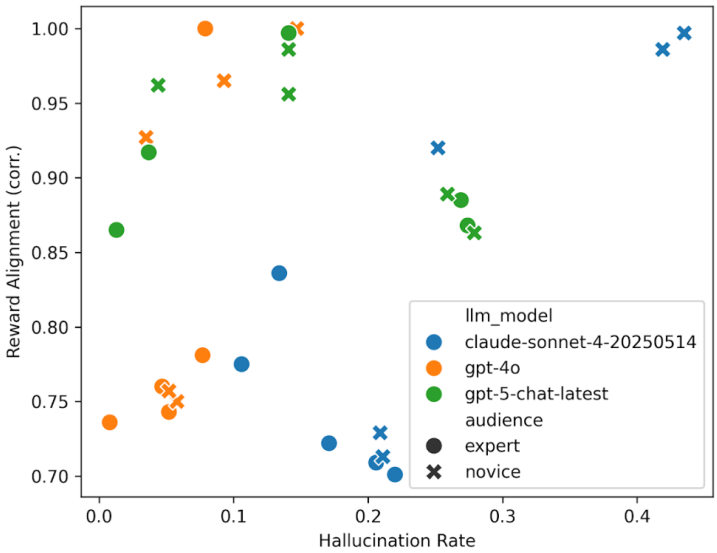

Figure 4.12: Hallucination rate versus reward alignment by provider and audience. Lower hallucination is generally associated with higher alignment, although some high-alignment cells retain moderate hallucination.

Table 4.14 summarises readability, clarity, lexical density, and hallucination rate across models, audiences, and prompting modes. Reported hallucination values are means of the per-explanation rate h, where $h = 0.1$ indicates that, on average, one in ten extracted feature-like mentions per explanation fell outside the decision's allowed feature vocabulary. Three patterns are most relevant for evaluation.

Table 4.14: Quantitative evaluation metrics for generated explanations across prompting modes.

Model	Audience	Few	Hist.	CF	Read.	Clarity	Lex. Dens.	Halluc.
Claude-Sonnet-4	Expert	No	No	No	33.99	0.362	0.809	0.106
		No	No	Yes	28.07	0.320	0.783	0.134
		No	Yes	No	38.30	0.337	0.718	0.206
		Yes	No	No	28.15	0.343	0.725	0.171
		Yes	Yes	No	39.64	0.350	0.721	0.220
	Novice	No	No	No	38.80	0.333	0.851	0.419
		No	No	Yes	43.73	0.407	0.822	0.435
		No	Yes	No	37.91	0.331	0.718	0.209
		Yes	No	No	50.54	0.467	0.861	0.252
		Yes	Yes	No	38.44	0.345	0.723	0.211

Continued on next page

Model (Provider)	Audience	Few	Hist.	CF	Read.	Clarity	Lex. Dens.	Halluc.
GPT-4o	Expert	No	No	No	37.22	0.366	0.835	0.008
		No	No	Yes	41.57	0.305	0.828	0.079
		No	Yes	No	40.51	0.324	0.724	0.047
		Yes	No	No	36.70	0.369	0.841	0.077
		Yes	Yes	No	40.49	0.320	0.723	0.052
	Novice	No	No	No	53.58	0.499	0.861	0.035
		No	No	Yes	52.49	0.384	0.839	0.147
		No	Yes	No	40.20	0.319	0.731	0.058
		Yes	No	No	58.70	0.508	0.873	0.093
		Yes	Yes	No	39.84	0.320	0.729	0.052
GPT-5 (Chat)	Expert	No	No	No	34.06	0.321	0.649	0.037
		No	No	Yes	42.76	0.365	0.722	0.141
		No	Yes	No	40.20	0.310	0.592	0.269
		Yes	No	No	45.56	0.453	0.742	0.013
		Yes	Yes	No	39.95	0.315	0.592	0.274
	Novice	No	No	No	54.73	0.469	0.728	0.044
		No	No	Yes	52.48	0.417	0.778	0.141
		No	Yes	No	40.02	0.318	0.584	0.259
		Yes	No	No	65.48	0.530	0.871	0.141
		Yes	Yes	No	39.27	0.314	0.596	0.279

First, novice framing consistently increases readability and clarity, particularly under few-shot prompting. GPT-4o and GPT-5 (Chat) achieve Flesch-style readability scores that place novice narratives in the "accessible" band, while maintaining high lexical density, in line with the readability targets in Section 3.2.3.3. This pattern is aligned with work on financial and XAI explanation design, which emphasises tailoring explanations to the needs and expertise of different stakeholder groups and shows that natural-language narratives for non-technical users can improve perceived clarity while still conveying the key model drivers (Weber et al. (2024); Zytek et al. (2024)).

Second, hallucination rates are lowest for GPT-4o across both audiences, especially in baseline and historical modes, whereas Claude Sonnet-4 and GPT-5 (Chat) exhibit higher variance under counterfactual prompting. Few-shot and historical prompts tend to reduce hallucination and allowed-mention violations, echoing broader findings that more structured prompts, retrieval augmentation, and tool-based grounding can improve factuality in financial question answering and explanation tasks (Kang and Liu (2023)).

Third, configurations with lower hallucination tend also to show higher reward align-

ment and lower allowed-mention violations (Figure 4.13), suggesting that the pipeline-level constraints described in Section 4.2.2 are effective in keeping narratives close to the underlying XAI substrate. This multi-metric view aligns with the recommendation in Zytek et al. (2024) that evaluation of LLM-based explanations should combine proxies for factuality, coverage of important factors, contextual grounding, and linguistic quality rather than relying on a single score.

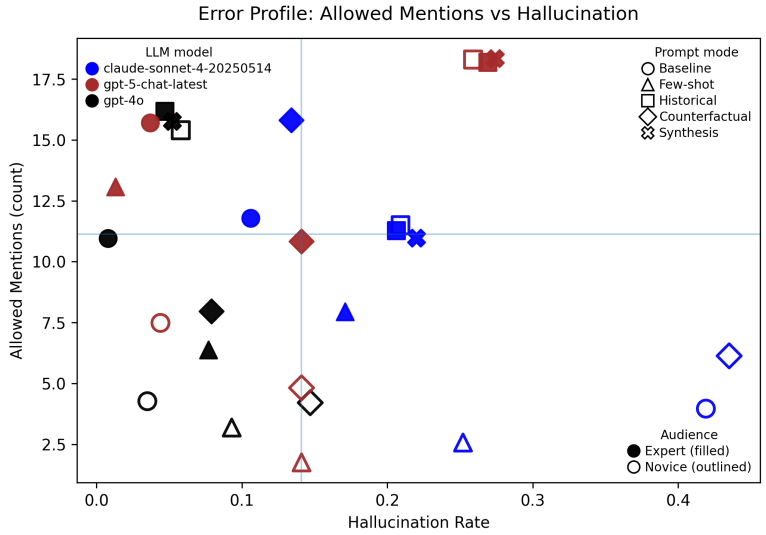

Figure 4.13: Error profile: hallucination rate versus allowed-mention violations (lower is better on both axes).

4.2.4 Automated Linguistic Metrics

To quantify lexical and structural similarity between generated explanations and pseudo-references constructed from the XAI corpus, METEOR, ROUGE-1, ROUGE-2, ROUGE-L, and corpus-level BLEU were computed for each provider, audience, and prompting mode. The main descriptive patterns are summarised here, with the compact leaderboard retained and the full per-cell table deferred to Appendix D.

Table 4.15 lists the five strongest provider–audience–prompting combinations ranked by METEOR, with ROUGE and BLEU used as tie-breakers. Across providers, historical and synthesis (few-shot+historical) prompts tend to achieve the highest overlap scores, especially for Claude Sonnet 4 and GPT-5 Chat, reflecting their more frequent reuse of regime and feature tokens from the factual substrate. Baseline and counterfactual prompts yield lower overlap on average, while novice prompts generally score slightly higher than ex-

pert prompts for a given provider and mode. The full per-configuration leaderboard is reported in Appendix D.

Table 4.15: Overall top-five cells by overlap (primary key = METEOR, tie-breakers = ROUGE-L then ROUGE-2).

Model	Audience	Prompt mode	METEOR	ROUGE-L	ROUGE-2	ROUGE-1	BLEU
Claude-Sonnet-4	Expert	Historical	0.227	0.182	0.082	0.219	3.80
Claude	Novice	Historical	0.227	0.183	0.083	0.219	3.99
GPT-4o	Novice	Baseline	0.258	0.158	0.082	0.186	2.20
GPT-4o	Expert	Historical	0.193	0.150	0.059	0.180	2.64
GPT-5 Chat	Expert	Historical	0.163	0.167	0.061	0.205	2.84

Novice prompts tend to increase overlap with the pseudo-references relative to expert prompts, especially for GPT-4o. Historical and synthesis prompts raise ROUGE and METEOR across models, reflecting closer alignment with the structured factual context. Few-shot prompting regularises style but does not always maximise overlap, as it encourages greater diversity. Counterfactual prompting lowers overlap on average, due to contrastive wording only partially reflected in the pseudo-references.

Lexical overlap metrics (BLEU, ROUGE-1/2/L, METEOR) provide a complementary view on explanation quality. As described in Section 3.2.3.1, pseudo-reference texts encode regime, action, key features, and outcome information derived from Experiment 1 artefacts, and scores computed with *sacrebleu*, *rouge_score*, and *nltk.meteor* quantify how closely generated explanations follow this structured reference.

Overall, BLEU and METEOR move together (Figure 4.14), indicating coherent behaviour across n-gram and semantics-aware overlap metrics. Novice-targeted, few-shot and historical prompts generally attain higher METEOR and ROUGE-L scores than expert prompts, reflecting more regularised phrasing and closer adherence to the reference schema. Counterfactual prompts exhibit deliberately lower lexical overlap due to contrastive constructions but remain reward-aligned, underscoring the limitation that overlap-based metrics may under-credit faithful paraphrases (Zhang et al. (2020b)).

It is important to note that overlap metrics primarily reward lexical similarity and can underestimate the quality of faithful paraphrases. Corpus-level BLEU is expectedly low in this single-reference setting, whereas METEOR and ROUGE-L are more informative because they incorporate stemming, synonymy, and sequence structure. In this study, lexical metrics are treated as auxiliary rather than primary indicators that complement

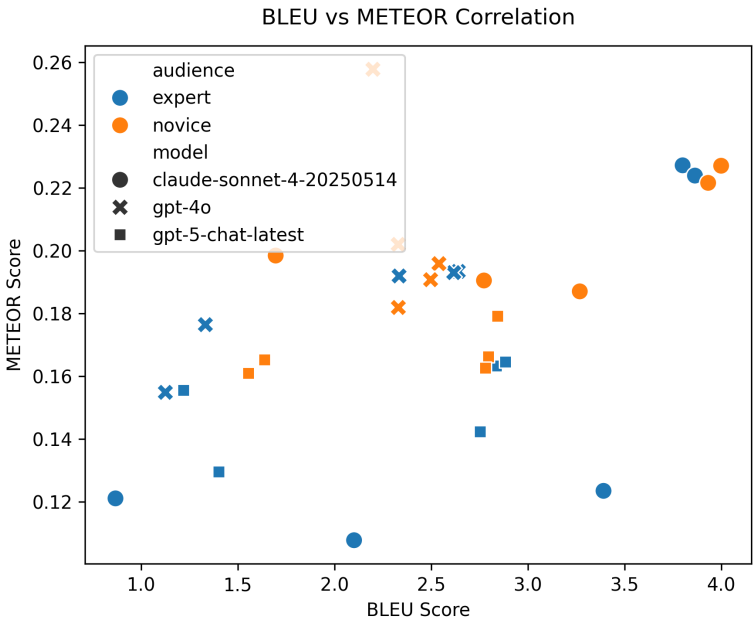

Figure 4.14: Association between BLEU and METEOR across runs. Points cluster along a positive trend, indicating metric coherence.

readability, clarity, hallucination, and reward alignment, supporting the conclusion that audience-conditioned prompts can deliver explanations that are stylistically coherent and well aligned with the underlying decision context.

4.2.5 User Study

4.2.5.1 Design and Constructs

The human centred evaluation examined how novice participants experienced the traditional visual artefacts (TE) from Experiment 1 and the LLM generated narratives (NLE) derived from the same underlying content. Seven explanation cards were constructed, each pairing a TE view (for instance, rolling feature attributions, surrogate policy trees, reward attribution plots, and regime timelines) with a matched NLE. For each card, participants first viewed the TE, then the NLE, and subsequently answered a small set of Likert scale questions (1–5) targeting perceived 'clarity', 'trust', 'transparency', 'usefulness', and 'actionability', as well as directly asking which format was more easily understandable (visual, narrative, both, or neither). A post study block then captured global ratings on clarity, trust, and actionability, as well as format preferences for clarity, trust, reasoning, feature identification, and preferred explanation style in a hypothetical trading assistant. The full

item wordings and response distributions are provided in Appendix F.

Twelve novice participants completed the study. Inclusion criteria required age ≥ 18, English proficiency, and some involvement in finance, understood as having studied economics or finance, working in a related industry, or having prior experience trading in the stock market. Participation was anonymous and uncompensated.

4.2.5.2 Overall Ratings

On average, participants rated the explanations as 'clear' ($\bar{M}_{\text{Clarity}} = 4.08$), with 'actionability' also favourable ($\bar{M}_{\text{Action}} = 3.92$) and 'trust' somewhat lower but still positive ($\bar{M}_{\text{Trust}} = 3.75$).

Card level items show the same pattern. Clarity oriented statements (for example, "The explanation was clear and easy to understand") averaged $\bar{M}_{\text{clarity,card}} = 4.38$ across cards, while trust oriented items (for example, "I would trust this explanation in a real trading scenario") averaged $\bar{M}_{\text{trust,card}} = 3.13$. This indicates that participants generally found the material easy to follow but were more cautious when judging whether they would rely on the explanations for real trading decisions. A majority selected the 'easy' or 'very easy' options on the global "Overall, how easy were the explanations to understand?" item, suggesting that most respondents regarded the combined explanation suite as easy to understand.

4.2.5.3 Format and Card Level Preferences

Post study format preferences are summarised in Table 4.16. Participants were asked which format they found most helpful for clarity, trust, reconstructing the agent's reasoning, and identifying important features.

Table 4.16: Post study format preferences by construct ($n = 12$).

Preference	Clarity	Trust	Reasoning	Features
Visual	6	5	0	5
Narrative	1	2	5	4
Both	5	5	7	3
Neither	0	0	0	0

For clarity, participants were split between visual only (6) and the combined format (5), with narratives alone rarely chosen. Trust was similarly divided between 'visual only' and 'both' (5 each), suggesting that some respondents were comfortable trusting a well designed chart, whereas others preferred corroboration from a narrative. For reasoning,

no participant preferred 'visuals' alone; seven chose 'both' and five chose 'narrative', indicating that text plays a central role in reconstructing the agent's logic. For feature importance, 'visual' regained a slight lead, but 'narrative' and combined formats still attracted substantial support, reflecting a desire for both graphical salience and verbal confirmation.

Card level preferences for understandability are reported in Table 4.17. For each of the content cards, participants selected which modality they found more easily understandable.

Table 4.17: Card level format preferences for understandability ($n = 12$).

Content	Visual	Narrative	Both	None
Rolling SHAP feature contributions (time varying reward influence)	0	2	9	1
Decision tree policy surrogate (Buy–Hold–Sell rules on indicators)	8	2	2	0
Integrated Gradients feature level reward contributions (stacked over time)	4	2	4	2
Attribution stability (Rolling IG heatmap of importance stability)	5	2	5	0
Reward attribution bar chart (feature level net contributions)	5	1	6	0
Market regime timeline (DJIA trend and regimes, 2009–2021)	3	1	8	0
Cross model feature importance comparison (PPO, DDPG, TD3, A2C indicators)	4	3	5	0

Two tendencies emerge. First, the combined format ('both') is the single most frequent choice across cards, particularly for temporally rich or context heavy content such as Rolling SHAP and the regime timeline. Secondly, preferences vary with the type of artefact; the decision tree card is dominated by the visual only option, whereas cards involving more complex temporal or contribution structures show stronger demand for the combined visual plus narrative presentation. Only three 'None' responses were recorded, all on the more complex attribution cards.

Finally, when asked which format they would prefer in a real trading assistant, eleven of twelve participants chose the combined visual plus narrative format and one chose narrative only. None preferred visuals alone.

4.2.5.4 Qualitative Impressions

Open ended comments broadly mirror these quantitative patterns. Several participants described the narrative as "help[ing] give context to the visual explanation" and noted

that having both formats made it easier to link specific time segments or indicator movements in the chart to statements in the text. Others remarked that some charts and rules were only straightforward to interpret “as long as you know what the indicators mean”, underscoring the impact of jargon and assumed prior knowledge on accessibility.

A recurring suggestion was that explanations should adapt to user expertise, with simpler wording and fewer abbreviations for novices and more detailed, indicator specific commentary for advanced users. Participants were generally positive about an assistant that explains trading decisions in this way, but stressed that they would use it to supplement, rather than replace, their own judgement and external information sources.

4.2.5.5 Interpretation of User Study

In summary, the user study indicates that the explanation suite is perceived as clear and practically useful, with moderate but not unconditional trust. Participants consistently preferred a combined visual plus narrative format, especially for richer temporal and attribution views, while relying more heavily on text to reconstruct reasoning. These findings provide a human centred complement to the automated metrics from Experiment 2.

These patterns can be interpreted in light of human centred XAI evaluation frameworks. The study evaluates whether the explanation suite improves perceived clarity, trust, and actionability for novice traders, consistent with human centred evaluation criteria in Babaei et al. (2022); Dikmen and Burns (2022); Doshi-Velez and Kim (2017); Hoffman et al. (2018). The questionnaire operationalises five constructs (clarity, trust, transparency, usefulness, and actionability) using short Likert scales, forced choice format preferences, and open ended questions. The design of the items is grounded in interpretability frameworks that distinguish between understanding, appropriate trust, and practical decision support (Doshi-Velez and Kim (2017); Hoffman et al. (2018); Lipton (2017)).

The composite scores in Section 4.2.5.2 indicate that the participants rated the explanations as clear on average (Clarity ≈ 4.1), with moderately positive ’actionability’ and slightly lower ’trust’. This pattern mirrors observations in financial applications and surveys, where explanations are viewed as helpful for understanding but stakeholders remain cautious about fully delegating decisions (Bussmann et al. (2020); Weber et al. (2024)).

Format preferences (Table 4.16 and Table 4.17) show a strong revealed preference for combined visual and narrative explanations, particularly when reconstructing reasoning or understanding time varying structure. No participant chose visuals alone as the preferred assistant format. This result is consistent with the integrative XAI plus NLP frameworks discussed in Section 2.4 (Arsenault et al. (2024); Zytek et al. (2024)), which argue that multi modal explanations can reduce cognitive load and improve usability compared with

purely graphical or purely textual displays.

Qualitative feedback supports this interpretation. Participants noted that narratives "add context" to charts but also warned about jargon and over-interpretation, echoing concerns about domain knowledge and trust in Dikmen and Burns (2022). The gap between high clarity and more cautious trust suggests that readable, coherent text is necessary but not sufficient for confidence. Users still expect close alignment between narrative emphasis, visual evidence, and domain intuition.

Given the modest sample size ($n = 12$), formal psychometric reliability estimates (Cronbach's α or McDonald's ω) would be unstable and are therefore not reported. As noted by Hoffman et al. (2018), explanation quality metrics and instruments should not be over-interpreted without adequate validation, and the present choice is in line with these concerns.

4.2.6 Summary and Discussion of Experiment 2

The evaluation of the LLM explanations yields three main conclusions that complement the automated metrics and user-study findings reported above.

Firstly, audience-aware prompting improves readability and clarity, particularly for novice participants, without substantially increasing hallucination or weakening reward alignment. This aligns with human-centred XAI guidance that explanations should be tailored to their audience and use case (Weber et al. (2024)) and with LLM-for-XAI studies that emphasise the importance of prompt design and narrative framing for effective communication (Zytek et al. (2024)).

Secondly, combined textual and visual explanations are strongly preferred to either modality alone. This reinforces the integrative frameworks surveyed in Section 2.4 and suggests that multimodal interfaces are a natural design point for XRL in finance (Arsenault et al. (2024); Weber et al. (2024); Zytek et al. (2024)), especially when users need to understand temporal structure and regime-dependent behaviour.

Finally, automated metrics (readability, lexical overlap, hallucination, reward alignment) and human-centred judgements (clarity, trust, actionability) point in broadly similar directions but are not interchangeable. Human participants are more sensitive to jargon, over-interpretation, and misalignment between narrative and charts than lexical metrics alone can capture, echoing the cautions raised in Zytek et al. (2024). In that sense, Experiment 2 positions the synthesis layer within recent work on LLMs for XAI (Zytek et al. (2024)) and finance-focused LLM risk and evaluation (Kang and Liu (2023); Lopez-Lira et al. (2025)), showing that grounded, audience-aware narratives can enhance interpretability while remaining closely tied to the underlying technical artefacts.

4.3 Overall Interpretation and Evaluation

Across both experimental stages, the results show that a model-agnostic multi-layer explainability framework can render RL-based trading agents substantially more transparent without sacrificing financial performance. Experiment 1 demonstrates that feature attribution, stability analysis, policy surrogates, and reward decomposition jointly provide a coherent picture of how agents use volatility- and momentum-sensitive indicators, how stable these signals remain under perturbation and regime shifts, how they translate into compact rule-like policies, and how they ultimately relate to realised rewards.

Experiment 2 extends this technical transparency into communicative interpretability. Structured XAI artefacts are transformed into audience-conditioned natural-language explanations that exhibit acceptable overlap with factual pseudo-references, favourable readability and clarity properties, and controllable hallucination rates. The user study indicates that novice participants find these explanations clear, moderately trustworthy, and actionable, with a strong preference for combined visual and narrative presentations when understanding why a trading decision was made. Together, these findings support the view that faithful, artefact-grounded narratives can bridge part of the gap between opaque RL agents and human users.

Taken together, the evaluations of both experiments indicate that the framework addresses both dimensions of XAI evaluation outlined in Section 2.6; technical fidelity to the underlying RL agents and human-centred interpretability for end-users.

On the technical side, the four-layer stack meets key criteria identified in systematic reviews of financial XAI (Weber et al. (2024); Yeo et al. (2023); Černevičienė and Kabašinskas (2024)). It combines fidelity and stability metrics grounded in established benchmarks (Xiong et al. (2024)), interpretable surrogates that capture policy behaviour without excessive complexity, and reward-level diagnostics that relate attributions to economically meaningful outcomes. Relative to studies that rely on a single explainer or predominantly visual inspection (Cong et al. (2021); Guan and Liu (2021); Kumar et al. (2022)), this framework offers a more systematic, multi-layer evaluation that is robust across methods, agents, and regimes.

On the human side, the LLM layer and user study extend recent proposals for integrating XAI and natural language generation in algorithmic decision-making (Arsenault et al. (2024); Zytek et al. (2024)).

The evaluation shows that grounded, audience-tailored narratives can raise perceived clarity and actionability, particularly when combined with visual artefacts, but that trust remains contingent on careful control of jargon, over-interpretation, and alignment with visible evidence. This aligns with observations in financial risk-management and credit-

scoring XAI (Babaei et al. (2022); Bussmann et al. (2020); Demajo et al. (2020)), where users value explanations but insist on cross-checking them against domain knowledge and independent data.

Overall, the evidence supports the central claim of the study; a layered XAI framework, evaluated with both quantitative metrics and human-centred studies, can help bridge the gap between technically faithful RL explanations and those that are usable and trustworthy for investors and other stakeholders. These integrated findings provide the basis for the concluding reflections on contributions, limitations, and avenues for future work in the following chapter.

5 Conclusions

5.1 Revisiting the Aims and Objectives

This study set out to enhance the transparency and interpretability of RL for algorithmic trading on DJIA daily data by combining a model-agnostic XAI framework with LLM-based natural language explanations. It delivered a four-layer technical framework that diagnoses and validates explanations across feature attribution, stability, policy surrogacy, and reward decomposition, together with a synthesis pipeline that grounds LLM outputs in XAI artefacts and evaluates them using both automated and human centred metrics.

Across the RL models on DJIA data, Layer 1 identified convergent behavioural drivers via SHAP, IG, GradientSHAP, and saliency. Layer 2 showed the expected masking pattern, with lower AIM and higher AUM, and lower RIS for smooth temporal attributions. Layer 3 produced compact surrogate trees with credible fidelity, indicating parts of the policy surface are compressible into symbolic rules. Layer 4 linked decomposed contributions to realised returns and revealed regime sensitive motifs consistent with Layer 1.

On the LLM layer, automated evaluation showed that few shot prompting produced the most stable readability across models and modes, while configurations with lower hallucination rates also achieved stronger reward alignment. This pattern is consistent with the view that tighter factual control over the explanation content helps maintain proximity to the underlying XAI substrate and preserves faithfulness to the agent's realised behaviour.

The exploratory user study complemented these quantitative findings. Participants rated the explanation suite as clear, moderately trusted, and practically useful, and generally preferred hybrid presentations that combined traditional visual explanations with natural language narratives over text only or visual only formats. Grounded narratives were particularly valued when reconstructing the agent's reasoning and understanding how salient features contributed to actions. Taken together, the automated metrics and user feedback indicate that the proposed approach maintains fidelity to RL behaviour while improving communicative accessibility for diverse audiences.

5.2 Critique and Limitations

The study is limited to a single equity universe (DJIA) and one historical window and a fixed random seed, which constrains external validity to other assets, horizons, and market microstructures. Although chronological splits and post-hoc leakage checks are used, alternative reward definitions, transaction-cost assumptions, and execution frictions could shift absolute performance and the apparent salience of drivers. Surrogate-tree fidelity also depends on the chosen discretisation and tree capacity, and different surrogate designs may alter the interpretability–fidelity trade-off.

On the explanation side, automated text metrics mainly capture style and lexical overlap and only approximate semantic faithfulness, while human judgements are based on a small, novice sample ($n = 12$) and descriptive rather than inferential analysis, so findings on clarity, trust, and usefulness should be viewed as indicative rather than generalisable. Finally, the LLMs are provider-hosted black-box models whose behaviour may drift with vendor updates, motivating periodic revalidation of prompts and evaluation controls.

Additionally, because the LLMs are used off the shelf, the study does not test whether lightweight domain adaptation would further improve faithfulness or audience alignment. While tuning may be beneficial for enforcing structure and reducing stylistic variability, it introduces new threats to validity (training data selection, leakage, and reduced cross-provider comparability). A rigorous treatment would require a human-validated reference set and an evaluation that prioritises factual traceability over lexical similarity.

Finally, the hallucination metric used for Experiment 2 is a whitelist-based proxy that measures out-of-scope feature invention relative to the provided decision record. While useful for enforcing and quantifying factual scope, it does not fully capture numerical correctness or whether the narrative's causal language is justified. Low hallucination rates should therefore be read as evidence of constraint adherence, not as a guarantee of full factual validity. As the whitelist-based hallucination proxy is lexicon-driven, it can misclassify some statements through imperfect matching, which weakens the reliability of the metric as an absolute measure. For this reason, the reported hallucination values are treated as indicative, with conclusions drawn primarily from consistent relative differences across controlled conditions and supported by complementary metrics and the human-centred findings.

5.3 Future Works

Future extensions include multi-asset, multi-market validation with intraday horizons and stress testing across regime shifts. The technical stack can add risk-aware variants and

constrained policies, causal and counterfactual XAI, and portfolio-level stability diagnostics. On the synthesis side, constrained decoding, retrieval grounding, and provenance tagging can reduce hallucination and strengthen linkage to XAI artefacts. Human-in-the-loop studies with experts, auditors, and regulators, together with longitudinal designs, would enhance ecological validity. Finally, aligning the pipeline with operational controls, audit trails, and AIA compliance would enable deployment in production settings.

5.4 Final Remarks

This work demonstrates a unified approach that combines technical faithfulness with human-centred usability in a financial trading case study. The layered XAI framework anchors explanations to RL behaviour via AIM/AUM, RIS, and surrogate fidelity, while the grounded LLM layer turns these diagnostics into concise natural-language narratives. Taken together, these complementary layers offer a practical template for transparent, auditable XRL in finance and a foundation for extending interpretable decision systems to additional assets, regimes, and user groups.

References

Agarwal, S., Iqbal, O., Buridi, S. A., Manjusha, M., and Das, A. Reinforcement Explanation Learning, November 2021. URL http://arxiv.org/abs/2111.13406. arXiv:2111.13406 [cs].

Aldridge, I. *High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems*. Wiley, 2013.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., and Herrera, F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99:101805, November 2023. ISSN 1566-2535. doi: 10.1016/j.inffus.2023.101805.

Allen, F. and Gale, D. *Comparing Financial Systems*. MIT Press, 2000. ISBN 0262011778.

Araci, D. Finbert: Financial sentiment analysis with pre-trained language models, 2019. URL https://arxiv.org/abs/1908.10063.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, December 2019. URL http://arxiv.org/abs/1910.10045.

Arsenault, P.-D., Wang, S., and Patenande, J.-M. A Survey of Explainable Artificial Intelligence (XAI) in Financial Time Series Forecasting, July 2024. URL http://arxiv.org/abs/2407.15909. arXiv:2407.15909 [cs].

Attanasio, G., Cagliero, L., and Baralis, E. Leveraging the explainability of associative classifiers to support quantitative stock trading. In *Proceedings of the International Workshop on Data Science for Macro-Modeling (DSMM)*, pages 1–6, 2020. doi: 10.1145/3401832.3402679.

Babaei, G., Giudici, P., and Raffinetti, E. Explainable artificial intelligence for crypto asset allocation. *Financial Research Letters*, 47:102941, 2022. doi: 10.1016/j.frl.2022.102941.

Bai, Y., Gao, Y., Wan, R., Zhang, S., and Song, R. A review of reinforcement learning in financial applications, 2024. URL https://arxiv.org/abs/2411.12746.

Bandi, H., Joshi, S., Bhagat, S., and Ambawade, D. Integrated technical and sentiment analysis tool for market index movement prediction, comprehensible using xai. In *2021 International Conference on Communication, Information and Computing Technology (ICCICT)*, pages 1–8. IEEE, 2021. doi: 10.1109/ICCICT50803.2021.9510124.

Bello, M., Bello, R., García, M.-M., Nowé, A., Sevillano-García, I., and Herrera, F. A three-level framework for llm-enhanced explainable ai: From technical explanations to natural language. *Information Systems Frontiers*, Dec 2025. ISSN 1572-9419. doi: 10.1007/s10796-025-10668-1.

Benhamou, É., Ohana, J.-J., Saltiel, D., Guez, B., and Ohana, S. Explainable AI (XAI) models applied to planning in financial markets. Research Paper 3862437, Université Paris-Dauphine, jun 2021. URL https://ssrn.com/abstract=3862437.

Bialkowski, J., Zastawniak, T., and Gissler, T. Do algorithmic traders improve liquidity? *International Review of Financial Analysis*, 17(5):890–899, 2008.

Bona, F. B. D., Dominici, G., Miller, T., Langheinrich, M., and Gjoreski, M. Evaluating explanations through llms: Beyond traditional user studies, 2024. URL https://arxiv.org/abs/2410.17781.

Brooke, J. SUS: A 'quick' and 'dirty' usability scale. In Jordan, P. W., Thomas, B., Weerdmeester, B. A., and McClelland, I. L., editors, *Usability Evaluation in Industry*, chapter 21, pages 189–194. Taylor and Francis, June 1996. ISBN 9780748404605.

Bussmann, N., Giudici, P., Marinelli, D., and Papenbrock, J. Explainable ai in fintech risk management. *Frontiers in Artificial Intelligence*, Volume 3 - 2020, 2020. ISSN 2624-8212. doi: 10.3389/frai.2020.00026.

Carta, S., Podda, A. S., Reforgiato Recupero, D., and Stanciu, M. M. Explainable ai for financial forecasting. In Nicosia, G., Ojha, V., La Malfa, E., La Malfa, G., Jansen, G., Pardalos, P. M., Giuffrida, G., and Umeton, R., editors, *Machine Learning, Optimization, and Data Science*, pages 51–69, Cham, 2022. Springer International Publishing. ISBN 978-3-030-95470-3.

Cartea, Jaimungal, S., and Penalva, J. *Algorithmic and High-Frequency Trading*. Cambridge University Press, 2015.

Choudhary, R., Gupta, R., and Shirani, H. Risk-adjusted deep reinforcement learning for portfolio management. *Quantitative Finance*, 25(1):45–62, 2025.

Cong, L. W., Tang, K., Wang, J., and Zhang, Y. AlphaPortfolio: Direct construction through deep reinforcement learning and interpretable AI. Working Paper 3554486, SSRN, aug 2021. URL https://ssrn.com/abstract=3554486.

Demajo, L. M., Vella, V., and Dingli, A. Explainable ai for interpretable credit scoring. In *Computer Science &; Information Technology (CS & IT)*, ACITY 2020. AIRCC Publishing Corporation, November 2020. doi: 10.5121/csit.2020.101516. URL http://dx.doi.org/10.5121/csit.2020.101516.

Deng, Y., Bao, F., Kong, Y., Ren, Z., and Dai, Q. Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3):653–664, 2017.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. Explanations based on the missing: Towards contrastive explanations with pertinent negatives, 2018. URL https://arxiv.org/abs/1802.07623.

Dikmen, M. and Burns, C. The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, 162:102792, 2022. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2022.102792.

Dispoto, G., Bonetti, P., and Restelli, M. "so, tell me about your policy...": Distillation of interpretable policies from deep reinforcement learning agents, 2025. URL https://arxiv.org/abs/2507.07848.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Easley, D., de Prado, M. L., and O'Hara, M. The microstructure of the 'flash crash': Flow toxicity, liquidity crashes, and the probability of informed trading. *Journal of Portfolio Management*, 37(2):118–128, 2011.

Fama, E. F. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970. ISSN 00221082, 15406261.

Ferreira, F. G. D. C., Gandomi, A. H., and Cardoso, R. T. N. Artificial intelligence applied to stock market trading: A review. *IEEE Access*, 9:30898–30917, 2021. doi: 10.1109/ACCESS.2021.3058133.

Fischer, T. and Krauss, C. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2):654–669, 2018.

Gao, Y., Xu, W., Cao, Z., and Yan, J. A framework for hierarchical deep reinforcement learning for portfolio management. *Expert Systems with Applications*, 164:113918, 2021.

García-Magariño, I. and Bravo-Agapito, J. DI2XAI: Dashboard for intelligent investment with practice over historic data using explainable artificial intelligence. SSRN, 2024. URL https://ssrn.com/abstract=5667638.

Giorgi, F., Silvestri, M., Campagnano, C., Silvestri, F., and Tolomei, G. Enhancing xai narratives through multi-narrative refinement and knowledge distillation. *ArXiv*, abs/2510.03134, 2025.

Gomez, O., Holter, S., Yuan, J., and Bertini, E. ViCE: Visual Counterfactual Explanations for Machine Learning Models, March 2020. URL http://arxiv.org/abs/2003.02428. arXiv:2003.02428.

Goodman, B. and Flaxman, S. European union regulations on algorithmic decision making and a "right to explanation". *AI Magazine*, 38(3):50–57, September 2017. ISSN 2371-9621. doi: 10.1609/aimag.v38i3.2741.

Guan, M. and Liu, X.-Y. Explainable deep reinforcement learning for portfolio management: An empirical approach, 2021. URL https://arxiv.org/abs/2111.03995.

Harris, L. *Trading and Exchanges: Market Microstructure for Practitioners*. Oxford University Press, 2003.

Hasbrouck, J. *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*. Oxford University Press, 2007.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., and Hussain, A. Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16, 08 2023. doi: 10.1007/s12559-023-10179-8.

Heuillet, A., Couthouis, F., and Díaz-Rodríguez, N. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214:106685, February 2021. ISSN 0950-7051. doi: 10.1016/j.knosys.2020.106685.

Hirchoua, O., Cottrell, M., and Gallinari, P. Deep reinforcement learning with risk-adjusted rewards for financial portfolio management. *Expert Systems with Applications*, 176:114784, 2021.

Hoffman, R., Klein, G., Mueller, S., and et al. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

Huotari, T., Savolainen, J., and Collan, M. Deep reinforcement learning agent for s&p 500 stock selection. *Axioms*, 9(4), 2020. ISSN 2075-1680. doi: 10.3390/axioms9040130.

IOSCO. Regulatory issues raised by the impact of technological changes on market integrity and efficiency. Technical report, International Organization of Securities Commissions, 2011.

Islam, S. R., Eberle, W., Ghafoor, S. K., and Ahmed, M. Explainable artificial intelligence approaches: A survey, 2021. URL https://arxiv.org/abs/2101.09429.

Izzo, C. On Explainable Deep Learning for Macroeconomic Forecasting and Finance - UCL Discovery, 2022. URL https://discovery.ucl.ac.uk/id/eprint/10162300/.

Jegadeesh, N. and Titman, S. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91, 1993. ISSN 00221082, 15406261.

Jeong, G. and Kim, H. Improving financial trading decisions using deep q-learning: Predicting the number of shares, action strategies, and transfer learning. *Expert Systems with Applications*, 117:125–138, 2019.

Jiang, Z., Xu, D., and Liang, J. A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem, July 2017. URL http://arxiv.org/abs/1706.10059. arXiv:1706.10059.

Jin, W., Fan, J., Gromala, D., Pasquier, P., and Hamarneh, G. Invisible users: Uncovering end-users' requirements for explainable ai via explanation forms and goals, 2023. URL https://arxiv.org/abs/2302.06609.

Kang, H. and Liu, X.-Y. Deficiency of large language models in finance: An empirical examination of hallucination. Papers 2311.15548, arXiv.org, Nov 2023. URL https://ideas.repec.org/p/arx/papers/2311.15548.html.

Kindleberger, C. P. and Aliber, R. Z. *Manias, panics, and crashes: a history of financial crises*. Palgrave Macmillan, 6th ed. edition, 2011. ISBN 9780230575974.

Kirilenko, A., Kyle, A. S., Samadi, M., and Tuzun, T. The flash crash: High-frequency trading in an electronic market. *Journal*

of Finance, 72(3):967–998, 2017.

Kumar, S., Vishal, M., and Ravi, V. Explainable Reinforcement Learning on Financial Stock Trading using SHAP, August 2022. URL https://arxiv.org/abs/2208.08790v1.

Lavie, A. and Agarwal, A. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, page 228–231, USA, 2007. Association for Computational Linguistics.

Leem, J.-H. and Kim, H. Action-specialized expert ensemble deep reinforcement learning for stock trading. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.

Li, P., Siddique, U., and Cao, Y. From Explainability to Interpretability: Interpretable Policies in Reinforcement Learning Via Model Explanation, January 2025. URL http://arxiv.org/abs/2501.09858. arXiv:2501.09858 [cs].

Liang, Y., Liu, Y., Wang, N., Yang, H., Zhang, B., and Wang, C. D. Fingpt: Enhancing sentiment-based stock movement prediction with dissemination-aware and context-enriched llms, 2025. URL https://arxiv.org/abs/2412.10823.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Lipton, Z. C. The mythos of model interpretability, 2017. URL https://arxiv.org/abs/1606.03490.

Liu, X.-Y., Xia, Z., Rui, J., Gao, J., Yang, H., Zhu, M., Wang, C. D., Wang, Z., and Guo, J. Finrl-meta: Market environments and benchmarks for data-driven financial reinforcement learning, 2022a. URL https://arxiv.org/abs/2211.03107.

Liu, X.-Y., Yang, H., Gao, J., and Wang, C. D. Finrl: deep reinforcement learning framework to automate trading in quantitative finance. In *Proceedings of the Second ACM International Conference on AI in Finance*, ICAIF '21, New York, NY, USA, 2022b. Association for Computing Machinery. ISBN 9781450391481. doi: 10.1145/3490354.3494366. URL https://doi.org/10.1145/3490354.3494366.

Lo, A. W. and MacKinlay, A. C. *A Non-Random Walk Down Wall Street*. Princeton University Press, 1999. ISBN 9780691092560. URL http://www.jstor.org/stable/j.ctt7tccx.

Lopez-Lira, A., Tang, Y., and Zhu, M. The memorization problem: Can we trust llms' economic forecasts?, 2025. URL https://arxiv.org/abs/2504.14765.

Lu, Y., Ouyang, S., and Zhou, K. Structured knowledge grounding for question answering, 2023. URL https://arxiv.org/abs/2209.08284.

Lucarelli, G. and Borrotti, M. A deep reinforcement learning approach for automated cryptocurrency portfolio management. *Expert Systems with Applications*, 173:114603, 2020.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Ma, S., Chen, Q., Wang, X., Zheng, C., Peng, Z., Yin, M., and Ma, X. Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making, March 2024. URL https://arxiv.org/abs/2403.16812v1.

Maree, C. and Omlin, C. W. Can interpretable reinforcement learning manage prosperity your way? *AI*, 3(2):526–537, June 2022. ISSN 2673-2688. doi: 10.3390/ai3020030.

Mern, J., Krishnan, S., Yildiz, A., Hatch, K., and Kochenderfer, M. J. Interpretable local tree surrogate policies, 2021. URL https://arxiv.org/abs/2109.08180.

Mersha, M., Lam, K., Wood, J., AlShami, A. K., and Kalita, J. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing*, 599:128111, September 2024. ISSN 0925-2312. doi: 10.1016/j.

neucom.2024.128111.

Mohammadshafie, S., Sadeghi, Z., and Wong, W.-K. Strategies of deep reinforcement learning agents in stock trading: A comparative analysis. *Journal of Behavioral and Experimental Finance*, 33:100746, 2024.

Mohseni, S., Zarei, N., and Ragan, E. D. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Trans. Interact. Intell. Syst.*, 11(3–4), September 2021. ISSN 2160-6455. doi: 10.1145/3387166.

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

Moody, J., Wu, L., Liao, Y., and Saffell, M. Performance functions and reinforcement learning for trading systems and portfolios. *Journal of Forecasting*, 17(5-6):441–470, 1998.

Murphy, J. J. *Technical Analysis of the Financial Markets*. New York Institute of Finance, 1999.

Müller, R., Schreyer, M., Sattarov, T., and Borth, D. Reshape: Explaining accounting anomalies in financial statement audits by enhancing shapley additive explanations, 2022. URL https://arxiv.org/abs/2209.09157.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, July 2023. ISSN 1557-7341. doi: 10.1145/3583558.

Naveed, S., Stevens, G., and Kern, D.-R. Explainable robo-advisors: Empirical investigations to specify and evaluate a user-centric taxonomy of explanations in the financial domain. In *IntRS@RecSys*, 2022. URL https://api.semanticscholar.org/CorpusID:252782003.

Neal, L. *The Rise of Financial Capitalism: International Capital Markets in the Age of Reason*. Studies in Macroeconomic History. Cambridge University Press, 1991.

Neuneier, R. Optimal asset allocation using adaptive dynamic programming. In *Advances in Neural Information Processing Systems*, volume 9, pages 952–958, 1996.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. URL https://aclanthology.org/P02-1040.

Paraschou, E., Arapakis, I., Yfantidou, S., Macaluso, S., and Vakali, A. Mind the xai gap: A human-centered llm framework for democratizing explainable ai, 2025. URL https://arxiv.org/abs/2506.12240.

Park, H. J., Kim, Y., and Kim, H. Y. Stock market forecasting using a multi-task approach integrating long short-term memory and the random forest framework. *Applied Soft Computing*, 114:108106, January 2022. ISSN 1568-4946. doi: 10.1016/j.asoc.2021.108106.

Pippas, N., Ludvig, E. A., and Turkay, C. The evolution of reinforcement learning in quantitative finance: A survey. *ACM Computing Surveys*, 57(11):1–51, June 2025. ISSN 1557-7341. doi: 10.1145/3733714.

Pring, M. J. *Technical Analysis Explained*. McGraw-Hill, 2002.

Quinn, B. Explaining AI in Finance: Past, Present, Prospects, June 2023. URL http://arxiv.org/abs/2306.02773. arXiv:2306.02773.

Rajani, N. F., McCann, B., Xiong, C., and Socher, R. Explain yourself! leveraging language models for commonsense reasoning, 2019. URL https://arxiv.org/abs/1906.02361.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL https://arxiv.org/abs/1908.10084.

Rong, Y., Leemann, T., Nguyen, T.-T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., and Kasneci, E. Towards

human-centered explainable ai: A survey of user studies for model explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2104–2122, April 2024. ISSN 1939-3539. doi: 10.1109/tpami.2023.3331846.

Samek, W., Wiegand, T., and Müller, K.-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, 2017. URL https://arxiv.org/abs/1708.08296.

Saranya, A. and Subhashini, R. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal*, 7:100230, 2023. ISSN 2772-6622. doi: https://doi.org/10.1016/j.dajour.2023.100230.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Singh, C., Inala, J. P., Galley, M., Caruana, R., and Gao, J. Rethinking interpretability in the era of large language models, 2024. URL https://arxiv.org/abs/2402.01761.

Srinivasan, R., Chander, A., and Pezeshkpour, P. Generating user-friendly explanations for loan denials using gans, 2019. URL https://arxiv.org/abs/1906.10244.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328, 2017.

Teymurzade, S. and Ašlepaczuk, R. Predicting djia, nasdaq and nyse index prices using arima and var models. Working Papers 2023-27, Faculty of Economic Sciences, University of Warsaw, 2023. URL https://ideas.repec.org/p/war/wpaper/2023-27.html.

The European Parliament. Regulation (EU) 2024/1689 of the european parliament and of the council, 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

Théate, T. and Ernst, D. An application of deep reinforcement learning to algorithmic trading. *Expert Systems with Applications*, 173:114632, July 2021. ISSN 0957-4174. doi: 10.1016/j.eswa.2021.114632.

Tsantekidis, A., Passalis, N., and Tefas, A. Diversity-driven knowledge distillation for ensemble reinforcement learning in financial trading. *Neurocomputing*, 423:1–12, 2021.

Tsay, R. S. *Analysis of Financial Time Series*. Wiley, 2010.

U.S. Securities and Exchange Commission. Technology and trading: Sec report. Technical report, U.S. SEC, 2013.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

Vo, N. and Ślepaczuk, R. Applying hybrid arima-sgarch in algorithmic investment strategies on s&p500 index. *Entropy (Basel)*, 24(2):158, jan 2022. ISSN 1099-4300. doi: 10.3390/e24020158.

Weber, P., Carl, K. V., and Hinz, O. Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature. *Management Review Quarterly*, 74(2):867–907, June 2024. ISSN 2198-1639. doi: 10.1007/s11301-023-00320-0.

Weng, C., Liu, Z., and Yan, Y. An attention-aware deep reinforcement learning framework for portfolio management. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):232–246, 2020.

Wu, C.-Y., Wei, C.-C., and Lee, C.-Y. Adaptive stock trading strategies with deep reinforcement learning methods. *Expert Systems with Applications*, 158:113573, 2020.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., and Mann, G. Bloomberggpt: A large language model for finance, 2023. URL https://arxiv.org/abs/2303.17564.

Xiong, Y., Hu, Z., Huang, Y., Wu, R., Guan, K., Fang, X., Jiang, J., Zhou, T., Hu, Y., Liu, H., Lyu, T., and Fan, C. Xrl-bench: A benchmark for evaluating and comparing explainable reinforcement learning techniques. In *Proceedings of the 30th*

ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24, page 6073–6082, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671595. URL https://doi.org/10.1145/3637528.3671595.

Yang, H., Liu, X.-Y., Zhong, S., and Walid, A. Deep Reinforcement Learning for Automated Stock Trading: An Ensemble Strategy. *SSRN Electronic Journal*, 2020. ISSN 1556-5068. doi: 10.2139/ssrn.3690996.

Yang, H., Liu, X.-Y., and Wang, C. D. Fingpt: Open-source financial large language models. *SSRN Electronic Journal*, June 2023. doi: 10.2139/ssrn.4489826. FinLLM at IJCAI 2023.

Yeo, W. J., van der Heever, W., Mao, R., Cambria, E., Satapathy, R., and Mengaldo, G. A Comprehensive Review on Financial Explainable AI, September 2023. URL http://arxiv.org/abs/2309.11960. arXiv:2309.11960 [cs, q-fin].

Yuan, Y., Wen, W., and Yang, J. Using Data Augmentation Based Reinforcement Learning for Daily Stock Trading. *Electronics*, 9(9):1384, September 2020. ISSN 2079-9292. doi: 10.3390/electronics9091384. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

Zhang, B., Yang, H., and Liu, X.-Y. Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models, June 2023. URL https://arxiv.org/abs/2306.12659v1.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with bert, 2020a. URL https://arxiv.org/abs/1904.09675.

Zhang, W., Liang, X., and Li, Y. Adaptive financial trading with deep reinforcement learning. *Applied Intelligence*, 50:188–201, 2020b.

Zou, J., Lou, J., Wang, B., and Liu, S. A novel deep reinforcement learning based automated stock trading system using cascaded lstm networks, 2023a. URL https://arxiv.org/abs/2212.02721.

Zou, J., Zhao, Q., Jiao, Y., Cao, H., Liu, Y., Yan, Q., Abbasnejad, E., Liu, L., and Shi, J. Q. Stock market prediction via deep learning techniques: A survey, 2023b. URL https://arxiv.org/abs/2212.12717.

Zytek, A., Pidò, S., and Veeramachaneni, K. LLMs for XAI: Future Directions for Explaining Explanations, May 2024. URL http://arxiv.org/abs/2405.06064. arXiv:2405.06064 [cs].

Çetin, E., Barrado, C., and Pastor, E. Explainability of Deep Reinforcement Learning Method with Drones. In *2023 IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC)*, pages 1–9, October 2023. doi: 10.1109/DASC58513.2023.10311156. URL https://ieeexplore.ieee.org/abstract/document/10311156. ISSN: 2155-7209.

Černevičienė, J. and Kabašinskas, A. Explainable artificial intelligence (XAI) in finance: a systematic literature review. *Artificial Intelligence Review*, 57(8):216, July 2024. ISSN 1573-7462. doi: 10.1007/s10462-024-10854-8.

A Reproducibility and Media Artefacts

A.1 Software Stack and Experimental Setup

Experiments 1 and 2 were implemented in Python using modular, reproducible pipelines executed on Google Colab Pro (Linux x86_64) with GPU acceleration (NVIDIA Tesla T4 and A100, 16–40 GB VRAM). Random seeds were fixed at 42 to ensure deterministic reproducibility across all runs.

Experiment 1

Implemented in Python 3.11 with *stable-baselines3* (2.2.1), *FinRL* (0.3.6), *PyTorch* (2.4.1)[1], *Captum* (0.7.0), and *SHAP* (0.46.0). Supporting libraries included *pandas*[2], *numpy*[3], *matplotlib*, and *scikit-learn*[4]. The workflow comprised dataset preparation (Dow 30 stocks), model training (PPO, DDPG, TD3, A2C), trajectory extraction, explainability computation (SHAP, Integrated Gradients, GradientSHAP, Saliency Maps), and evaluation (AIM, AUM, RIS, fidelity, Sharpe, Calmar).

Experiment 2

Conducted in Python 3.12, using *pandas*, *numpy*, and *matplotlib* for data processing; *openai* and *anthropic* for LLM access; and *nltk*[5], *textstat*[6], and *rouge_score* for linguistic evaluation metrics. *tqdm*[7] was used for batch monitoring. Each model run used deterministic decoding (*temperature = 0.0*, *max_tokens* $\leq$ *400*) and generated outputs serialised into `.jsonl` files within structured directories.

[1] Information about PyTorch: `https://pytorch.org/`
[2] Pandas documentation: `https://pandas.pydata.org/`
[3] `https://numpy.org/`
[4] `https://scikit-learn.org/stable/`
[5] `https://www.nltk.org/`
[6] `https://textstat.org/`
[7] GitHub implementation: `https://github.com/tqdm/tqdm`

Unlike the RL agents in Experiment 1, which are trained end to end, the LLMs in Experiment 2 are used off the shelf via API, with all customisation occurring at the prompt level rather than through parameter fine-tuning.

A.2 Repository Contents and Media Files

All intermediate artefacts were stored on and loaded from Google Drive for ease of access. All output data files, figures, and source code used in Experiments 1 and 2 are archived in a project repository on GitHub[8].

Table A.1 summarises the main files and directories relevant to reproducing the experiments and analyses.

Table A.1: Overview of repository contents and media artefacts.

File / folder	Description
`README.md`	Top-level documentation on the structure, software dependencies, and instructions for Experiments 1 and 2.
`Source Code/`	Top-level source directory containing the implementation and outputs of both experiments: **Experiment 1/** **Experiment 1 - Explainability Framework.ipynb** – main notebook for data preparation, training of the RL agents, and generation of explainability artefacts. **Experiment_1_Results/** – processed data, trained model checkpoints, attribution logs, stability metrics, policy-surrogate outputs, and reward-attribution summaries. **Experiment 2/** **Experiment 2 - Language-Based Interpretability.ipynb** – main notebook for prompt construction, LLM runs, and automated evaluation of explanations. **Experiment_2/** – generated explanations, prompt templates, evaluation outputs, and synthesis reports.
`Survey_Responses/`	Anonymised pre-study, in-study, and post-study questionnaire responses from the human-centred evaluation.

[8]The repository is currently hosted at: `https://github.com/NathanPortelli/Enhancing-Transparency-and-Interpretability-in-AI-Driven-Algorithmic-Trading`.

B Literature Review Studies

B.1 Summary of Reinforcement Learning Studies in Financial Markets

Table B.1: Summary of key RL applications in financial markets, including methods, datasets, evaluation metrics, performance, and major conclusions.

Reference	RL Method	Data	Key Metrics	Performance	Conclusions
Neuneier (1996)	Adaptive DP (Q-learning)	S&P 500	Return, Drawdown	Outperformed buy-hold over 1970-94	First demonstration of RL improving asset allocation vs. benchmarks.
Moody et al. (1998)	Recurrent Policy Gradient	S&P 500 + T-bills (monthly)	Sharpe ratio, Wealth	Sharpe $\approx$ 1.5 vs 0.7 benchmark; avoided 1987 crash	RL-trained "trading systems" beat buy-hold; introduced differential Sharpe ratio reward.
Deng et al. (2017)	A3C (with LSTM)	Chinese index futures (tick data)	Annualised Return, Sharpe	Sharpe 1.02 vs 0.78 baseline; +32% return	Deep recurrent RL captured temporal patterns; outperformed linear models.

(continued…)

Reference	RL Method	Data	Key Metrics	Performance	Conclusions
Jiang et al. (2017)	Policy Gradient (portfolio weights)	Crypto portfolio (daily)	Cumulative return, MDD	4-asset portfolio return 197% vs 170% market	Proposed DeepPortfolio with softmax policy - higher returns with risk penalty in reward.
Jeong and Kim (2019)	DQN (with action sizing)	KOSPI 200 index (daily)	Profit ($), Accuracy	+9.7% profit vs baseline; 60% win rate	DQN predicted optimal trade sizes; transfer learning improved generalisation.
Lucarelli and Borrotti (2020)	Dueling Double DQN	Crypto (4 coins, hourly)	Sharpe, ROI	Best variant Sharpe 0.20; ROI +7.5% in high-vol regime	Multi-agent DQN for portfolios; robust profits in volatile crypto markets.
Wu et al. (2020)	DDPG (with GRU encoder)	US stocks (15, daily 8yr)	Sortino, Cum. Return	Sortino 1.45 (vs 0.5 baseline); +200% return on best stock	Two adaptive strategies beat Turtle Trading; GRU-based DDPG adapts to non-stationary phases.
Yang et al. (2020)	Ensemble (PPO, DDPG, SAC)	US stocks (30, daily 7yr)	Annual Return, Sharpe	Return 13% vs 4% S&P; Sharpe 1.30	Ensemble of DRL agents outperformed single agents; improved return stability.
Huotari et al. (2020)	DQN variant	S&P 500 stock picking (daily)	Total Return, Hit ratio	Portfolio +328% vs 257% index	DQN selected best subset of stocks; higher total return than S&P.

(continued…)

Reference	RL Method	Data	Key Metrics	Performance	Conclusions
Weng et al. (2020)	SAC + Attention	Crypto (20 assets, 30-min)	Cumulative return	22× initial capital in 2mo	Attention-enhanced SAC exploited cross-asset signals, yielding high growth.
Leem and Kim (2020)	Multi-DQN Ensemble	KOSPI stock (daily)	Return, MDD	+29.4% return vs 3.6% index; MDD -5%	Specialised DQNs per action improved performance, limited drawdowns.
Théate and Ernst (2021)	"Trading DQN" (TDQN)	Global stocks (daily)	Sharpe, Calmar	Sharpe up to 1.1 (vs 0.3 benchmark); Calmar 3.5	Custom DQN maximising Sharpe significantly improved returns; validated via backtests.
Tsantekidis et al. (2021)	DDPG Ensemble + Distillation	Forex (EURUSD, hourly)	Profit, Drawdown	+35% profit vs best single agent; DD decreased	Diversity-driven ensemble distilled into single agent retained ensemble performance.
Hirchoua et al. (2021)	Actor-Critic (with curiosity)	US stocks (daily)	Sharpe, Sortino	Sharpe 1.12 vs 0.8 baseline; Sortino +15%	Risk-driven curiosity reward improved risk-adjusted returns.
Gao et al. (2021)	Hierarchical DQN	Chinese stocks (daily)	Annual Return	44% vs 18% (flat DQN)	Sector + stock hierarchy yielded higher returns; proved hierarchy benefits.
Mohammadshafie et al. (2024)	A2C, PPO, SAC, DDPG, TD3 (comparison)	US stocks (daily)	Cum. Reward, Trades	A2C best reward (100); PPO/SAC most trades	A2C earned highest wealth; PPO/SAC hyperactive; DDPG/TD3 held assets longer.

(continued…)

Reference	RL Method	Data	Key Metrics	Performance	Conclusions
Choudhary et al. (2025)	Multi-Objective (3× DRL)	Global indices (daily)	Sharpe, Calmar	Sharpe 1.4 (+8% vs best single)	CNN-fused policy combining return, Sharpe, and drawdown agents outperformed single-goal.

B.2 Summary of Selected XAI Approaches in Financial Markets

Table B.2: Summary of Selected XAI Approaches in Financial Markets

Reference	Methodology	Data/Domain	Metrics	Findings and Conclusion
Attanasio et al. (2020)	Associative classifier (rule-based) for stock trading signals; inherent interpretability	Historical stock market data (U.S. equities)	Trading accuracy, profit factor	Generated human-readable IF–THEN trading rules. Showed that a rule-based model can support quantitative trading decisions without significant loss in performance, providing transparency in strategy rationale.

(continued…)

Reference	Methodology	Data/Domain	Metrics	Findings and Conclusion
Bandi et al. (2021)	Hybrid model combining technical indicators and sentiment analysis; explanations via feature importance (XAI toolkit)	Market index movement prediction (India, index levels with news sentiment)	Prediction accuracy (F1, accuracy)	Integration of news sentiment with technical features improved prediction accuracy over technicals-only models. XAI analysis revealed that sentiment features (e.g. news polarity) were among top predictors, instilling trust that the model's gains are driven by intuitive market information.
Benhamou et al. (2021)	Gradient boosting trees for regime change (crash) prediction; post-hoc SHAP explanations for feature impact	S&P 500 futures with 150 macro/technical features (1970s–2020)	Classification accuracy, timing of crash prediction (out-of-sample)	GBDT outperformed other ML methods in predicting equity crashes. SHAP identified a mix of pro-cyclical and counter-cyclical features driving crash risk. Provided local explanations for the 2020 crash (e.g. tech sector metrics acted as contrarian signals), giving practitioners a clear rationale for forecasts of market regime shifts.

(continued…)

Reference	Methodology	Data/Domain	Metrics	Findings and Conclusion
Kumar et al. (2022)	Deep Reinforcement Learning (DQN) for stock trading; SHAP values explain the agent's actions	Stock price data for multiple companies (Yahoo Finance, 2013–2018)	Cumulative return, Sharpe ratio of the RL agent	The RL trading agent achieved profitable performance (beating a buy-and-hold baseline). SHAP explanations for the DQN's policy highlighted intuitive drivers: moving average trends and volatility indices strongly influenced buy/sell decisions. This demonstrated that the agent's strategy aligns with traditional trading signals, aiding user acceptance.

(continued…)

Reference	Methodology	Data/Domain	Metrics	Findings and Conclusion
Cong et al. (2021)	AlphaPortfolio: Deep RL-based portfolio allocation with policy distillation (Lasso surrogate) and polynomial feature analysis	U.S. stock market portfolio (multiple assets) with monthly rebalancing	Annualised return, Sharpe ratio, alpha	Achieved high out-of-sample performance (Sharpe > 2.0, $\sim$13% annual alpha). The distilled linear model retained $\sim$5 key features (e.g. interest rates, volatility) explaining most decisions. Polynomial sensitivity analysis identified which features drove returns, making the black-box RL strategy transparent. Conclusion: interpretable policy rules can be extracted without sacrificing returns, enabling traders to understand and trust the AI strategy.

(continued…)

Reference	Methodology	Data/Domain	Metrics	Findings and Conclusion
Babaei et al. (2022)	Explainable portfolio optimisation for crypto assets using XGBoost; TreeSHAP for global and local explanation	Cryptocurrency market data (top assets, 2016–2021)	Portfolio return, risk (volatility), Sharpe ratio	Proposed an AI asset allocation that slightly outperformed a naive benchmark. The novelty lies in explainability: SHAP values provided clear justification for allocations (e.g. high-momentum coins got larger weights). Regulators and investors could verify that allocations obeyed rational patterns (such as diversifying in high-volatility periods), demonstrating compliance and building trust in robo-advisory contexts.

C Experiment 1 Data, Feature Set, Configuration and Performance

This appendix documents the data, feature engineering, asset universe and training configuration used for the RL agents in Experiment 1, as well as the overall financial performance of the models. It complements the methodological description in Sections 3.1.2 and 3.1.3 by providing full tabular summaries of the processed features and environment parameters. The underlying price and volume series are retrieved via the *YahooDownloader* module in *FinRL* and processed through the *FeatureEngineer* component, which applies standard TA-Lib technical indicators to the Dow 30 universe over the 2009–2021 sample (Liu et al. (2022b)).

C.1 Market and Indicator Features

Table C.1 lists the core columns in the processed market dataset. In the long format used during data engineering, each record corresponds to a single asset–day pair with the following fields:

During environment construction, these long-format records are pivoted into a wide state representation. For each trading day, the agent observes a concatenation of per-asset indicator vectors and portfolio variables,

$$S_t = [p_{t,1}, \ldots, p_{t,n}, v_{t,1}, \ldots, v_{t,n}, b_t],$$

where $p_{t,i}$ contains the normalised indicators for asset i, $v_{t,i}$ is the number of shares held in asset i, and b_t is the remaining cash balance (see Section 3.1.2 and Section 3.1.3). In the explainability layers, these features appear as asset-specific columns such as `AAPL_rsi_30` or `DIS_dx_30`. Additional derived descriptors, including Bollinger band upper and lower bounds, are computed from the `close` series during the attribution analysis and give rise to features such as `JNJ_boll_ub` and `CRM_boll_lb` in the Layer 1 and Layer 4 panels.

Table C.1: Market-level and technical-indicator features used in Experiment 1.

Field	Description
`date`	Trading day in `YYYY-MM-DD` format. Serves as the temporal index for all assets.
`tic`	Stock ticker identifier for the corresponding Dow 30 constituent (see Table C.2).
`open`	Adjusted opening price (USD) for the asset on that trading day.
`high`	Adjusted intraday high price (USD).
`low`	Adjusted intraday low price (USD).
`close`	Adjusted closing price (USD) used as the base for indicator calculations and portfolio valuation.
`volume`	Number of shares traded during the day.
`day`	Integer day-of-week index (0–4) indicating Monday to Friday, used to capture simple calendar effects.
`macd`	Moving Average Convergence Divergence indicator computed from closing prices with the standard short, long, and signal windows, capturing trend and momentum.
`rsi_30`	30-day Relative Strength Index, measuring overbought and oversold conditions based on recent gains and losses.
`cci_30`	30-day Commodity Channel Index, capturing deviations of price from its moving average as a normalised oscillator.
`dx_30`	30-day Directional Movement Index component, summarising the strength of directional trends over the lookback horizon.
`turbulence`	Cross-sectional turbulence index computed over the Dow 30 universe, measuring how unusual the current return vector is relative to its recent historical distribution and acting as a market-wide risk indicator.

C.2 Dow 30 Ticker Universe

The asset universe follows the fixed Dow Jones Industrial Average (Dow 30) constituents used in the *FinRL* NeurIPS 2018 example, held constant over the 2009–2021 period for reproducibility (Liu et al. (2022b)). The list is shown in Table C.2.

This fixed composition provides a liquid, large-cap universe spanning multiple sectors of the United States economy. It also matches the default Dow 30 configuration in *FinRL*, simplifying comparison with prior deep RL trading studies and ensuring that differences in performance and explainability can be attributed to modelling choices rather than to changes in the underlying asset set.

Table C.2: Dow 30 ticker universe used in Experiment 1.

Ticker	Company	Sector (approximate)
AXP	American Express Co.	Financials / Consumer Finance
AMGN	Amgen Inc.	Health Care / Biotechnology
AAPL	Apple Inc.	Information Technology / Hardware
BA	The Boeing Company	Industrials / Aerospace
CAT	Caterpillar Inc.	Industrials / Machinery
CSCO	Cisco Systems Inc.	Information Technology / Networking
CVX	Chevron Corporation	Energy / Integrated Oil and Gas
GS	The Goldman Sachs Group Inc.	Financials / Investment Banking
HD	The Home Depot Inc.	Consumer Discretionary / Retail
HON	Honeywell International Inc.	Industrials / Conglomerate
IBM	International Business Machines Corp.	Information Technology / IT Services
INTC	Intel Corporation	Information Technology / Semiconductors
JNJ	Johnson & Johnson	Health Care / Pharmaceuticals
KO	The Coca-Cola Company	Consumer Staples / Beverages
JPM	JPMorgan Chase & Co.	Financials / Banking
MCD	McDonald's Corporation	Consumer Discretionary / Restaurants
MMM	3M Company	Industrials / Diversified
MRK	Merck & Co. Inc.	Health Care / Pharmaceuticals
MSFT	Microsoft Corporation	Information Technology / Software
NKE	NIKE Inc.	Consumer Discretionary / Apparel
PG	Procter & Gamble Co.	Consumer Staples / Household Products
TRV	The Travelers Companies Inc.	Financials / Insurance
UNH	UnitedHealth Group Inc.	Health Care / Managed Care
CRM	Salesforce Inc.	Information Technology / Cloud Software
VZ	Verizon Communications Inc.	Communication Services / Telecoms
V	Visa Inc.	Financials / Payments
WBA	Walgreens Boots Alliance Inc.	Consumer Staples / Pharmacy Retail
WMT	Walmart Inc.	Consumer Staples / Retail
DIS	The Walt Disney Company	Communication Services / Media
DOW	Dow Inc.	Materials / Chemicals

C.3 Environment Configuration (*StockTradingEnv-v2*)

For reproducibility, Table C.3 summarises the key environment parameters used for all agents in Experiment 1. These settings correspond to the *StockTradingEnv-v2* configuration in FinRL, with minor adjustments as described in Section 3.1.3.

Table C.3: Environment configuration (*StockTradingEnv-v2*) for RL agents.

Component	**Configuration**	**Notes**
General	$h_{\max} = 100$; initial capital = \$1,000,000; reward scaling = 1×10^{-3}; turbulence threshold = 250	FinRL-v2 baseline with adjusted reward scale.
Transaction Costs	Buy/sell fee = 0.1% per transaction	Slightly reduced cost to encourage trading activity.
Action Space	Continuous; dimension = number of stocks	Non-discrete trading actions.
Cash Penalty	0.1 (proportion of idle cash penalised)	Encourages capital deployment.
Risk Wrapper	30-day rolling volatility penalty = 0.05	Implements risk-adjusted reward shaping.

C.4 Model Training Configuration

Table C.4 lists the principal hyperparameters for each RL algorithm. All agents were trained for 600k environment steps using *stable-baselines3* as backend, under the common environment defined above.

Table C.4: Model training configuration for RL agents.

Algorithm	Key Parameters	Learning Rate	Steps	Training Duration
A2C	$n_{\text{steps}} = 5$, $ent\_coef = 0.005$, $vf\_coef = 0.5$, $gae\_\lambda = 1.0$	7×10^{-4}	On-policy (RMSProp)	600k steps
PPO	$n_{\text{steps}} = 2048$, $clip\_range = 0.2$, $ent\_coef = 0.01$, $n\_epochs = 10$, $gae\_\lambda = 0.95$	3×10^{-4}	Batch = 256	600k steps
DDPG	$\tau = 0.005$, $\gamma = 0.99$, buffer = 10^6, batch = 128	1×10^{-3}	OU noise $\sigma = 0.1$	600k steps
TD3	$\tau = 0.005$, $\gamma = 0.99$, delay = 2, noise clip = 0.5, buffer = 10^6	1×10^{-3}	Gaussian noise $\sigma = 0.1$	600k steps

C.5 Evaluation of Financial Performance

The evaluation component of Experiment 1 assesses both the financial effectiveness and the interpretability quality of each RL trading agent, ensuring that improved transparency does not come at the expense of trading performance. Financial performance follows standard quantitative-finance conventions, using annualised return, cumulative return, annualised volatility, Sharpe, Sortino, Calmar, and Omega ratios to capture profitability, risk-adjusted performance, drawdown sensitivity, and tail behaviour. All metrics were computed using functions within the FinRL PyFolio pipeline[1].

For reference, the Sharpe ratio summarises risk-adjusted performance as

$$SR = \frac{E[r_t]}{\sigma(r_t)},$$

assuming a zero risk-free rate, while the Calmar ratio relates annualised return to maximum drawdown,

$$\text{Calmar} = \frac{AR}{|\text{Max Drawdown}|},$$

with all other metrics following their standard PyFolio definitions. Each metric was evaluated for every trained agent at checkpoints, providing a consistent basis for assessing stability and convergence of financial performance across models.

[1]Base PyFolio can be found here: `https://github.com/quantopian/pyfolio`

C.5.1 FinRL Replication: Financial and Trading Performance

Four RL agents (PPO, DDPG, TD3, A2C) were trained on the DJIA dataset (2009–2021) under identical market, reward, and transaction-cost conditions. Table C.5 reports standard risk and return metrics on the held-out test window (July 2020 to June 2021), alongside a DJIA buy-and-hold benchmark, and Figure 4.1 shows the corresponding cumulative-return trajectories.

Table C.5: Financial performance on the test window (Jul 2020–Jun 2021).

Model	Ann. Ret. (%)	Cum. Ret. (%)	Vol. (%)	Sharpe	Calmar	Stab.	MDD (%)	Omega	Sortino	Skew	Kurt.	Tail	VaR (%)
DJIA (BH)	33.56	33.25	14.57	2.06	3.76	0.95	8.93	1.41	2.93	-0.37	4.08	1.08	-1.48
A2C	46.65	48.44	17.87	2.23	6.14	0.94	7.60	1.42	3.53	-0.01	2.72	1.16	-1.85
PPO	38.66	40.10	16.12	2.11	1.71	0.60	22.67	1.41	3.56	0.40	3.36	1.31	-1.41
DDPG	32.86	34.07	16.55	1.80	4.12	0.91	7.98	1.33	2.82	0.08	2.86	1.07	-1.63
TD3	38.09	39.51	14.82	2.25	4.35	0.83	8.76	1.44	3.62	0.07	3.13	1.28	-1.34

Notes. Metrics follow standard PyFolio definitions; the risk-free rate is assumed 0% for Sharpe and Sortino; transaction cost is 10 bps per trade on buys and sells. "Stab." denotes the equity-curve stability coefficient (R^2 of a linear fit on log cumulative equity); "Tail" is the tail ratio; VaR is one-day 95% historical VaR reported as a negative percentage.

Across risk-normalised measures, three of the four RL agents (A2C, PPO, TD3) outperform the DJIA benchmark, with DDPG slightly below the benchmark on Sharpe but still ahead on cumulative return. TD3 achieves the highest Sharpe and Sortino scores with volatility close to the DJIA, while A2C combines strong risk-adjusted performance with the highest Calmar ratio, reflecting the most favourable return–drawdown balance. PPO delivers competitive returns but with a substantially deeper maximum drawdown. Overall, these outcomes are consistent with prior FinRL-based studies on DJIA daily trading (Liu et al. (2022b); Yang et al. (2020)), confirming that the replicated environment provides a valid baseline for the subsequent explainability analysis.

Annualised returns, volatility, Sharpe ratios, and drawdowns fall within the same order of magnitude as the reference study, with A2C closely matching the published risk–return profile and PPO and TD3 achieving similar Sharpe ratios ($\approx$2.1–2.3). Minor deviations are consistent with stochastic training variability and alternative random seeds, and confirm that the replicated FinRL environment provides a realistic and profitable baseline.

D Automated Linguistic Metrics for Experiment 2

This appendix reports the full set of lexical overlap metrics for Experiment 2, complementing the summary in Section 4.2.4. For each LLM provider, audience type (novice, expert), and prompting mode, METEOR, ROUGE-1, ROUGE-2, ROUGE-L, and corpus-level BLEU scores are provided, computed against the pseudo-references derived from the XAI corpus. Higher values indicate greater lexical and structural overlap with the factual substrate for METEOR and ROUGE; BLEU is reported as a descriptive corpus-level indicator.

Table D.1: Descriptive overlap metrics by model, audience, and prompting mode. Higher is better for METEOR/ROUGE; BLEU is corpus level.

Model (Provider)	**Group**	**Prompt mode**	**METEOR**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BLEU**
Claude-Sonnet-4 (Anthropic)	Expert	Baseline	0.108	0.152	0.035	0.118	2.10
		Counterfactual	0.123	0.165	0.047	0.129	3.39
		Historical	**0.227**	**0.219**	**0.082**	**0.182**	3.80
		Few-shot	0.121	0.148	0.022	0.119	0.87
		Synthesis (Few-shot+Hist.)	0.224	0.217	0.082	0.183	3.86
	Novice	Baseline	0.190	0.153	0.055	0.123	2.77

Continued on next page

Model (Provider)	Group	Prompt mode	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
		Counterfactual	0.187	0.154	0.059	0.120	3.27
		Historical	**0.227**	**0.219**	**0.083**	**0.183**	**3.99**
		Few-shot	0.198	0.142	0.055	0.119	1.70
		Synthesis (Few-shot+Hist.)	0.222	0.218	0.084	0.184	3.93
GPT-4o (OpenAI)	Expert	Baseline	0.192	0.173	0.045	0.132	2.33
		Counterfactual	0.155	0.132	0.024	0.102	1.13
		Historical	0.193	0.180	**0.059**	**0.150**	**2.64**
		Few-shot	0.176	0.152	0.039	0.120	1.33
		Synthesis (Few-shot+Hist.)	0.193	0.178	0.057	0.149	2.62
	Novice	Baseline	**0.258**	0.186	**0.082**	**0.158**	2.20
		Counterfactual	0.182	0.156	0.039	0.119	2.33
		Historical	0.196	0.176	0.058	0.146	**2.54**
		Few-shot	0.202	0.159	0.068	0.129	2.33
		Synthesis (Few-shot+Hist.)	0.191	0.177	0.056	0.147	2.50
GPT-5 Chat (OpenAI)	Expert	Baseline	0.142	0.175	0.044	0.147	2.75
		Counterfactual	0.130	0.144	0.024	0.115	1.40
		Historical	**0.163**	**0.205**	**0.061**	**0.167**	**2.84**
		Few-shot	0.156	0.156	0.030	0.128	1.22

Continued on next page

Model (Provider)	Group	Prompt mode	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
		Synthesis (Few-shot+Hist.)	0.165	0.205	0.062	0.170	2.88
	Novice	Baseline	**0.179**	0.195	0.056	0.136	2.84
		Counterfactual	0.165	0.156	0.033	0.108	1.64
		Historical	0.166	**0.205**	**0.063**	**0.168**	**2.80**
		Few-shot	0.161	0.137	0.032	0.109	1.56
		Synthesis (Few-shot+Hist.)	0.163	0.204	0.061	0.170	2.78

E LLM Synthesis Framework Content

This appendix documents the guardrails, anonymisation scheme, base terms, prompt templates and corpus schema used in the LLM-based synthesis framework.

E.1 Domain-Specific Base Terms for Factuality Diagnostics

The following table lists the domain-specific base terms used to build the factuality diagnostics and leakage checks for the LLM explanations.

Table E.1: Domain-specific base terms used in factuality diagnostics.

Category	**Representative Terms**
Core indicators	macd, rsi, cci, adx, dx, volatility, turbulence
Indicator components	index, channel, commodity, relative, strength, moving, average
Trading and RL terminology	bull, bear, sideways, reward, action, policy, regime, portfolio, risk, trend, signal, agent, trading, model, feature, return, profit, loss, explanation, stability, fidelity, reinforcement, learning, market, environment
Generic explanatory / domain-neutral terms	price, stock, indicator, decision, considered, shown, rising, falling, buy, sell, hold, position, market condition, movement, momentum, trendline, signal strength, feature importance, reward signal

Continued on next page

Category	Representative Terms (continued)
Neutral verbs / adjectives	indicating, suggested, observed, likely, possible, increase, decrease, rise, fall, change, fluctuation, upward, downward, period, quarter
Temporal and contextual tokens	day, month, year, session
Behavioural / signal descriptors	oversold, overbought, strong, weak, positive, negative, steady, fluctuating, decline, volatility index, tendency, direction
Strategy or decision-level terms	strategy, approach, decision making, holdings, confidence, pattern, signal line, crossover, trend reversal
FinRL-Meta (RL and DRL concepts)	agent, environment, gym style environment, markov decision process, state, action, reward, policy, policy gradient, deep q learning, dqn, ddpg, ppo, hyperparameter tuning, ensemble strategy, population based training, generational evolution, tournament based evolution, curriculum learning, simulation to reality gap
FinRL-Meta (Financial and trading terms)	algorithmic trading, backtesting, signal to noise ratio, snr, dataops, sentiment data, historical data, survivorship bias, information leakage, paper trading, ohlcv, technical indicators, market frictions, market crash, volatility index, vix, limit order book, lob, smart beta index, liquidation, trade execution

E.2 Faithfulness Guardrail and Anonymisation

E.2.1 Future-Information Guard

The following excerpt shows the faithfulness-oriented system-level guard used to constrain each LLM call to the quantitative context supplied by the explainability framework:

```
STRICT CONSTRAINTS:
• Use only the indicators, metrics, and values provided below.
• Do not reference external knowledge, future events, or market news.
• Do not infer information not contained in the supplied context.
• Mention only the features and drivers explicitly listed in
```

'Important Features' and 'Reward Drivers'.
- If some information is missing, still explain using any available indicators or drivers.
- Never predict future outcomes or financial results.
- Keep reasoning factual, short, and limited to the given inputs.

E.2.2 Anonymisation Scheme

Before prompt construction, instrumented outputs are anonymised to avoid leakage of identifiable assets and dates. The anonymisation step applies the following rules:

- Each Dow 30 ticker symbol is deterministically mapped to a neutral alias of the form `Asset_01`, `Asset_02`, ..., `Asset_30`.
- All four-digit calendar years in the range 1900–2099 are replaced with the token `[year]`.
- Quarter markers (`Q1`–`Q4`) are replaced with the token `[quarter]`.
- Full and abbreviated month names (`Jan`, `January`, ..., `Dec`, `December`) are replaced with the token `[month]`.
- The same transformation is applied to both free-text fields and feature lists before these are interpolated into the templates below.

E.3 Prompt Templates

This section lists the effective prompt templates used for each audience and mode. For each case, we show the structured context record presented as the user message, followed by the corresponding system message that controls style and faithfulness.

E.3.1 Novice-Investor Explanation Template

The following excerpt shows the context record used to instantiate the novice-level explanation prompt:

```
{
  ## Reinforcement Learning Explainability Context (Layers A–D)
  Model: {model} | Checkpoint: {checkpoint}
  Regime: {regime}
```

```
  Action: {action} | Reward: {reward}

  Important Features (Layer A): {allowed_features}
  Stability (Layer B - AIM/AUM/RIS):
    {aim}/{aum}/{ris}
  Policy Fidelity (Layer C):
    Fidelity={policy_fidelity:.3f}
    | Key Indicators: {policy_top_features}
    | Rules: {policy_rules}
  Reward Drivers (Layer D):
    {top_reward_features}
    | Exposed Features: {reward_exposed_features}
    | Mean Attribution: {reward_attr_mean}
    | Mean Corr. w/ Return: {reward_corr_mean}

  **Task**
  Explain in simple English:
  1. Why the agent likely took this action, referring to the listed
     indicators and reward drivers.
  2. How the market regime may have influenced its logic.
  3. Whether the reward was consistent with the indicators' signals.
  4. Stay within {target_words} words, producing a short,
     readable paragraph.
}
```

The corresponding system message, which defines style and faithfulness constraints for novice readers, is:

```
{
  "system": (
    "You are an AI assistant explaining the behaviour of "
    "reinforcement learning (RL) trading agents to a *novice "
    "investor* with no technical background.\n"
    "Explain decisions clearly, factually, and in plain language.\n"
    "Use only the indicators and drivers provided.\n"
    "Do not invent or assume causes.\n"
    "Follow these clarity guidelines:\n"
```

```
    "- One main idea per sentence, average length under 15 words.\n"
    "- Use simple connectors (because, when, if) to show reasoning.\n"
    "- Replace jargon with plain phrases (e.g., 'indicator showing "
    "strength of price movement').\n"
    "- Aim for 100-120 words but keep going if needed to finish "
    "your point.\n"
    "- Ensure Gunning Fog index $\approx$ 8-10 for novice clarity.\n"
    "- Mention at least one 'Important Feature'; include a reward "
    "driver if available.\n"
    "- If reasoning cannot be derived from inputs, write "
    "'Insufficient information.'"
  )
}
```

E.3.2 Expert-Analyst Explanation Template

For expert financial analysts, the user-level context is instantiated as:

```
{
  ## Analytical Context (Layers A-D)
  Model: {model} | Checkpoint: {checkpoint}
  Regime: {regime}
  Action: {action} | Reward: {reward}

  Feature Attribution (A): {allowed_features}
  Stability (B):
    AIM={aim}, AUM={aum}, RIS={ris}
  Policy Interpretability (C):
    Fidelity={policy_fidelity:.3f},
    TopFeatures={policy_top_features},
    Rules={policy_rules}
  Reward Attribution (D):
    Key Drivers={top_reward_features},
    Exposed Features: {reward_exposed_features},
    Mean Attribution: {reward_attr_mean},
    Mean Corr. w/ Return: {reward_corr_mean}
```

```
  **Task**
  Produce a concise analytical explanation covering:
  1. Which features and reward drivers explain the action.
  2. Whether the decision logic aligns with stable, interpretable
     reasoning (AIM/AUM/Fidelity).
  3. If the reward outcome reflects correct directional alignment.
  4. Limit output to {target_words} words.
}
```

The associated system message is:

```
{
  "system": (
    "You are an assistant writing structured analytical explanations "
    "for *financial analysts*.\n"
    "Describe the reasoning of an RL trading model using only "
    "provided framework data.\n"
    "Be precise, quantitative, and grounded in listed features, "
    "avoiding speculation.\n"
    "Follow these clarity rules:\n"
    "- One main clause per sentence.\n"
    "- Keep average sentence length under 18 words.\n"
    "- Mention at least two important indicators and one reward "
    "driver.\n"
    "- Correlate the explanation with stability metrics "
    "(AIM/AUM) and policy fidelity.\n"
    "- Avoid AI-specific jargon (no 'embedding', 'neuron', "
    "or 'weights').\n"
    "- Target Gunning Fog 10-12 for expert readability.\n"
    "- Write 100-120 words of coherent analytical prose but keep "
    "going if needed to finish your point."
  )
}
```

E.3.3 Counterfactual Explanation Template

For counterfactual "what-if" explanations, the user message is:

```
{
  ## Counterfactual Explanation Context

  Available Features:
    Important Features: {allowed_features}
    Reward Drivers: {top_reward_features}

  **Task**
  Using the context above, state one counterfactual:
  - Choose one feature from {allowed_features} or {top_reward_features}.
  - Describe briefly how a realistic increase or decrease could
    change the action.
  - Keep to {target_words} words.
}
```

The corresponding system instructions are:

```
{
  "system": (
    "You generate a counterfactual explanation for a trading agent's "
    "action.\n"
    "Describe how a realistic change in one provided indicator could "
    "alter the decision.\n"
    "Do not introduce unlisted features or external events."
  )
}
```

E.3.4 Cross-Model Synthesis Template

The synthesis prompt aggregates interpretability findings across multiple agents. The user message is:

```
{
  ## Comparative Explainability Summary

  Framework Layers:
    A - Feature Attribution
    B - Stability
```

```
    C - Policy Interpretability
    D - Reward Attribution

  Models: PPO, DDPG, TD3, A2C.

  **Task**
  1. Summarise interpretability and stability for each model.
  2. Highlight which has strongest feature grounding and reward
     alignment.
  3. Recommend the most transparent agent for {audience}-level
     readers.
  4. Keep within {target_words} words.
}
```

The system message for this mode is:

```
{
  "system": (
    "You synthesise interpretability findings across multiple RL "
    "agents (PPO, DDPG, TD3, A2C).\n"
    "Focus on clarity, stability, and transparency rather than "
    "profit.\n"
    "Explicitly reference Layers A–D when comparing agents."
  )
}
```

E.3.5 Historical-Context (E-Layer) Template

When a short historical description is available, the prompt extends Layers A–D with an additional E-layer context block:

```
{
  ## Contextual Explainability (E-Layer)
  Model: {model} | Checkpoint: {checkpoint}
  Regime: {regime}
  Action: {action} | Reward: {reward}

  Key Features (A): {allowed_features}
```

```
  Stability (B):
    AIM={aim}, AUM={aum}, RIS={ris}
  Policy Interpretability (C):
    Fidelity={policy_fidelity:.3f},
    TopFeatures={policy_top_features},
    Rules={policy_rules}
  Reward Drivers (D): {top_reward_features}

  Historical Context (curated): {historical_context}

  **Task**
  1. Explain the behaviour using only Layers A-D and the given
     context.
  2. Link features and reward drivers to relevant contextual clues.
  3. Avoid speculation or forward-looking statements.
  4. Limit to {target_words} words.
}
```

The system instructions for this historical mode are:

```
{
  "system": (
    "You are explaining an RL agent's decision using its "
    "explainability data (Layers A-D) and a short historical "
    "context provided below.\n"
    "Use only the supplied context and indicators--no external "
    "knowledge.\n"
    "If the context is insufficient, say 'Insufficient information.'"
  )
}
```

E.4 Corpus Record Schema

```
{
  "model": "<string: RL agent id, e.g. 'a2c' | 'ppo' | 'ddpg' | 'td3'>",
  "date": <int: UNIX timestamp in milliseconds>,
  "tic": "<string: underlying DJIA ticker>",
```

```
"regime": "<string: 'bull' | 'bear' | 'sideways'>",

"action": <int: discrete action id, e.g. 0=SELL, 1=HOLD, 2=BUY>,
"reward": <float: one-step reward for this decision>,
"account_value": <float: portfolio account value after the step>,

// Local technical indicators for the active asset
"macd": <float: moving average convergence/divergence>,
"rsi_30": <float: 30-day relative strength index>,
"cci_30": <float: 30-day commodity channel index>,
"dx_30": <float: 30-day directional index>,
"turbulence": <float: market turbulence score>,

// Layer 1: attribution and importance metrics
"AIM": <float: masking-based importance metric>,
"AUM": <float: masking-based unimportance metric>,
"RIS": <float: perturbation stability score>,

// Layer 1: global and rolling attribution strings
"attr_global_features":
  "<pipe-separated list of 'feature_name (weight)'>",
"attr_rolling_features":
  "<pipe-separated list of 'feature_name (weight)' "
  "for the rolling window>",

// Layer 2-3: policy surrogate summaries
"policy_fidelity":
  <float: surrogate accuracy vs. base agent>,
"policy_top_features":
  "<pipe-separated list of dominant feature groups, "
  "e.g. 'imp_turbulence|imp_rsi_30|...'>",
"policy_rules":
  "<text representation of the extracted decision tree, "
  "with indentation and class labels>",

// Layer 4: reward attribution summaries
"reward_top_features":
```

```
    "<pipe-separated list of motifs contributing most to return>",
  "reward_exposed_features":
    "<pipe-separated list of features with high exposure>",
  "reward_corr_features":
    "<pipe-separated list of features with high return correlation>",
  "reward_attr_mean":
    <float: average absolute reward attribution>,
  "reward_corr_mean":
    <float: average absolute return correlation>,
  "reward_method":
    "<comma-separated list of attribution methods combined, "
    "e.g. 'GradSHAP,IG,SHAP,Saliency'>",

  // Feature vocabulary and guard-rail control
  "allowed_features":
    "<pipe-separated list of feature names and (optionally) "
    "their weights that may be referenced in the explanation>"
}
```

F Questionnaires Content

The Research Ethics and Data Protection (REDP) Form for this study was submitted to the Faculty Research Ethics Committee and was acknowledged (Application ID ICT-2025-00022).

F.1 Pre-Study Questionnaire

Source: Google Forms (Pre-Study Questionnaire):
`https://forms.gle/yh5Tty4mVXzCexM28`

F.1.1 Information Letter

My name is Nathan Portelli, and I am a Master's student pursuing a degree in Artificial Intelligence at the University of Malta. As part of my research dissertation titled *"Enhancing Transparency and Interpretability in AI-Driven Algorithmic Trading,"* supervised by Dr. Vincent Vella, I am conducting a study to evaluate the effectiveness of Explainable AI (XAI) methods in improving the interpretability of AI-driven trading strategies.
Study Overview: The study aims to investigate how Explainable AI techniques, including traditional XAI outputs (e.g., feature importance visualisations) and Natural Language Processing (NLP)-based narrative explanations, can enhance the transparency and usability of AI models in algorithmic trading. Specifically, I am exploring whether these methods improve trust, decision-making, and accessibility for both expert and non-expert users in financial systems.
Your Involvement: Your participation would involve reviewing AI-generated trading explanations and providing feedback on their clarity, usefulness, and effectiveness through a structured questionnaire. The process will take approximately 1 hour, and your insights will contribute to advancing the field of explainable AI in financial technology.
Confidentiality and Data: All data collected will be anonymised and treated with strict confidentiality. Your responses will be securely stored and used solely for research pur-

poses. Any identifying information will be removed before publication or presentation of the findings.
Voluntary Participation: Participation in this study is entirely voluntary. You are free to withdraw at any time without providing a reason, and your decision will not affect your relationship with the University of Malta. Should you choose to withdraw, your data will be erased where possible, unless anonymisation or research objectives prevent its removal.
Your input is invaluable to this research, as it will help shape more transparent and trustworthy AI systems in finance. If you have any questions or concerns, please contact me at *nathan.portelli@um.edu.mt* or my supervisor, Dr. Vincent Vella, at *vvell04@um.edu.mt*.
Thank you for considering participation in this study.
* **Indicates required question**

F.1.2 Consent Form

I, the undersigned, give my consent to take part in the study conducted by Nathan Portelli. This consent form specifies the terms of my participation in this research study.

1. I have been given written and/or verbal information about the purpose of the study; I have had the opportunity to ask questions and any questions that I had were answered fully and to my satisfaction. I also understand that I am free to accept to participate, or to refuse or stop participation at any time without giving any reason and without any penalty. Should I choose to participate, I may choose to decline to answer any questions asked. In the event that I choose to withdraw from the study, any data collected from me will be erased as long as this is technically possible (for example, before it is anonymised or published), unless erasure of data would render impossible or seriously impair achievement of the research objectives, in which case it shall be retained in an anonymised form.

2. I understand that I have been invited to participate in a usability study, in which I will evaluate AI-generated trading explanations. This will involve reviewing both traditional Explainable AI (XAI) outputs (e.g., feature importance visualisations, saliency maps) and Natural Language Processing (NLP)-based narrative explanations and then completing a questionnaire assessing these explanations based on predefined criteria (e.g., trust, interpretability, and ease of understanding). The evaluation process will take approximately 30-45 minutes and will be conducted at a location and time convenient for me.

3. I understand that my participation does not entail any known or anticipated risks.
4. I understand that there are no direct benefits to me from participating in this study. However, this research may benefit others by enhancing the transparency and interpretability of AI-driven trading strategies, ultimately making financial AI systems more accessible and trustworthy for traders and investors.
5. I understand that, under the General Data Protection Regulation (GDPR) and national legislation, I have the right to access, rectify, and where applicable, ask for the data concerning me to be erased.
6. I understand that all data collected that does not form part of the publication will be permanently erased following the completion of this study and the publication of results.
7. I am aware that my identity and personal information will not be revealed in any publications, reports, or presentations arising from this research.
8. I have been provided with a copy of the information letter and understand that I will also be given a copy of this consent form.
9. I am aware that the study will be held online; the researcher will use Zoom and will activate the Require Encryption for 3rd party endpoints SIP/H-323 function. The researcher will only audio record the session if permission is provided in (11).
10. I am aware that, by marking the first tick box below, I am giving my consent for this usability study to be audio recorded and converted to text as it has been recorded (transcribed). Mark only one oval: (=:) I agree to this usability study being audio recorded. (=:) I do not agree to this usability study being audio recorded.
11. I am aware that extracts from my interview may be published (e.g., in a dissertation, academic journals) and/or presented (e.g., during conferences, meetings), either in anonymous form or using a pseudonym (a made-up name or code, e.g., respondent A). Mark only one oval: (=:) I would like to review extracts of my interview transcript that the researcher would like to publish and/or present before these are published/presented. (=:) I would not like to review extracts of my interview transcript that the researcher would like to publish and/or present before these are published/presented.

By continuing this questionnaire, you declare the following (tick the boxes to agree):

☐ I have been provided the Information Letter and have read about the nature of the study, what my involvement entails, and the data management policy.

☐ I have had the opportunity to ask any questions which have been answered satisfactorily.

☐ I understand that I can withdraw my participation at any time and for any reason.

☐ I have read and understood the above statements and agree to participate in this study.

F.1.3 Participant Background

What best describes your professional background? Mark only one oval: (=:) Student (Finance, Economics, or Related Field) (=:) Professional Trader / Investment Analyst (=:) Financial Services Professional (e.g. Banking, Insurance, FinTech) (=:) Technology Professional (e.g. Data Science, Software Engineering, AI) (=:) Academic / Researcher (=:) Other Business Professional (e.g. Management, Consulting) (=:) Other:

How frequently do you make trading decisions? Mark only one oval: (=:) Daily (=:) Weekly (=:) Monthly (=:) Rarely (=:) Never

Have you previously used any algorithmic or automated trading platforms or systems? Mark only one oval: (=:) Yes (=:) No
If yes, which tools or platforms do you use? _____

Have you worked with any explainability techniques (e.g., SHAP, LIME, saliency maps)? Mark only one oval: (=:) Yes (=:) No (=:) Not sure

Have you ever received an AI-based trading recommendation that you distrusted or ignored? Mark only one oval: (=:) Yes (=:) No
If yes, briefly describe why you did not trust it: _____

F.1.4 Understanding of Explainability and Transparency

Have you heard of Explainable AI (XAI) before? Mark only one oval: (=:) Yes (=:) No
If yes, how would you define it in your own words? _____

In your opinion, should AI systems be required to explain the rationale behind their trading decisions? Mark only one oval: (=:) Yes (=:) No (=:) Not sure
Why or why not? _____

Have you ever felt that AI or algorithmic trading systems were too opaque or difficult to understand? Mark only one oval: (=:) Yes (=:) No
If so, what made them hard to interpret? _____

F.1.5 Expectations Toward Explanation Systems

Which type of explanation would you find most useful when reviewing a trading decision? Mark only one oval: (=:) Visual (charts, feature-importance plots) (=:) Text-based (natural-language summaries) (=:) Combination of both (=:) Other:

What level of detail do you expect from an AI explanation? Mark only one oval: (=:) Minimal (1-2 key points) (=:) Moderate (=:) In-depth (with evidence and reasoning) (=:) Other:

How would you prefer explanations to be structured? (tick all that apply)

- ☐ Risk or confidence breakdowns
- ☐ Comparison to historical performance
- ☐ Feature attribution or factor weights
- ☐ Decision-path logic (why Buy/Sell/Hold)
- ☐ Plain-language summary
- ☐ Other:

Do you believe AI explanations should adapt their complexity to the user's experience level? Mark only one oval: (=:) Yes (=:) No (=:) Unsure

F.1.6 Reactions to AI Explanations

If an AI explanation contradicted your own analysis, how would you react? Mark only one oval: (=:) Re-evaluate my stance (=:) Ignore the AI (=:) Seek more information (=:) Other:

If an AI explanation was unclear or overly complex, what would you do? Mark only one oval: (=:) Ignore it (=:) Seek human interpretation (=:) Request simplified output (=:) Other:

Would you feel comfortable basing a trading decision solely on an AI-generated explanation if it were clear and well-supported? Mark only one oval: (=:) Yes (=:) No (=:) Maybe

Why or why not? _____

How much control would you like over the depth of the explanation? Mark only one oval: (=:) None (=:) Basic filtering (=:) Ability to toggle detail (=:) Full customisation (=:) Other:

F.1.7 Information Ranking

Please rank the following from the most important information to the least important information

Table F.1: Importance Ratings for Explanation Components

Component	**Least Important**	**Not Very Important**	**Important**	**Very Important**	**Extremely Important**
Feature Importance (what factors the model used)	◯	◯	◯	◯	◯
Regime Awareness (how logic changes under conditions)	◯	◯	◯	◯	◯
Policy Rationale (why certain actions were taken)	◯	◯	◯	◯	◯
Profit Drivers (which factors drove results)	◯	◯	◯	◯	◯
Human-Readable Summary (plain-language explanation)	◯	◯	◯	◯	◯

F.1.8 Final Thoughts

What would make AI explanations most useful to you in practice? _____
Do you have any concerns about using AI explainability tools in trading? _____

F.2 Pre-Study Responses

This section summarises responses from the pre-study questionnaire ($N = 12$), organised by question rather than by individual respondent.

Consent and Participation

Audio recording (Q11). All 12 respondents completed the consent section. A clear majority agreed to the usability study being audio recorded and transcribed, with a small minority opting out.

Review of transcript extracts (Q12). Most participants indicated that they would like to review any interview extracts before publication or presentation, while a minority preferred not to review extracts in advance.

Tick-box consent (Q13). All participants indicated that they had received the information letter, had the opportunity to ask questions, understood they could withdraw at any time, and agreed to take part in the study.

Participant Background

Age range (Q: "What is your age range?", Pre-Study.4). The age distribution was:

- 18–23: 5 participants
- 24–34: 5 participants
- 35–45: 1 participant
- 55 or older: 1 participant

Professional background (Q: "What best describes your professional background?", Pre-Study.5). Reported backgrounds were:

- Financial Services Professional (e.g. Banking, Insurance, FinTech): 5
- Technology Professional (e.g. Data Science, Software Engineering, AI): 3
- Student (Finance, Economics, or Related Field): 3
- Other Business Professional (e.g. Management, Consulting): 1

Trading frequency (Q: "How frequently do you make trading decisions?", Pre-Study.6).

- Weekly: 4
- Daily: 3
- Rarely: 3
- Monthly: 2

Familiarity with AI-driven trading (Q: "How familiar are you with AI-driven trading strategies?", Pre-Study.7).

- Moderately familiar: 7
- Slightly familiar: 4
- Very familiar: 1

Use of algorithmic / automated platforms (Q: "Have you previously used any algorithmic or automated trading platforms or systems?", Pre-Study.8–9).

- Yes: 2
- No: 10

Those who answered "Yes" mentioned:

- Revolut RoboAdvisor (two mentions, with slight spelling variation).

Experience with explainability techniques (Q: "Have you worked with any explainability techniques (e.g. SHAP, LIME, saliency maps)?", Pre-Study.10).

- No: 6
- Not sure: 3
- Yes: 3

Confidence interpreting technical indicators (Q: Pre-Study.11). Self-reported confidence in interpreting indicators such as RSI, MACD, moving averages, and volume was:

- Moderately confident: 8
- Slightly confident: 3
- Very confident: 1

Prior distrust of AI-based trading recommendations (Q: Pre-Study.12–13).

- Have previously distrusted or ignored an AI-based recommendation: 2
- Have not: 10

Reasons for distrust included:

- Concerns that the recommendation did not fully account for geopolitical and macroeconomic instability at the time, raising doubts that such factors were properly incorporated.
- Perception that the recommendation was based on outdated knowledge.

Understanding of Explainability and Transparency

Awareness of XAI (Q: "Have you heard of Explainable AI (XAI) before?", Pre-Study.14–15).

- Yes: 3
- No: 9

Those familiar with XAI described it in their own words along themes such as:

- AI systems often operate as a "black box", and XAI methods help expose the underlying logic or "show the working" behind outputs in understandable language.
- XAI helps people understand why AI makes particular decisions or predictions, thereby building trust.
- Explanations should accompany outputs, highlighting the reasoning in a transparent, human-readable way rather than providing only a bare result.

Requirement for explanations (Q: "In your opinion, should AI systems be required to explain the rationale behind their trading decisions?", Pre-Study.16–17).

- Yes: 12
- No / Not sure: 0

Justifications emphasised:

- The need for transparency in high-stakes financial decisions and regulatory / auditability requirements.
- The importance of understanding underlying drivers to assess robustness, avoid blind trust, and detect potential errors.
- The view that explanations support accountability and informed human oversight.

Perceived opacity of AI or algorithmic systems (Q: Pre-Study.18–19).

- Have found such systems too opaque or difficult to understand: 9
- Have not: 3

When systems felt opaque, respondents mentioned factors such as:

- Limited visibility into the underlying model, data, or decision rules.
- Overly technical presentation of outputs without sufficient contextualisation.
- Difficulty in connecting numerical or signal-based outputs to intuitive reasoning.

Expectations Toward Explanation Systems

Preferred explanation type (Q: "Which type of explanation would you find most useful?", Pre-Study.20).

- Combination of both visual and text-based explanations: 8
- Visual (charts, feature-importance plots): 3
- Text-based (natural-language summaries): 1

Expected level of detail (Q: Pre-Study.21–22).

- In-depth (with evidence and reasoning): 10
- Moderate detail: 2
- Minimal: 0

Preferred structure of explanations (Q: "How would you prefer explanations to be structured? (tick all that apply)", Pre-Study.23). Participants could select multiple options. Across the 12 respondents, the most frequently selected structural components were:

- Plain-language summary.
- Risk or confidence breakdowns.
- Feature attribution or factor weights.
- Decision-path logic (why Buy / Sell / Hold was chosen).
- Comparison to historical performance.

Most respondents selected a combination of these, indicating a preference for multi-layered explanations rather than a single view.

Adaptation to user experience (Q: "Do you believe AI explanations should adapt their complexity to the user's experience level?", Pre-Study.24).

- Yes: 11
- Unsure: 1
- No: 0

Reactions to AI Explanations

If an AI explanation contradicted own analysis (Q: Pre-Study.25).

- Seek more information: 7
- Re-evaluate own stance: 4
- Ignore the AI: 1
- Other: 0

If an AI explanation was unclear or overly complex (Q: Pre-Study.26).

- Request simplified output: 7
- Seek human interpretation: 4
- Ignore it: 1

Comfort basing a decision solely on a clear AI explanation (Q: Pre-Study.27–28).

- Maybe: 7
- Yes: 3
- No: 2

Reasons included:

- Supportive participants stressed that a well-supported, transparent, and empirically grounded explanation could provide sufficient confidence for decision-making.
- More cautious participants preferred to use AI as an input alongside their own analysis, citing concerns around data quality, model assumptions, and unmodelled risks.

Desired control over explanation depth (Q: Pre-Study.29).

- Ability to toggle detail: 8
- Full customisation: 2
- Basic filtering: 1
- None: 1

Importance Ratings for Explanation Components

Participants rated each explanation component on a five-point importance scale (Least, Not Very, Important, Very, Extremely). Table F.2 summarises the distribution of responses; the original placeholder counts have been replaced with the observed frequencies from the 12 respondents.
Overall, human-readable summaries and policy rationale were most frequently rated as "Extremely Important", while regime-awareness and feature importance tended to cluster around "Important" and "Not Very / Least Important" for some participants.

Final Thoughts

What would make AI explanations most useful in practice? (Q: Pre-Study.35). Participants described a range of desiderata, including:

- Clear linkage between explanations and real trading decisions, including concrete examples and highlighting risk / reward trade-offs.

Table F.2: Importance ratings for explanation components (Pre-Study, $N = 12$).

Component	Least	Not Very	Important	Very	Extremely
Feature Importance (what factors the model used)	3	3	3	2	1
Regime Awareness (how logic changes under conditions)	3	4	3	1	1
Policy Rationale (why certain actions were taken)	2	1	4	2	3
Profit Drivers (which factors drove results)	2	2	3	3	2
Human-Readable Summary (plain-language explanation)	2	1	3	1	5

- Simple, jargon-free language for high-level explanations, with the option to drill down into technical detail when needed.
- Integration of asset performance, related news, identified risks, and scenario-based justifications that describe the conditions under which a recommendation is expected to pay off.
- Reliable, up-to-date information sources and access to live metrics.

Concerns about using AI explainability tools in trading (Q: Pre-Study.36). Eight participants provided qualitative concerns, including:

- Risk of inaccurate, biased, or misleading explanations, particularly if the underlying data are poor or the model is mis-specified.
- Worries that users may over-rely on AI or offload responsibility to the system, especially if explanations appear persuasive but are not well-founded.
- Fear that non-transparent or overly complex explanations could discourage adoption or foster mistrust.

- Concerns about copy-trading dynamics and herd behaviour if many users follow similar AI-guided strategies.
- The possibility that polished explanations can obscure the true complexity and uncertainty of financial markets.

F.3 User Study Questionnaire

Source: Google Forms (User Study Questionnaire):
`https://forms.gle/1te8s8ExW9kLLtcpalKn9Jx3AQMoLFeKj7Pxcl7ZKZ_O`

F.3.1 Introduction

Thank you for taking part in this short study on how people understand explanations from AI-driven trading systems.
You will see several examples showing how different trading agents explain their decisions.
Each example includes a short description and a simple chart or diagram.
After reading each one, you will answer a few quick questions about how clear, useful, or trustworthy the explanation felt to you.
There are no right or wrong answers, we only care about your impressions.

F.3.1.1 What this study is about

Modern trading systems often use Reinforcement Learning (RL), a kind of AI that learns by trial and error to make trading decisions.
Each agent, such as PPO, DDPG, or TD3, learns its own strategy:

- PPO (Proximal Policy Optimisation): tends to make steady, balanced decisions and avoids large swings.
- DDPG (Deep Deterministic Policy Gradient): is more adaptive, reacting to small market changes.
- TD3 (Twin Delayed DDPG): focuses on reducing noise and avoiding overly risky moves.

These models can perform well but are often hard to understand as they do not easily show why they made a buy, hold, or sell choice.

That is where Explainable AI (XAI) comes in. XAI methods turn complex model behaviour into clear, human-readable explanations, using visuals (charts) or narratives (short texts).

F.3.1.2 Why we are doing this

This study aims to learn how traders like you understand and trust these explanations. By comparing visuals and written narratives, we can discover which styles help people interpret AI trading models more easily.

F.3.1.3 What you will do

1. View the visual explanation image and the short description, and attempt to decipher it.
2. Read the narrative description.
3. Answer 5 short questions after each example.
4. Finish with a short reflection on which format felt clearer or more trustworthy.

Your responses are anonymous and used only for academic research.
* **Indicates required question**

F.3.2 Rolling SHAP Feature Contributions

This example shows how the PPO trading model's key market indicators changed in importance over time.
You will see how certain signals sometimes helped performance while others reduced it, reflecting how the agent adapted to shifting conditions.
Visual Explanation: (Figure F.1)
Narrative Description: From October to November the agent faces sustained negative pressure, led by DIS cci 30 and reinforced by a smaller drag from BA rsi 30. This is partly offset by a persistent positive pull from BA boll ub, which remains one of the most supportive factors throughout the period.
In December the balance stabilises as the adverse effect of DIS cci 30 diminishes. Entering January the positive side compresses briefly, then from February the negatives deepen again, largely through HD dx 30 and a growing influence from BA rsi 30.
The most favourable phase arrives in April and May. BA boll ub strengthens markedly and is joined by steady lifts from DOW boll ub and CVX boll lb, with intermittent additional

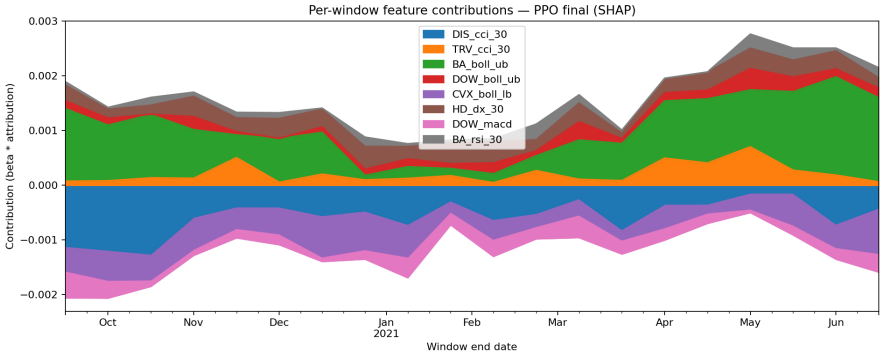

Figure F.1: Rolling SHAP feature contributions for PPO, showing how each indicator's influence on returns changed through time. Positive bands reflect features that helped performance, while negative ones reduced it.

support from TRV cci 30. This combination produces the strongest net positive tendency in the series.

By June the supportive effects recede and the aggregate influence weakens, with BA rsi 30 and HD dx 30 remaining as the main opposing signals, yielding a more even but lower-magnitude mix of contributions.

Questions* Mark only one oval per row.

Table F.3: Perceived Clarity and Trustworthiness for PPO Example

	Strongly Disagree	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
The explanation was clear and easy to understand.	◯	◯	◯	◯	◯
The visual helped me identify which indicators were important.	◯	◯	◯	◯	◯
I understood how the model's reasoning changed over time.	◯	◯	◯	◯	◯
The chart and text together helped me interpret the decision.	◯	◯	◯	◯	◯
I would trust this explanation in a real trading scenario.	◯	◯	◯	◯	◯

Which explanation is more easily understandable?* Mark only one oval: (=:) Visual explanation (=:) Narrative description (=:) Both equally (=:) None (=:) Other:

Any quick thoughts on this example? _____

F.3.3 Decision Tree Policy

This section shows how the TD3 trading model made its buy, hold, or sell decisions using combinations of indicators.
It highlights how certain conditions, like momentum or volatility levels, triggered specific trading actions.
Visual Explanation: (Figure F.2)

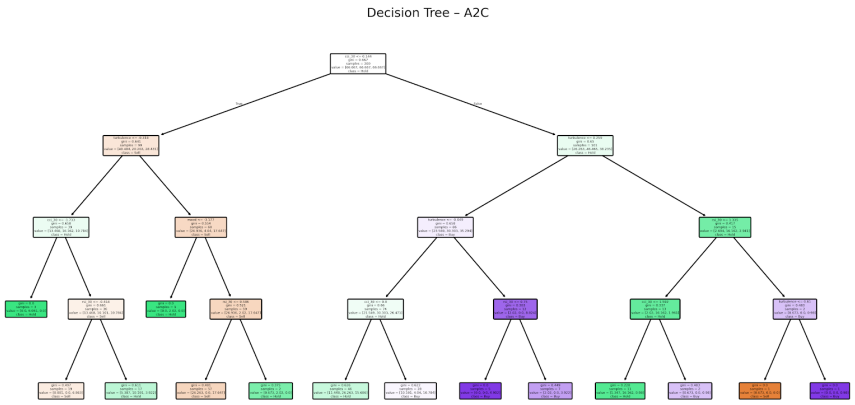

Figure F.2: Decision-tree view showing how the trading model linked indicators such as MACD, CCI, and turbulence influences the decision for Buy/Hold/Sell actions.

Narrative Description: The surrogate rules first test `cci 30` at 0.144. When `cci 30` is at or below this level, the rules emphasise adverse conditions. If `turbulence` is at or below -0.318, two outcomes appear: when `cci 30` falls further to -1.733 or below, the decision settles into *Hold*; otherwise the next check is `rsi 30` at -0.414, where lower values lead to *Sell* and higher values return to *Hold*. If `turbulence` is above -0.318, the rules look to `macd` at -3.177: values at or below this level point clearly to *Sell*. If `macd` is higher than -3.177, the final check is `rsi 30` at 0.586, with lower readings keeping the outcome mostly *Sell*, and higher readings allowing *Hold*.
When `cci 30` is above 0.144, the rules favour more constructive outcomes provided `turbulence` does not rise too much. With `turbulence` at or below 0.259, the next test is `rsi 30` at 0.75, but both sides of this split end in *Buy*, so relatively calmer conditions here support *Buy*. If `turbulence` exceeds 0.259, the decision turns on `rsi 30` at 1.335. At or below this value, a further check on `cci 30` at 1.592 typically maintains *Hold* outcomes on both sides. Above 1.335, a small branch considers `turbulence` at 0.61, where lower values yield *Buy* and higher values produce a rare *Sell*.

Overall, `cci 30` provides the first, coarse separation of situations. On the lower side of `cci 30`, very negative `turbulence`, depressed `macd`, or low `rsi 30` push the decision towards *Sell*, unless `cci 30` is extremely low, in which case *Hold* can prevail. On the higher side of `cci 30`, outcomes tend towards *Buy* when `turbulence` remains contained and `rsi 30` is not excessive; once `turbulence` rises, the rules rely on finer thresholds of `rsi 30` and `cci 30`, usually resulting in *Hold*, with smaller pockets of *Buy* and occasional *Sell* when `turbulence` is high.
Questions* Mark only one oval per row.

Table F.4: Perceived Clarity and Logic for TD3 Policy Example

	Strongly Disagree	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
I could follow how the model's decision process works.	◯	◯	◯	◯	◯
The rule-based structure helped me understand why it bought or sold.	◯	◯	◯	◯	◯
The visual made it clear what each indicator threshold represents.	◯	◯	◯	◯	◯
The narrative complemented the visual explanation well.	◯	◯	◯	◯	◯
I would find this kind of explanation trustworthy for understanding model rules.	◯	◯	◯	◯	◯

Which explanation is more easily understandable?* Mark only one oval: (=:) Visual explanation (=:) Narrative description (=:) Both equally (=:) None (=:) Other:
Any part of the decision logic that was unclear? _____

F.3.4 Integrated Gradients Reward Attribution

This part illustrates how individual market indicators contributed to the TD3 model's overall trading rewards.
It breaks down which factors supported profitable outcomes and which ones led to lower returns.
Visual Explanation: (Figure F.3)
Narrative Description: From October to December, the net contribution is predominantly negative. The largest adverse effect is associated with WMT rsi 30, with additional drag from MRK cci 30 and a smaller negative component from TRV rsi 30.

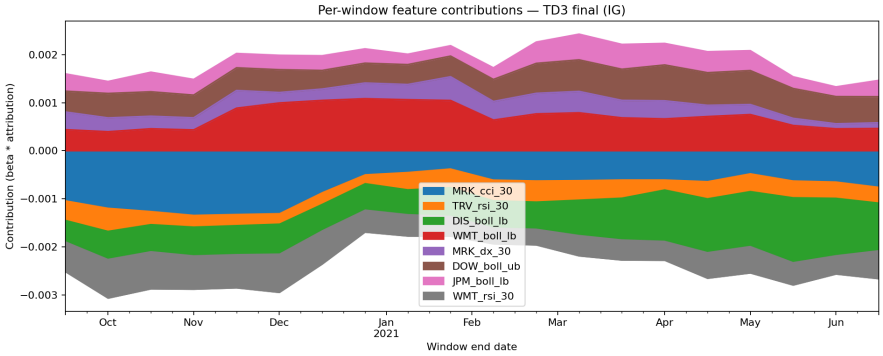

Figure F.3: Feature-level reward contributions for TD3 across time, showing how indicators like DIS boll_lb and WMT boll_lb shaped returns.

From January, the magnitude of the negative terms declines as WMT rsi 30 and MRK cci 30 move closer to neutral. Over the same period, positive contributors strengthen: WMT boll lb becomes the most persistent source of uplift, DOW boll ub provides a stable positive base, JPM boll lb adds further support, and MRK dx 30 supplies a modest positive increment between January and March.

The most favourable intervals occur around March and again in late April, when the positive contributors—principally WMT boll lb, DOW boll ub, JPM boll lb, and to a lesser extent MRK dx 30—align to produce the highest net contributions.

By June, the absolute magnitudes across features are smaller, indicating weaker overall explanatory influence and a more even balance between supportive and opposing signals.

Questions* Mark only one oval per row.

Which explanation is more easily understandable?* Mark only one oval: (=:) Visual explanation (=:) Narrative description (=:) Both equally (=:) None (=:) Other:

What stood out most in this example? _____

F.3.5 Attribution Stability, Rolling Integrated Gradients

Here, you will see how consistently the DDPG model relied on different indicators across time.

Stable signals suggest strong, recurring influences, while fluctuating ones reflect changing or uncertain importance.

Visual Explanation: (Figure F.4)

Narrative Description: The trading agent's importance weights concentrate on a small group of indicators, with clear temporal phases. Early in the horizon the signals are modest

Table F.5: Perceived Clarity and Interpretation for TD3 Reward Attribution Example

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The explanation clearly showed which indicators helped or hurt performance.	◯	◯	◯	◯	◯
The visual made it easy to tell when the model performed better or worse.	◯	◯	◯	◯	◯
I understood why some features had positive and others negative influence.	◯	◯	◯	◯	◯
The description helped me connect the features to the market conditions.	◯	◯	◯	◯	◯
I could imagine using this type of chart to evaluate trading models.	◯	◯	◯	◯	◯

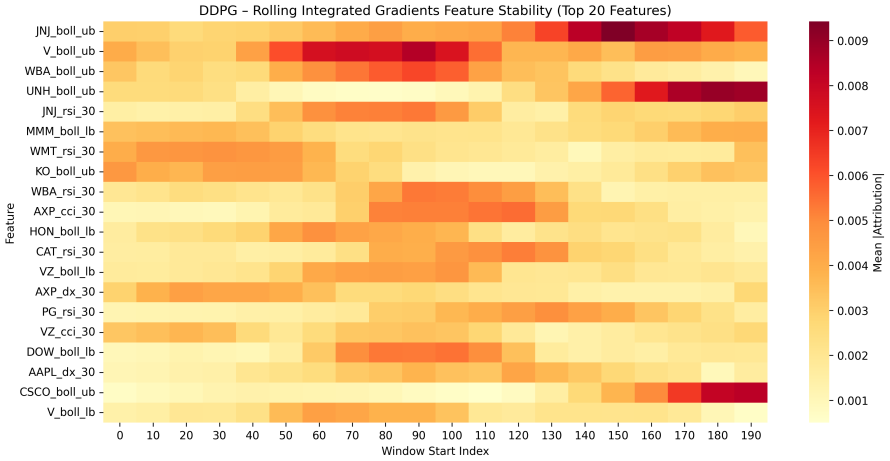

Figure F.4: Rolling stability of the top 20 features for DDPG, showing how consistent each indicator's importance was over time.

across the board, with only mild emphasis on lower-band momentum such as MMM boll lb and short-term strength from WMT rsi 30.

From roughly the middle of the period the focus intensifies around upper Bollinger bands, most notably V boll ub, which becomes the dominant and most persistent source of importance for an extended stretch. During the same interval the agent also gives sustained attention to AXP cci 30, WBA rsi 30, and DOW boll lb, indicating a secondary cluster of

features that matter when the main signal is active.
Towards the end, the pattern sharpens again with strong late surges in WBA boll ub and UNH boll ub, and brief spikes in AAPL dx 30 and CSCO boll ub. By contrast, features like JNJ rsi 30, HON boll lb, and CAT rsi 30 remain comparatively weak and intermittent throughout.
Overall, the model repeatedly returns to upper-band Bollinger signals for several constituents, augmented at times by CCI and RSI measures, with the most stable and pronounced reliance occurring in the mid to late portion of the study window.
Questions* Mark only one oval per row.

Table F.6: Perceived Stability and Comprehension for DDPG Attribution Example

	Strongly Disagree	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
I could understand how stable or variable each indicator was.	◯	◯	◯	◯	◯
The visual helped me see when the model focused on certain features.	◯	◯	◯	◯	◯
The description made it easier to interpret what darker or lighter areas mean.	◯	◯	◯	◯	◯
The explanation felt simple and clear to follow.	◯	◯	◯	◯	◯
I would find this information useful for judging a model's consistency.	◯	◯	◯	◯	◯

Which explanation is more easily understandable?* Mark only one oval: (=:) Visual explanation (=:) Narrative description (=:) Both equally (=:) None (=:) Other:
Any observations about which features seemed most stable? _____

F.3.6 How Market Indicators Affected Reward

This section shows how various indicators directly impacted the DDPG model's reward outcomes.
It helps reveal which signals guided successful trades and which ones were associated with poorer performance.
Visual Explanation: (Figure F.5)

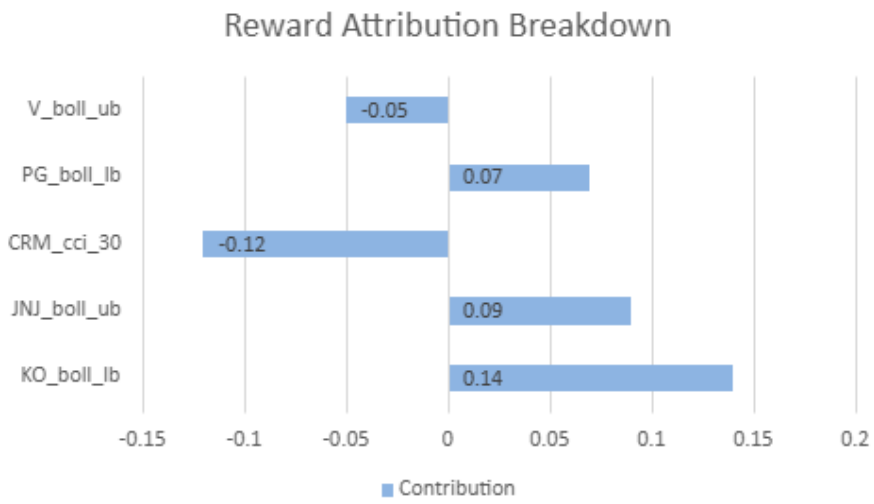

Figure F.5: How specific market indicators contributed to the model's overall reward outcomes.

Narrative Description: Reward attribution indicates a mixed set of influences with positives outweighing negatives. The largest uplift comes from KO boll lb (+0.14), followed by JNJ boll ub (+0.09) and PG boll lb (+0.07). Offsetting these, CRM cci 30 exerts the strongest adverse effect (-0.12), with a smaller negative from V boll ub (-0.05). Taken together, gains associated with KO and JNJ dominate the aggregate impact, while the CRM CCI signal is the principal drag.

Questions* Mark only one oval per row.

Which explanation is more easily understandable?* Mark only one oval: (=:) Visual explanation (=:) Narrative description (=:) Both equally (=:) None (=:) Other:

What was the most useful part of this explanation? _____

F.3.7 Market Regime (Dow Jones, 2009-2021)

This part provides market context, showing how broad conditions shifted between growth (bull), decline (bear), and sideways periods.

Understanding these trends helps interpret how trading models behave in different environments.

Visual Explanation: (Figure F.6)

Narrative Description: Between January 2009 and June 2021, the Dow Jones Industrial Average showed a long, upward trajectory with several periods of pause and correction. From 2009 onwards, the index recovered steadily from earlier lows, forming a consistent rising trend that continued for most of the decade.

Table F.7: Perceived Understanding and Realism for DDPG Reward Effects Example

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The explanation helped me understand which indicators improved or reduced rewards.	◯	◯	◯	◯	◯
The text made it clear how the agent reacted to strong or weak signals.	◯	◯	◯	◯	◯
I could follow which stocks or features had the biggest effect.	◯	◯	◯	◯	◯
The reasoning felt realistic and easy to understand.	◯	◯	◯	◯	◯
This explanation helped me see how the model connects indicators.	◯	◯	◯	◯	◯

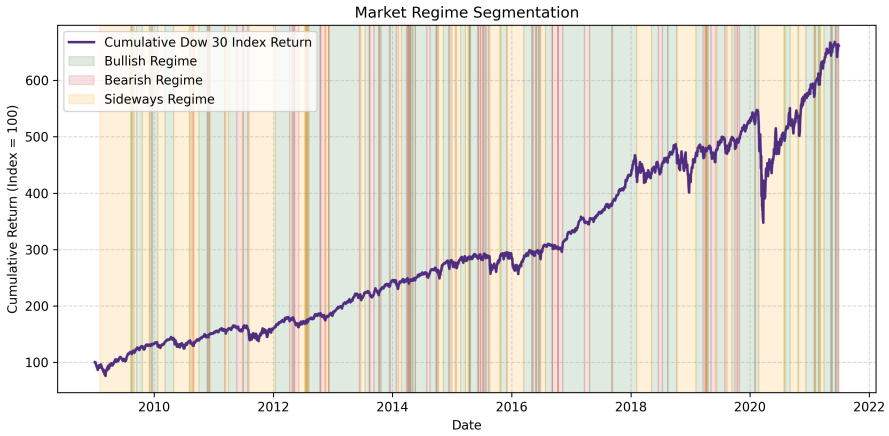

Figure F.6: Market trend of the Dow Jones between 2009 and mid-2021, showing long bullish stretches with short corrections.

Short-term declines appeared as brief dips, but each was followed by renewed upward movement, keeping the broader direction positive.

Around 2015-2016, the index flattened for a while, indicating slower momentum before rising again through 2017 and 2018.

A sharp fall occurred in early 2020, visible as a deep trough, after which the index recovered quickly and reached new highs by mid-2021.

Overall, the market was dominated by long bullish stretches, interrupted by only a few

short corrections and one major downturn with a strong rebound.
Questions* Mark only one oval per row.

Table F.8: Perceived Clarity of Market Regime Example

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The description clearly summarised the market behaviour over time.	◯	◯	◯	◯	◯
The chart made it easy to spot major rises and drops.	◯	◯	◯	◯	◯
The text was clear and simple to follow.	◯	◯	◯	◯	◯
I understood how the overall market trend looked across this period.	◯	◯	◯	◯	◯
This context helps me interpret how trading models might behave differently.	◯	◯	◯	◯	◯

Which explanation is more easily understandable?* Mark only one oval: (=:) Visual explanation (=:) Narrative description (=:) Both equally (=:) None (=:) Other:
Any additional thoughts about the market trend? _____

F.3.8 Feature Importance

This section compares the main indicators used by several trading models, PPO, DDPG, TD3, and A2C.
It shows which signals were most influential across models and how their decision focus differs.
Visual Explanation: (Figure F.7)
Narrative Description: Across the four models, the indicators show varied levels of influence.
RSI 30 and DX 30 emerge as consistently strong contributors, suggesting that relative strength and directional movement carried significant weight in the models' decision processes.
MACD and CCI 30 have moderate importance, indicating that trend-following and cyclical momentum signals were considered but not dominant.
Turbulence appears influential mainly in certain configurations, implying that market stability or disorder had differing impacts depending on the model.

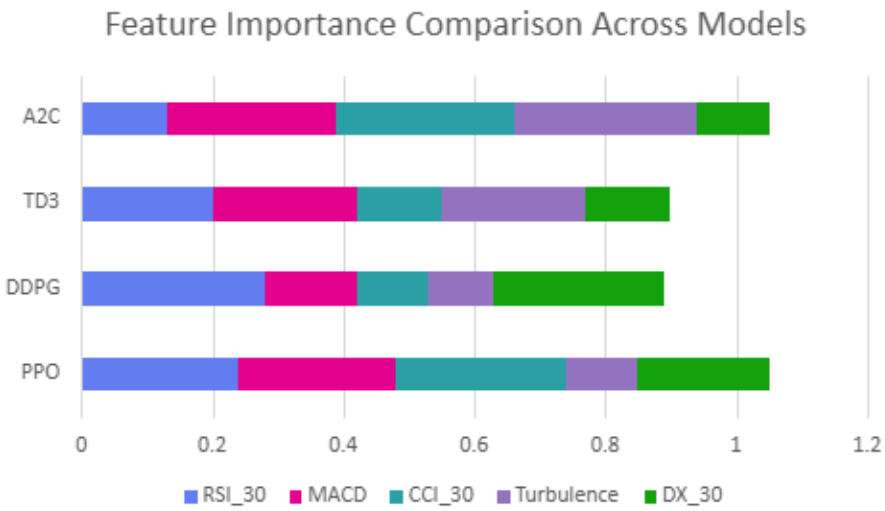

Figure F.7: Feature importance comparison showing how RSI 30, DX 30, MACD, and CCI 30 varied in influence across PPO, DDPG, TD3, and A2C.

Among the models, A2C relied more heavily on cyclical and stability measures, TD3 distributed its focus more evenly across indicators, and PPO and DDPG leaned more on momentum-based signals.

Overall, the data suggests that while all models drew from a common technical set, their weighting patterns reflect subtle but distinct strategic behaviours.

Questions* Mark only one oval per row.

Table F.9: Perceived Comparability Across RL Models

	Strongly Disagree	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
The comparison made it easy to see how models differ in what they focus on.	◯	◯	◯	◯	◯
The text clearly described each model's main strengths.	◯	◯	◯	◯	◯
The visual helped me compare features between models.	◯	◯	◯	◯	◯
I could understand which features were common across all models.	◯	◯	◯	◯	◯
I would find this type of comparison helpful when choosing a model.	◯	◯	◯	◯	◯

Which explanation is more easily understandable?* Mark only one oval: (=:) Visual ex-

planation (=:) Narrative description (=:) Both equally (=:) None (=:) Other:
Which model seemed most understandable or balanced? _____

F.3.9 Closing Reflection

Thank you for reviewing the examples. Please answer these final questions about your overall impressions.
Which format felt most trustworthy overall?* Mark only one oval: (=:) Visual only (=:) Text only (=:) Both equally (=:) None (=:) Other:

Any additional comments or observations about the explanations provided? _____

F.4 User-Study Responses

This section summarises the main questionnaire conducted during the explanation-viewing stage ($N = 12$). Results are grouped by example and then by closing questions, with responses aggregated per question.

Example 1: Rolling SHAP Feature Contributions (PPO)

Likert items (clarity, visual support, temporal reasoning, combined view, trust). Across the five statements:

- "The explanation was clear and easy to understand." *Agree* or *Strongly Agree*: 8 participants; 3 Neutral; 1 Strongly Disagree.
- "The visual helped me identify which indicators were important." *Strongly Agree*: 5; *Agree*: 4; 2 Disagree; 1 Neutral.
- "I understood how the model's reasoning changed over time." *Agree*: 7; *Strongly Agree*: 2; 2 Neutral; 1 Disagree.
- "The chart and text together helped me interpret the decision." *Agree*: 7; *Strongly Agree*: 3; 2 Neutral.
- "I would trust this explanation in a real trading scenario." *Agree*: 7; *Strongly Agree*: 2; 3 Neutral.

Preferred format for understandability.

- Both equally: 9
- Narrative description: 2
- None: 1

Open comments. Participants noted that:

- The narrative and visual complemented each other when read together.
- The time axis and terminology initially felt technical; the explanation became clearer once the narrative was read carefully.
- A prior "key" or legend was helpful for interpreting colours and bands.
- For first-time investors, the combined explanation was informative but might still be dense without further simplification.

Example 2: Decision Tree Policy (TD3)

Likert items (process understanding, rule structure, thresholds, complementarity, trust).

- "I could follow how the model's decision process works." *Strongly Agree*: 5; *Agree*: 4; 3 Neutral.
- "The rule-based structure helped me understand why it bought or sold." *Agree*: 7; *Strongly Agree*: 2; 2 Neutral; 1 Disagree.
- "The visual made it clear what each indicator threshold represents." Responses were mixed, with a majority Agree / Strongly Agree, some Neutral, and a minority Disagree, reflecting that threshold labelling was helpful for some but still confusing for others.
- "The narrative complemented the visual explanation well." Most respondents agreed or strongly agreed; a small number were Neutral or Disagree.
- "I would find this kind of explanation trustworthy for understanding model rules." Again, most participants agreed or strongly agreed, with only a small minority neutral or less positive.

Preferred format for understandability.

- Both equally: 9
- Visual explanation: 2
- Narrative description: 1

Open comments. Participants highlighted:

- The decision tree visual made the high-level branching structure clear.
- Some struggled with the exact interpretation of numerical thresholds and normalised indicator scales.
- The narrative helped clarify when branches led to Buy, Hold, or Sell, especially where the visual became dense.

Example 3: Integrated Gradients Reward Attribution (TD3)

Likert items (feature contributions, performance phases, positive/negative influence, link to conditions, usability). Across the five statements, responses showed:

- Majority *Agree* or *Strongly Agree* that the explanation clearly showed which indicators helped or hurt performance.
- Majority agreement that the visual made it easy to see periods of better or worse performance.
- Broad agreement that participants understood why some features had positive versus negative influence.
- Agreement that the description connected features to market conditions.
- Agreement that such charts could be useful for evaluating trading models, with only isolated neutral or negative responses.

Preferred format for understandability.

- Both equally: 7
- Narrative description: 3
- Visual explanation: 1
- None: 1

Open comments. Participants remarked that:

- The visual was particularly helpful once the narrative described which coloured bands corresponded to which stocks and indicators.
- The idea of "net contribution" over time was intuitive once explained but could still be heavy for non-technical users.

Example 4: Attribution Stability – Rolling Integrated Gradients (DDPG)

Likert items (stability, focus, interpretation of heatmap, simplicity, usefulness). Patterns across the five statements were:

- Most respondents agreed they could understand which indicators were stable or variable.
- The visual heatmap was widely seen as helpful for spotting periods of concentrated focus on specific features.
- The narrative explanation was seen as helpful in clarifying what darker and lighter areas meant (sustained versus intermittent importance).
- Participants generally rated the explanation as simple and clear to follow.
- Most agreed that stability information would be useful in judging a model's consistency.

Preferred format for understandability.

- Both equally: 8
- Visual explanation: 2
- Narrative description: 2

Open comments. Respondents observed that:

- Upper-band Bollinger indicators stood out as particularly stable in the mid-late part of the window.
- Some indicators seemed consistently weak or intermittent, matching the narrative description.

Example 5: How Market Indicators Affected Reward (DDPG)

Likert items (understanding reward drivers, reactions to signals, biggest effects, realism, connectivity). Across the five statements:

- Most participants agreed the explanation clarified which indicators improved or reduced rewards.
- They generally agreed that the text made the agent's reaction to strong versus weak signals understandable.
- Respondents reported that they could follow which stocks or features had the largest effect.
- The reasoning was typically judged realistic and easy to understand, though some respondents remained neutral.
- The explanation was seen as helping to understand how indicators are combined in the model.

Preferred format for understandability.

- Both equally: 8
- Narrative description: 2
- Visual explanation: 2

Open comments. Participants identified:

- The numerical ranking of positive and negative contributions (e.g. KO, JNJ, PG vs CRM, V) as especially helpful.
- A preference for clear sign and magnitude labelling when reading such bar charts.

Example 6: Market Regime (Dow Jones, 2009–2021)

Likert items (summary clarity, visual trend, textual clarity, understanding overall trend, usefulness of context). Responses were strongly positive:

- Almost all respondents agreed or strongly agreed that the description clearly summarised market behaviour over time.

- The chart was widely considered easy for spotting major rises and drops.
- The text was viewed as clear and simple to follow.
- Participants reported that they understood the overall trend (long bull stretches with corrections and a sharp COVID-19 downturn).
- Most agreed that this context helps interpret how trading models behave differently across regimes.

Preferred format for understandability.

- Both equally: 7
- Visual explanation: 3
- Narrative description: 2

Open comments. Comments reinforced that:

- The long bullish trajectory with a significant but brief downturn (COVID-19) was clearly visible.
- Contextualisation of regimes was helpful for framing model performance and risk.

Example 7: Feature Importance Across RL Models

Likert items (comparability, text clarity, visual comparison, common features, usefulness for model choice).

- "The comparison made it easy to see how models differ in what they focus on." Most participants agreed or strongly agreed, with very few neutral responses.
- "The text clearly described each model's main focus." Agreement was again high, with a small minority neutral.
- "The visual helped me compare features between models." Most respondents agreed, with isolated neutral or less positive ratings.
- "I could understand which features were common across all models." Agreement or strong agreement dominated.
- "I would find this type of comparison helpful when choosing a model." Most participants endorsed its usefulness.

Preferred format for understandability.

- Both equally: 6
- Visual explanation: 4
- Narrative description: 2

Most understandable or balanced model. Participants were asked which model seemed most understandable or balanced. Responses were split across PPO, DDPG, TD3, and A2C, with no single model dominating, reflecting different preferences for stability versus responsiveness.

Closing Reflection

Format felt most trustworthy overall (closing Q). When asked which format felt most trustworthy overall:

- Many participants selected “Both equally”, indicating that narratives and visuals reinforce one another.
- A non-trivial subset preferred visual-only or text-only, reflecting individual differences in preferred modality.

Overall ease of understanding. When rating overall ease of understanding, most participants selected the top two options on the scale (“Easy” or “Very easy”), indicating that the explanations were generally accessible, though a minority reported more moderate difficulty.

What would make explanations clearer or more useful? Open responses suggested:

- Slightly longer or more detailed narratives where necessary, especially for more complex visuals.
- Avoidance of jargon and greater use of clear, simplified language.
- Adding more structure, such as bullet-point summaries or tables, to highlight key takeaways.
- Tighter integration between text and visuals, explicitly referencing elements in the figure in the narrative.
- Providing more context about the model and indicators for less technical users.

Other feedback about the study. Participants described the study as interesting but noted that the material is inherently technical. Some suggested that a real-world setting might include more fundamental or macroeconomic context and potentially different asset classes or indices.

F.5 Post-Study Questionnaire

Source: Google Forms (Post-Study Questionnaire):
`https://forms.gle/1siSgr-tMEclz2SOvO6fOAcfOgAlfSATvM_1rdrG3wPg`

F.5.1 Introduction

Thank you for completing the explanation-viewing stage.
This final questionnaire asks about your experience comparing visual and text-based (narrative) explanations of AI trading decisions.
Please respond based on your overall impressions across all examples.
* **Indicates required question**

F.5.2 Overall Usability and Comprehension

Participants also rated overall usability and comprehension on the same five-point Likert scale (*Strongly disagree*, *Disagree*, *Neutral*, *Agree*, *Strongly agree*). The items were:

- The explanations were easy to read and interpret.
- The information shown was balanced — neither too detailed nor too brief.
- I could understand what influenced each trading action (Buy / Hold / Sell).
- I could follow how the AI's reasoning related to market indicators (e.g., RSI, MACD).
- The sequence of visual to text explanation felt logical and clear.

F.5.3 Comparison of Visual vs Narrative Explanations

Which format helped you understand the AI's decision most clearly?* Mark only one oval: (=:) Visual explanation (=:) Narrative explanation (=:) Both equally (=:) Neither

Which format did you find more trustworthy?* Mark only one oval: (=:) Visual explanation (=:) Narrative explanation (=:) Both (=:) Neither

Which format communicated the reasoning behind the AI's action more effectively?* Mark only one oval: (=:) Visual explanation (=:) Narrative explanation (=:) Both (=:) Neither

Which format better highlighted what features mattered most in the decision?* Mark only one oval: (=:) Visual explanation (=:) Narrative explanation (=:) Both (=:) Neither

Which format would you prefer to see in a trading assistant?* Mark only one oval: (=:) Visual only (=:) Narrative only (=:) Both combined (=:) Neither

F.5.4 Perceived Clarity, Trust, and Actionability

Participants rated their agreement with a set of statements on a five-point Likert scale with the options: *Strongly disagree*, *Disagree*, *Neutral*, *Agree*, and *Strongly agree*. The items were:

- The explanations increased my confidence in the AI's decisions.
- I trusted the AI's reasoning after reviewing the explanations.
- The explanations provided actionable insights that could inform trading choices.
- I could see why certain indicators or factors led to profit or loss outcomes.
- Seeing both explanation types together improved my overall understanding.

F.5.5 Preferences for Future Explanations

How much detail do you prefer in future AI explanations?* Mark only one oval: (=:) Minimal (1-2 key points) (=:) Moderate (=:) In-depth with evidence (=:) Other:

How should explanations adapt to different users?* Mark only one oval: (=:) Stay identical for all (=:) Adjust wording and depth to user experience (=:) Unsure (=:) Other:

Would you like to control explanation depth (e.g., toggle between summary and detailed view)?* Mark only one oval: (=:) Yes (=:) No (=:) Unsure

F.5.6 Open-Ended Questions

What did you find most useful in the explanations? _____
What, if anything, was confusing or unnecessary? _____

How could the explanations be improved to support real trading decisions? _____
Did you notice differences in how visuals and text emphasised key factors? Please describe. _____
Would you use an AI assistant that explains its decisions in this way? Why / why not? _____

F.6 Post-Study Responses

This section summarises responses to the post-study questionnaire ($N = 12$), which captured overall impressions after viewing all explanation examples.

Overall Usability and Comprehension

Ease of reading and interpretation (Post-Study.1).

- Strongly Agree: 2
- Agree: 7
- Neutral: 3

No participants disagreed that the explanations were easy to read and interpret.

Balance of information (Post-Study.2).

- Strongly Agree: 3
- Agree: 5
- Neutral: 2
- Disagree: 2

Most respondents felt that the explanations were neither too detailed nor too brief, though a small minority found them imbalanced.

Understanding what influenced each trading action (Post-Study.3).

- Strongly Agree: 3
- Agree: 8
- Neutral: 1

Linking AI reasoning to indicators (Post-Study.4).

- Strongly Agree: 4
- Agree: 7
- Neutral: 1

Sequence from visual to text (Post-Study.5).

- Strongly Agree: 5
- Agree: 5
- Neutral: 2

The visual-to-text sequence was widely perceived as logical and clear.

Comparison of Visual vs Narrative Explanations

Format that most clearly conveyed the AI's decision (Post-Study.6).

- Both equally: 7
- Narrative explanation: 3
- Visual explanation: 2
- Neither: 0

Most trustworthy format (Post-Study.7).

- Both: 7
- Narrative explanation: 3
- Visual explanation: 2
- Neither: 0

Format that communicated reasoning most effectively (Post-Study.8).

- Narrative explanation: 5
- Both: 4
- Visual explanation: 3
- Neither: 0

Format that better highlighted important features (Post-Study.9).

- Visual explanation: 5
- Narrative explanation: 4
- Both: 3

Preferred format in a trading assistant (Post-Study.10).

- Both combined: 11
- Narrative only: 1
- Visual only: 0
- Neither: 0

There was a very strong preference for combined visual and narrative explanations.

Perceived Clarity, Trust, and Actionability

Items in this section were rated on a numerical 1–5 scale, where higher scores indicate stronger agreement.

Increased confidence in the AI's decisions (Post-Study.11). Most respondents selected 4 or 5, indicating that explanations tended to increase their confidence, with only a small number selecting 3 (neutral).

Trust in the AI's reasoning (Post-Study.12). Scores again clustered around 4 and 5, with occasional neutral responses and very few lower scores.

Actionable insights for trading choices (Post-Study.13). The majority gave ratings of 4 or 5, suggesting that participants felt they could derive practical guidance from the explanations, though one respondent adopted a more moderate (3) stance.

Seeing why indicators led to profit or loss (Post-Study.14).

- Rating 4: 7
- Rating 5: 3
- Rating 3: 2

Benefit of seeing both explanation types together (Post-Study.15).

- Rating 5: 9
- Rating 4: 2
- Rating 3: 1

Participants overwhelmingly felt that the combination of visuals and narratives improved their overall understanding.

Preferences for Future Explanations

Preferred level of detail (Post-Study.16).

- In-depth with evidence: majority of respondents
- Moderate: some respondents
- Minimal: none

Participants generally preferred detailed explanations that include supporting evidence.

Adaptation to different users (Post-Study.17).

- Adjust wording and depth to user experience: clear majority
- Stay identical for all: few
- Unsure / Other: very few

Control over explanation depth (Post-Study.18).

- Yes (would like to control depth, e.g. toggle summary vs detail): majority
- Unsure: some
- No: very few

Open-Ended Responses

Most useful aspects of the explanations (Post-Study.19). Participants highlighted:

- The combination of visuals and narratives, especially when the text explicitly interpreted the charts.
- Clear identification of which indicators and stocks contributed most to returns or losses.
- Structured breakdowns (e.g. by regime, feature, or reward component) that made complex behaviour more digestible.

What was confusing or unnecessary (Post-Study.20). Points raised included:

- Some terminology and indicator names felt too technical for less experienced users.
- Certain visuals were dense and required careful reading of the narrative to interpret.
- Occasional repetition between text and charts could be streamlined.

Improvements for real trading decisions (Post-Study.21). Suggested enhancements:

- More explicit and concise bullet-point summaries highlighting key reasons for each decision.
- Clearer links to risk metrics and real-world scenarios.
- Additional context on timeframes, asset characteristics, and macro events.
- Customisable views that allow users to focus on specific indicators or time windows.

Differences between visuals and text (Post-Study.22). Respondents observed that:

- Visuals were better at highlighting patterns (e.g. trends, spikes, stability) and relative magnitudes.
- Narratives were better at explaining causal reasoning, translating technical details into plain language.
- In several cases, the two modalities emphasised similar factors but in complementary ways, with the visual making the structure salient and the text clarifying interpretation.

Willingness to use an AI assistant of this kind (Post-Study.23). Participants gave a range of views, including:

- Several would use such an assistant, seeing it as a way to make trading more approachable and to supplement their own analysis.
- Some would use it cautiously, only as an additional input and not as the sole decision-maker, emphasising the need for independent research and fundamental analysis.
- A few were unsure or reluctant, noting that explanations may still omit important real-world factors and that stock prices can be influenced by elements beyond technical indicators.
- Others indicated that they would consider using such tools to reduce research time or to explore new trading opportunities, provided that reliability and data quality are demonstrated.