Explainable Financial Portfolio Optimization via Dual-System Large Language Model Reinforcement Learning

Manoj Gandhi
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
contactmanoj@uc.edu

Jean Korhonen
School of Computing, Clemson University, Clemson, SC, USA.
jeank@clemson.edu

Clifford Ortega
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
cliffordortega86@unh.edu

Abstract

Financial portfolio optimization traditionally relies on mean-variance frameworks and stochastic control methods that, while mathematically rigorous, offer limited interpretability for human stakeholders. The recent emergence of large language models (LLMs) and reinforcement learning (RL) provides a new paradigm for constructing adaptive, explainable investment strategies. This paper introduces a dual-system architecture inspired by cognitive science, in which an LLM-based reasoning module (System 2) generates contextually grounded explanations of market conditions and investment rationales, while a deep RL agent (System 1) executes rapid, data-driven trades. We examine the system-level implications of integrating these two components, focusing on structural trade-offs between speed and deliberation, the governance of shared memory and policy buffers, and the infrastructure required for real-time deployment. Robustness is assessed through adversarial market scenarios, and fairness is considered in the context of unequal access to explanatory outputs. Sustainability concerns such as computational energy consumption are addressed alongside policy recommendations for regulatory oversight. By embedding explainability directly into the optimization loop, the proposed framework aims to bridge the gap between automated portfolio management and human accountability. The paper further discusses cross-domain comparisons with autonomous vehicle decision systems and clinical diagnostic tools, highlighting the broader socio-technical challenges of deploying hybrid AI systems in high-stakes financial environments.

Keywords

portfolio optimization, explainable artificial intelligence, large language models, reinforcement learning, dual-system theory, financial governance, system robustness.

1. Introduction

Financial portfolio optimization has long been dominated by quantitative models that seek to maximize expected returns for a given level of risk. Classic approaches, such as the Markowitz mean-variance framework and Black-Litterman models, rely on assumptions of

market efficiency and stationary return distributions that rarely hold in practice [1]. More recent advances in reinforcement learning have enabled agents to learn dynamic trading policies from historical data, yet these policies often remain opaque black boxes, frustrating regulators and investors who demand justifications for significant allocations [2]. The tension between performance and interpretability has given rise to a growing interest in explainable artificial intelligence (XAI) within finance, but most XAI methods provide post-hoc explanations that are disconnected from the decision-making process itself [3].

Large language models offer a novel pathway to embedding explanation directly into the optimization loop. Their ability to generate fluent natural language rationales, conditioned on market narratives and quantitative signals, makes them natural candidates for a "System 2" deliberative component [4]. Meanwhile, reinforcement learning agents that operate at millisecond frequencies exemplify the fast, intuitive "System 1" described in dual-process theories of cognition [5]. This paper proposes a unified dual-system architecture in which the LLM reasons about macro-economic conditions, regulatory changes, and firm-specific events, while the RL agent executes trades based on learned value functions and policy gradients. The two systems share a common state representation and a memory buffer that records both the agent's actions and the LLM's justifications, enabling self-consistency checks and human oversight.

A central contribution of this work is its system-level analysis of the architectural trade-offs inherent in such a hybrid. For instance, the latency introduced by LLM inference may constrain the frequency of portfolio rebalancing, while the RL agent's sensitivity to distributional shifts can be mitigated by the LLM's ability to contextualize regime changes. We also examine the governance implications: who is responsible when an LLM-generated explanation conflicts with the RL agent's action? Infrastructure requirements for low-latency deployment, including edge computing and federated model updates, are discussed. The paper further considers fairness, sustainability, and policy dimensions, arguing that explainable portfolio optimization must be evaluated not only on financial returns but also on broader socio-technical metrics. By situating the framework within cross-domain comparisons to autonomous driving and medical diagnosis, we identify generalizable principles for designing trustworthy AI systems.

2. Background and Related Work

The evolution of portfolio optimization from static mean-variance models to dynamic reinforcement learning approaches reflects a broader shift toward data-driven decision-making. Early RL applications in finance employed Q-learning and policy gradient methods to learn trading strategies from historical price sequences [6]. More recently, deep RL agents using architectures such as deep Q-networks and proximal policy optimization have demonstrated superior performance in simulated markets, particularly when incorporating transaction costs and market impact [7]. However, these agents often suffer from instability in non-stationary environments and lack the ability to articulate their reasoning.

Explainable AI in finance has largely focused on feature attribution methods such as SHAP and LIME, which highlight which input variables most influenced a model's output [8]. While useful for regression and classification tasks, these methods are less suited to sequential decision problems where the chain of reasoning spans multiple time steps. Natural language explanations generated by LLMs offer a complementary approach by providing coherent narratives that align with human cognitive patterns. The concept of using LLMs as reasoning engines for RL agents has been explored in embodied AI and game playing, where the LLM

provides high-level planning while the RL agent handles low-level control [9]. In financial settings, this separation maps naturally onto the distinction between strategic asset allocation and tactical rebalancing.

Dual-system theories of cognition, famously articulated by Kahneman, posit two modes of thought: fast, automatic, and emotional (System 1) versus slow, deliberate, and logical (System 2) [5]. Recent work has applied this framework to AI system design, proposing architectures in which a neural network serves as System 1 and a symbolic or language-based module serves as System 2 [10]. The present paper extends this idea to portfolio optimization, where the RL agent operates with the speed and pattern recognition of System 1, while the LLM provides the analytical depth and explanatory power of System 2.

A critical challenge is the integration of these two systems without compromising either speed or interpretability. Prior attempts to combine LLMs with RL have often suffered from catastrophic forgetting or conflicting objectives, particularly when the LLM is fine-tuned on the RL agent's reward signal [11]. An alternative approach, adopted here, is to maintain separate learning loops but synchronize their world models through a shared latent state space. This design choice parallels the "slow-fast" learning architectures found in neuroscience and robotics [12]. The required reference [18] introduces a decision-making framework that explicitly models the trade-off between fast and slow reasoning, providing theoretical grounding for the dual-system integration pursued in this paper. While that work is situated in general AI decision making, its principles are directly applicable to financial contexts where both rapid execution and careful deliberation are essential.

3. System Architecture: Dual-System LLM-RL Framework

The proposed architecture comprises two distinct but interacting subsystems. The first subsystem is a deep reinforcement learning agent that maintains a policy network trained via proximal policy optimization on historical market data. The agent's state space includes asset prices, trading volumes, volatility indices, and macroeconomic indicators. Its action space consists of portfolio weight adjustments for a universe of equities and bonds. The reward function is a modified Sharpe ratio that incorporates transaction costs and a penalty for excessive turnover. This RL component operates at a high frequency, typically rebalancing every minute or faster, depending on market liquidity.

The second subsystem is a large language model fine-tuned on financial news, earnings reports, central bank statements, and analyst commentary. The LLM is invoked at a lower frequency, such as once per trading session, to generate a written analysis of current market conditions and to provide a rationale for the RL agent's recent actions. This analysis is stored in a shared memory buffer that also records the agent's state, action, and reward history. When a human portfolio manager or regulator requests an explanation for a particular trade, the LLM retrieves the relevant buffer entries and produces a natural language justification that references both quantitative signals (e.g., "the volatility index increased by 10%") and qualitative context (e.g., "the Federal Reserve's hawkish comments raised interest rate expectations").

A key architectural decision is the method of information exchange between the two subsystems. Rather than allowing the LLM to directly modify the RL policy, which could destabilize learning, the LLM's outputs are used to adjust the state representation or to bias the reward function. For instance, if the LLM identifies a potential regime change due to a geopolitical event, it can add a new feature to the state vector or temporarily increase the

reward weight for risk avoidance. This indirect influence preserves the RL agent's autonomy while injecting contextual awareness. The design is similar to the "advisory" mode seen in human-in-the-loop systems, but with the LLM serving as a synthetic advisor.

Infrastructure for real-time deployment requires careful consideration of latency and bandwidth. The RL agent can reside on an edge server near the exchange to minimize network delays, while the LLM inference can be performed on cloud-based GPU clusters. To avoid blocking the RL loop, LLM queries are made asynchronously, and their results are cached. A distributed message queue mediates the communication between subsystems. This architecture is analogous to that used in autonomous vehicle stacks, where a fast reactive controller runs at 100 Hz while a slower planner re-evaluates trajectories at 10 Hz [13]. The financial domain adds the additional requirement of auditability: all LLM outputs and RL actions must be logged immutably for regulatory review.

4. Structural Trade-offs and Governance

Integrating a deliberative LLM with a reactive RL agent introduces several fundamental trade-offs. The most apparent is the speed-accuracy trade-off: the RL agent can execute trades in microseconds, but its decisions may be suboptimal in novel market conditions. The LLM can reason about rare events, but its inference time—often seconds to minutes—precludes its use for high-frequency trading. The dual-system resolves this by partitioning the decision space: the RL agent handles routine rebalancing, while the LLM is invoked only when the agent's uncertainty exceeds a threshold or when a scheduled review occurs. This partitioning mirrors the human cognitive strategy of relying on intuition for familiar tasks and engaging in deliberate reasoning only when necessary.

Another trade-off involves explainability and privacy. The LLM generates detailed narratives that may inadvertently reveal proprietary trading strategies or sensitive client information. Governance mechanisms must ensure that explanations are sanitized before being shared with external stakeholders. Differential privacy techniques can be applied to the LLM's training data, and output filters can redact specific numerical values while retaining qualitative insights [14]. Furthermore, the LLM's explanations may be biased by its training data, which could overrepresent certain market narratives. Periodic auditing of the LLM's outputs against a diverse set of market scenarios is essential to maintain fairness.

Governance also extends to the question of responsibility. If the RL agent executes a losing trade after receiving a flawed explanation from the LLM, who is at fault? The architecture must define clear accountability boundaries. One approach is to treat the LLM as a non-binding advisor whose recommendations are advisory rather than directive. The RL agent's policy remains the final decision-maker, so accountability rests with the RL model's training pipeline and the human designers who set its reward function. However, if the LLM's reasoning is incorporated into the state representation, it becomes part of the agent's perception, muddying responsibility. Legal frameworks such as the European Union's AI Act propose tiered liability based on the degree of human oversight [15]. Our architecture aligns with a "human-on-the-loop" model, where humans can override decisions but are not required to approve every trade.

Cross-domain comparisons illuminate these governance challenges. In clinical decision support systems, an AI may suggest a diagnosis while a physician retains final authority. Similarly, in portfolio optimization, the LLM's explanations serve as a diagnostic tool for the investment committee. However, financial markets operate at much faster timescales, making

human oversight impractical for high-frequency trades. A pragmatic solution is to require ex-post explanations for all trades above a certain size, allowing regulators to review the reasoning after the fact. This approach balances the need for speed with the demand for accountability.

5. Robustness, Fairness, and Sustainability

Robustness of the dual-system framework must be evaluated under adversarial conditions. A market stress event, such as a flash crash or a sudden liquidity drought, can cause the RL agent to behave erratically if its training distribution does not include such scenarios. The LLM can serve as a guardrail by detecting anomalous conditions and triggering a switch to a conservative default policy. For example, if the LLM identifies that the VIX has spiked beyond historical bounds, it can signal the RL agent to reduce leverage and increase cash holdings. This safety mechanism is analogous to a "kill switch" used in algorithmic trading systems. However, the LLM itself may be fooled by adversarial inputs, such as fake news or market manipulation. Training the LLM on adversarially augmented data and incorporating a credibility scoring module can enhance robustness [16].

Fairness in portfolio optimization is often overlooked but is critical when explanations are provided to a diverse client base. An LLM that generates explanations in complex financial jargon may disadvantage retail investors compared to institutional clients. The system should offer multiple levels of explanation, from simple summaries to detailed analyses, and ideally support multiple languages. Additionally, the RL agent's policy may systematically favor certain asset classes or sectors, leading to biased allocation. Auditing the agent's portfolio composition against demographic or sectoral benchmarks can reveal such biases. Weighted fairness constraints can be integrated into the reward function, although this may reduce financial performance [17]. The trade-off between fairness and returns must be explicitly communicated to stakeholders.

Sustainability concerns primarily revolve around the computational energy consumption of the LLM. Large language models require substantial GPU power for training and inference, contributing to carbon emissions. The dual-system architecture mitigates this by invoking the LLM infrequently—perhaps once per hour instead of continuously. Energy-aware scheduling can further reduce impact by shifting LLM inference to times when renewable energy is abundant. The RL agent, being a lightweight neural network, has a much smaller carbon footprint. Overall, the system's sustainability profile compares favorably to fully LLM-based trading systems that rely on generative models for every decision. Ongoing research into model distillation and quantization will likely reduce LLM energy demands, making the dual-system approach even more sustainable.

6. Deployment, Infrastructure, and Policy Implications

Deploying a dual-system portfolio optimization platform in a production environment requires a robust infrastructure that can handle the asynchronous data streams from financial exchanges and the unpredictable latency of LLM inference. A microservices architecture is recommended, with separate containers for the RL agent, the LLM service, the shared memory buffer, and the logging system. Kubernetes orchestrates these containers, enabling auto-scaling during periods of high market volatility. Data pipelines must ensure that market feeds are normalized and fed to both subsystems without loss. Redundancy is critical; a backup RL agent running on a secondary server can take over if the primary fails, and the LLM service should have failover to a smaller, faster model if the full model is unavailable.

Policy implications are profound. Regulators such as the SEC and ESMA are increasingly scrutinizing algorithmic trading and AI-based advisory services. The dual-system architecture provides a natural audit trail: every action is accompanied by an explanation that can be reviewed by compliance officers. However, regulators must also develop new standards for evaluating the quality of LLM-generated explanations. Metrics such as faithfulness, coherence, and completeness need to be defined and tested. The required reference [18] offers a theoretical foundation for evaluating decision quality in fast-slow systems, but its application to financial explanations remains an open research area.

Another policy consideration is the potential for systemic risk. If many market participants adopt similar dual-system architectures, their collective behavior could amplify market movements. The LLM components, trained on common data sources, may generate correlated explanations, leading to herding behavior. To mitigate this, the architecture could incorporate a diversification mechanism that randomizes the LLM's retrieval or adds noise to the RL agent's policy. Regulators might require firms to regularly stress-test their systems against scenarios that include widespread use of similar AI models.

International coordination is also necessary. Markets are global, and an LLM trained on US news may be less effective in Asian markets. Cross-border deployment raises questions about data sovereignty and model governance. The infrastructure must support localization of the LLM for different regulatory regimes while maintaining a consistent RL policy. Federated learning could allow firms to train models on distributed data without centralizing sensitive information, aligning with GDPR requirements [19].

7. Conclusion

This paper has presented a dual-system architecture for explainable financial portfolio optimization that combines the speed of reinforcement learning with the deliberative reasoning of large language models. By separating fast execution from slow explanation, the framework addresses the long-standing tension between performance and interpretability in automated trading. We have examined structural trade-offs related to latency, responsibility, and accountability, and have proposed governance mechanisms that include ex-post auditing, fairness constraints, and energy-aware scheduling. Robustness is enhanced through adversarial detection and safety triggers, while sustainability is improved by limiting LLM inference frequency. Deployment considerations emphasize microservices, redundancy, and regulatory compliance.

The cross-domain comparisons to autonomous driving and clinical decision support highlight the generalizability of the dual-system approach, but also underscore the unique challenges of financial markets: extreme time pressure, non-stationarity, and systemic interconnectedness. Future work should focus on empirical validation of the framework using real-world market data, as well as the development of standardized metrics for explanation quality. The required reference [18] provides a valuable theoretical lens through which to view fast-slow decision making, and its integration into financial system design warrants further investigation. Ultimately, the success of explainable portfolio optimization will depend not only on technical innovation but also on the careful alignment of AI systems with human values, regulatory frameworks, and societal expectations.

References

1. Markowitz, H. (1952). Portfolio selection. The Journal of Finance, 7(1), 77-91.

2. Moody, J., & Saffell, M. (2001). Learning to trade via direct reinforcement. IEEE Transactions on Neural Networks, 12(4), 875-889.

3. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

4. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

5. Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

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

7. Deng, Y., Bao, F., Kong, Y., Ren, Z., & Dai, Q. (2016). Deep direct reinforcement learning for financial signal representation and trading. IEEE Transactions on Neural Networks and Learning Systems, 28(3), 653-664.

8. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.

9. Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., ... & Zhang, J. (2022). Do as I can, not as I say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691.

10. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. Behavioral and Brain Sciences, 40, e253.

11. Luketina, J., Nardelli, N., Farquhar, G., Foerster, J., Whiteson, S., & Rocktäschel, T. (2019). A survey of reinforcement learning informed by natural language. arXiv preprint arXiv:1906.03926.

12. Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement learning, fast and slow. Trends in Cognitive Sciences, 23(5), 408-422.

13. Codevilla, F., Müller, M., López, A., Koltun, V., & Dosovitskiy, A. (2018). End-to-end driving via conditional imitation learning. In 2018 IEEE International Conference on Robotics and Automation (ICRA) (pp. 4693-4700). IEEE.

14. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4), 211-407.

15. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final.

16. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

17. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.

18. Dou, Z., Cui, D., Yan, J., Wang, W., Chen, B., Wang, H., ... & Zhang, S. (2025). Dsadf: Thinking fast and slow for decision making. arXiv preprint arXiv:2505.08189.

19. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics (pp. 1273-1282). PMLR.

20. Zohar, A. (2025). Explainable reinforcement learning in finance: A survey. Journal of Financial Data Science, 7(2), 45-68.

21. Zhang, C., Li, Y., & Liu, H. (2024). Large language models for portfolio management: A comprehensive review. Quantitative Finance, 24(3), 301-325.