Research on Decision Biases of Generative-AI-Driven Robo-Advisory and Investor Protection Mechanisms

Xiuling Zhou[1,a], Shuang Diao[2,b,*]

[1]Trinity Collge Dublin, Dublin, Ireland

[2]Guangzhou University of Applied Sciences, Zhaoqing, China

[a]zhoux1@tcd.ie, [b]1012616950@163.com

Keywords: Generative AI; Robo-advisory; Decision bias; Causal robustness; Investor protection; Explainability; Reinforcement learning; Robust optimization; Suitability; Compliance

Abstract: This paper proposes a governance-ready framework for identifying and mitigating decision biases introduced by generative AI in robo-advisory, integrating explainability constraints, causal robustness, and compliance executability within a layered architecture spanning data, features, forecasting, optimization, and risk governance. Counterfactual evaluation and constrained decision optimization coordinate return, risk, and compliance objectives, and a semi-synthetic, empirically calibrated study indicates that, at equal risk budgets, generative-AI-enhanced advisory improves allocation diversity, communication personalization, and scenario responsiveness while remaining susceptible to prompt injection, hallucination, overconfidence, and interaction-induced risk-preference drift. An integrated protection bundle composed of decision evidence cards, policy corridors, robust optimization, and human-in-the-loop governance reduces erroneous extrapolation and strategy whipsawing, enhances suitability alignment, and strengthens tail-risk control under stress, offering a pragmatic deployment roadmap.

1. Introduction

1.1 Background and Significance

In recent years, generative artificial intelligence has begun to permeate financial advisory workflows, extending its influence from investor education and interactive Q&A services to more complex tasks such as risk profiling, portfolio allocation guidance, and narrative scenario generation [1]. This shift reflects the increasing demand for personalized, scalable, and accessible advisory services. However, the deployment of generative models introduces non-trivial challenges. Generative uncertainty, distributional shifts between training and serving environments, and dynamic feedback loops created by client interaction all give rise to novel decision biases and operational risks. Traditional safeguards designed for deterministic or statistical models are insufficient to address these challenges. To ensure investor protection and regulatory compliance, a "protection-first" system design becomes essential. Such a design must convert language-driven influences into verifiable evidence, stabilize recommendations through policy corridors that prevent abrupt strategy shifts, and implement human oversight for high-impact or low-confidence cases. In this way, generative AI can be harnessed responsibly while upholding obligations of suitability, explainability, auditability, and accountability.

1.2 Research Status at Home and Abroad

Research on robo-advisory has steadily advanced, with established paradigms based on mean–variance optimization, risk parity approaches, and goal-based investing frameworks. These have increasingly been enhanced through Bayesian learning, reinforcement learning, and other machine learning techniques, providing more adaptive allocation strategies. Internationally, studies have begun to explore the potential of generative AI, yet most applications remain limited to content automation, document drafting, or compliance triage, rather than core decision-making processes. Within the domestic research landscape, systematic integration of generative models into advisory

  DOI: 10.25236/gemmsd.2025.087

pipelines is still nascent. Several key gaps remain: the measurement of interaction-induced preference drift, the quantification of prompt injection and hallucination effects on allocation stability, the operationalization of verifiable explanations and audit trails, and the engineering of robust optimization methods combined with human–AI co-governance in real-time loops. Addressing these gaps is crucial for transforming generative AI from a peripheral tool into a core enabler of trustworthy, regulation-aligned robo-advisory. This paper responds to these needs by proposing an implementable and evaluation-ready framework, designed to meet both practical market requirements and stringent regulatory review.

2. Theoretical Framework and Problem Definition

2.1 Objectives and Constraints

The advisory framework is designed with the overarching objective of maximizing the probability of achieving client-specific financial goals while carefully constraining downside risks. This requires a dual focus: on the one hand, ensuring that portfolios remain within defined risk budgets; on the other, embedding ethical and regulatory constraints such as suitability, disclosure, fairness, and privacy. Importantly, the system acknowledges that language and framing may exert significant behavioral influence on investors, potentially biasing decisions. To safeguard against misuse, "hard guardrails" are established. These include enforcing smooth strategy transitions rather than abrupt shifts, mandating transparent disclosure of potential conflicts of interest, and requiring auditable outputs that allow both internal compliance and external regulators to review decision processes. Together, these constraints ensure that optimization results are not only mathematically robust but also practically enforceable in real-world financial advisory settings.

2.2 Bias Taxonomy and Propagation

Bias is categorized across multiple dimensions. At the model level, vulnerabilities arise from corpus shift, fragile extrapolation over long financial sequences, underrepresentation of rare events, and tendencies toward hallucination or unjustified overconfidence [2]. At the interaction level, risks emerge from adversarial prompt injection, suggestive phrasing that induces biased behavior, and emotional contagion through tone. Data-level biases include distortions caused by backtest timing choices and survivorship effects. At the execution stage, neglected transaction costs, market slippage, and overly frequent rebalancing can erode returns. These biases are not isolated: they propagate in a closed feedback loop where generated recommendations alter inferred client preferences, which then guide optimization and execution. The resulting outcomes feed back into model retraining and template updates, potentially amplifying existing distortions or, conversely, enabling corrective adjustments when biases are detected and mitigated.

2.3 System Architecture

To address these challenges, a five-layer architecture is proposed. The data layer consolidates diverse inputs, including KYC profiles, transaction and holdings data, execution costs, interaction records, complaints, and external market and event streams. The feature layer extracts and encodes client-specific dimensions such as risk capacity, investment horizon, cash flow cadence, exposure to volatility and tail risks, as well as transaction cost and tax footprints. The forecasting layer generates multi-perspective return–risk projections, stress testing under macro and market shocks, and counterfactual scenarios to explore alternative strategies. Building on this, the optimization layer performs robust multi-period allocation and rebalancing under explicit execution and liquidity constraints. Finally, the risk and compliance layer ensures accountability through explanation modules, evidence cards for each decision, corridor-based risk thresholds, anomaly escalation protocols, and immutable audit trails that enable both human review and regulatory oversight. This layered structure provides resilience, transparency, and adaptability, balancing advanced analytics with enforceable compliance safeguards.

3. Data and Feature Engineering

3.1 Sources and Governance

The data foundation of the advisory system draws on both internal and external sources. Internally, onboarding processes provide KYC and suitability records, complemented by structured risk questionnaires, detailed transaction and position histories, order execution logs, fee and commission data, as well as interaction summaries and records of complaints or client churn [3]. Externally, the system ingests a broad set of market signals, including asset prices, corporate fundamentals, options-implied volatility and skew, interest rate and credit spread data, ESG controversies, news sentiment, and macroeconomic calendars. To ensure analytical consistency, master-data management and temporal harmonization are applied at daily or weekly cadences, creating a synchronized view of client and market states. Governance mechanisms are embedded through privacy-preserving computation frameworks, alongside immutable logs that record every prompt-template version, model snapshot, active constraint, and decision context. This ensures end-to-end traceability, facilitates rollback when errors or anomalies occur, and aligns data handling with regulatory and ethical standards.

3.2 Feature Construction

Features are engineered across client, market, and interaction dimensions. On the client side, measures include risk budgets expressed through Value-at-Risk (VaR) and Expected Shortfall (ES), proxies for loss aversion based on drawdown sensitivity, liquidity preferences derived from transaction history, funding gaps relative to financial goals, investment horizons, and indicators of behavioral stability inferred from trading patterns [4]. Market-side features incorporate cross-asset volatility and correlation structures, jump risks and skewness in returns, time-varying risk premia, and liquidity or impact cost curves that reflect execution constraints. Interaction-side features capture qualitative dimensions such as sentiment polarity, client comprehension confidence, prompt stability across sessions, and flags for contexts vulnerable to manipulation or misinterpretation [5]. To avoid information leakage, strict temporal lags and event-time stamping are applied, while quantile capping and stable scaling are used to enhance robustness against extreme outliers.

3.3 Labeling and Validation Protocol

Supervised targets extend beyond simple returns, covering multi-horizon distribution forecasts, maximum drawdowns and tail-loss measures, goal-attainment probabilities triggered by specific milestones, and behavioral outcomes such as client retention or likelihood of complaints. Validation follows a rolling-origin design, with blocking by macroeconomic regime and client segment to prevent overfitting. Performance is evaluated using multiple statistical criteria—WAPE or RMSE for forecast accuracy, CRPS for distribution calibration, Brier scores and AUC/PR-AUC for classification, and coverage rates for prediction intervals. Beyond predictive accuracy, operational decision metrics are tracked [6]. These include shifts along the Pareto frontier of risk–return trade-offs, drawdown depth and recovery times, portfolio turnover and associated costs, compliance hit rates, consistency of generated explanations, and acceptance levels in human review processes. Such multi-layer validation ensures both quantitative rigor and operational reliability in live advisory contexts.

4. Modeling Methodology

4.1 Generative AI with Retrieval Augmentation

To ensure accuracy, transparency, and regulatory compliance, the advisory system employs retrieval-augmented generation (RAG) constrained to a carefully curated and continuously updated knowledge base [7]. This controlled corpus includes product specifications, risk policies, fee and tax schedules, and authoritative interpretations of applicable regulations. Instead of free-form text generation, the model retrieves and grounds responses in these verified documents. A fact-verification module cross-checks every citation, and in cases where evidence is insufficient or

the query falls outside scope, the system abstains from producing speculative recommendations. Outputs follow a standardized five-part structure—recommendation, rationale, risks, costs, and alternatives—emitted through function-calling templates. This format facilitates systematic review, comparability across cases, and downstream auditability, thus aligning generative AI with both operational and supervisory requirements.

4.2 Causally Robust Bias Identification

Bias detection is not limited to descriptive error analysis but grounded in causal inference. Recommendations are modeled as conditional functions of client state variables (e.g., risk capacity, liquidity needs) and contemporaneous market conditions [8]. To estimate the effect of recommendation changes, doubly robust estimators and causal forests are deployed, effectively addressing selection bias and time-varying confounding. Diagnostics enhance credibility: overlap checks confirm sufficient support across treatment conditions, placebo tests in pre-periods guard against spurious correlations, and sensitivity analyses measure robustness to exogenous shocks such as changes in fee schedules or regulatory calendar events. Event-study methods further track pre- and post-intervention trajectories, clarifying dynamic causal effects. On the interaction side, controlled A/B evaluations probe how variations in prompt phrasing influence client preference formation, allowing quantification of "preference drift." Adversarial testing ensures resilience against manipulative or injected prompts, a critical component of safeguarding advisory integrity.

4.3 Robust Allocation and Execution

Portfolio optimization is framed as a multi-period problem of maximizing expected utility or goal-attainment probability, subject to multiple real-world constraints. Risk is managed using Conditional Value-at-Risk (CVaR), while turnover, transaction taxes, and allocation smoothness are explicitly bounded. Bayesian and distributionally robust optimization formulations capture parameter uncertainty in expected returns, covariances, and execution cost curves [9]. To prevent excessive portfolio volatility, policy corridors define upper and lower bounds for both asset weights and aggregate risk budgets. These constraints allow for week-over-week adjustments within limits, and in cases of temporary shocks, deviations are permitted but decay automatically back toward the baseline trajectory. Execution processes translate advisory outputs into tradeable orders, incorporating market impact and slippage modeling. Orders may be staged across time to minimize costs, with all execution steps recorded in audit trails at the order level, ensuring that recommendations remain explainable, reproducible, and compliant with best-execution obligations.

4.4 Investor Protection and Compliance Governance

Investor protection principles are integrated as first-order design constraints [10]. Suitability guardrails link recommendation tiers to KYC profiles and empirically demonstrated client tolerance levels, with any advice exceeding bounds escalated for secondary confirmation and accompanied by a cooling-off mechanism to prevent impulsive acceptance. Each piece of advice is accompanied by an "evidence card" detailing decision drivers, data sources, confidence intervals, alternative strategies, rejection reasons, disclosures of risks and costs, and a record of triggered constraints. Defenses against hallucinations and prompt manipulation include retrieval whitelists, static prompt linting, runtime guardrails, and abstention thresholds to block unreliable outputs. Human-in-the-loop workflows handle high-impact or low-confidence cases, providing an additional safeguard where algorithmic recommendations may be insufficient. Beyond individual interactions, fairness and compliance are reinforced through disparate-impact analysis, ensuring that recommendations do not inadvertently disadvantage protected groups. Restricted causal pathways prevent the inappropriate use of sensitive attributes, while immutable logs of all advisory steps enable independent audit and regulatory verification. Collectively, these measures establish a governance framework where technological innovation coexists with rigorous investor protection.

5. Empirical Evaluation and Case Study

5.1 Experimental Setup

The experimental evaluation is conducted in a semi-synthetic environment that blends authentic market data with controlled simulations. Historical multi-asset return series are integrated with cross-sectional factor data to capture realistic co-movement structures. Liquidity and transaction cost models are incorporated to reflect market frictions, while exogenous shocks—including policy windows, central bank announcements, and volatility spikes—are introduced to test robustness under stress [11]. Three advisory paradigms are compared: a rules-based baseline reflecting traditional allocation heuristics; a discriminative robo-advisory system built on gradient boosting models and Bayesian mean–variance optimization; and an advanced generative-AI-enhanced robo that integrates retrieval augmentation, distributionally robust optimization, and compliance guardrails. Beyond static evaluation, the system is tested under interactive scenarios, both with and without adversarially designed prompts, to assess vulnerability to manipulation. All experiments are run across rolling horizons of 12 to 52 weeks, allowing dynamic assessment of cumulative performance, stability, and compliance adherence.

5.2 Results

The results indicate that the generative-AI-enhanced framework delivers superior outcomes across multiple dimensions. For equalized Conditional Value-at-Risk (CVaR) budgets, it achieves higher annualized excess returns and greater probabilities of meeting client-specific goals. Drawdown recovery times are shortened, and turnover is stabilized within corridor bounds, reducing trading costs. Stress-test periods reveal further advantages: narrative-based scenario explanations and corridor constraints prevent abrupt portfolio shifts, significantly lowering week-to-week weight variance compared to discriminative baselines. Interaction robustness also improves. Without defenses, adversarial prompts induce notable deviations from intended strategy and heighten complaint risks. By contrast, retrieval whitelists, evidence cards, and structured disclosure protocols markedly reduce such deviations. Additional benefits are observed in compliance and client experience: explanation consistency improves, reviewer acceptance rates increase, suitability violations decline, and proxies for client retention and satisfaction show measurable gains.

5.3 Ablation and Sensitivity

Ablation studies confirm the contribution of each architectural element. Removing retrieval augmentation and fact-verification modules leads to an uptick in hallucinations and spurious extrapolations, which translate into deeper drawdowns and higher complaint likelihood. Eliminating policy corridors results in elevated portfolio volatility, excessive turnover, and increased trading costs. Weakening causal debiasing mechanisms encourages momentum-chasing behavior, undermining robustness during exogenous shocks. Sensitivity analyses explore parameter tuning: expanding robustness radii and tightening chance-constraint thresholds enhance resilience but may reduce achievable returns, whereas looser specifications boost short-term gains at the expense of higher stress vulnerability. These findings highlight a clear trade-off between performance and stability, offering a calibrated design space for policymakers and practitioners to balance investor protection with market competitiveness.

5.4 Managerial Insights

From a managerial standpoint, the experiments yield several practical lessons. Embedding generative AI into a structured Sales and Operations Planning (S&OP) cadence—supported by tri-level dashboards for performance, risk, and compliance—creates transparent oversight and actionable escalation thresholds [12]. Prioritization should focus on foundational elements such as data quality, execution-cost modeling, and curated knowledge-base governance before attempting to integrate broad sentiment data from unverified sources. Asset-class context matters: flexible corridor mechanisms are well suited for liquid, high-substitutability assets, whereas illiquid or

concentrated exposures may be better managed through non-price levers such as allocation limits or liquidity buffers. Standardized evidence cards, enriched with alternative options and rejection reasoning, reduce client misunderstanding and mitigate behavioral biases. Finally, robust model change management, combined with retrospective audits anchored by immutable logs, enhances accountability, builds trust, and eases regulatory inspection. Collectively, these insights demonstrate that careful system design can harness generative AI's potential while upholding fiduciary and compliance obligations.

6. Conclusion

This study develops and evaluates a bias-aware, protection-first framework for generative-AI-driven robo-advisory. The proposed approach addresses the dual challenge of leveraging the personalization power of large language models while safeguarding against generative uncertainty, interaction-induced bias, and operational risk. By constraining advisory generation through retrieval and fact verification, the system ensures that recommendations remain grounded in authoritative sources. Causally robust estimators provide a principled basis for identifying and mitigating biases, while distributionally robust optimization strengthens portfolio allocation against parameter uncertainty and rare-event shocks. Policy corridors, combined with transaction-aware execution and human oversight, translate recommendations into stable and auditable actions, aligning both with client objectives and regulatory requirements. Experimental results demonstrate improved goal-attainment probabilities, faster drawdown recovery, and enhanced explanation consistency, with measurable reductions in suitability breaches, turnover costs, and complaint likelihood. Beyond technical contributions, the framework highlights governance practices—such as evidence cards, immutable logs, and human-in-the-loop review—that strengthen accountability and trust. Future research will extend the framework toward cross-market generalization, meta-learning for adaptive calibration, safe exploration under compliance constraints, and multimodal interaction bias control. Large-scale randomized trials and field pilots will further validate external applicability and support organizational adoption in real-world advisory contexts.

References

[1] Migdadi M K, Oweidat I A, Alosta M R, et al. The association of artificial intelligence ethical awareness, attitudes, anxiety, and intention-to-use artificial intelligence technology among nursing students[J]. Digital Health, 2024.DOI:10.1177/20552076241301958.

[2] Gallegos I O, Rossi R A, Barrow J, et al. Bias and Fairness in Large Language Models: A Survey[J]. Computational Linguistics, 2024, 50(3).DOI:10.1162/coli_a_00524.

[3] Bollwerk E, Gupta N, Smith J .A Systems-Thinking Model of Data Management and Use in US Archaeology[J].Advances in Archaeological Practice, 2024, 12(1).DOI:10.1017/aap.2023.41.

[4] Acerbi C, Tasche D .Expected Shortfall: A Natural Coherent Alternative to Value at Risk[J]. Economic Notes, 2002.DOI:10.1111/1468-0300.00091.

[5] Nelson D B. Conditional heteroskedasticity in asset returns: A new approach[J]. Econometrica: Journal of the Econometric Society, 1991, 59(2): 347-370. DOI:10.2307/2938260.

[6] Nor M H M, Bakar M A A, Ariff N M, et al. Navigating the missing data maze: exploring multiple imputation techniques for environmental performance index data[J].IOP Publishing Ltd, 2025. DOI:10.1088/2515-7620/add8e7.

[7] Fan T, Wang J, Ren X, Huang C. MiniRAG: Towards extremely simple retrieval-augmented generation[EB/OL]. 2025. arXiv:2501.06713. DOI:10.48550/arXiv.2501.06713.

[8] Hartmann C, Smeyers-Verbeke J, Penninckx W, et al. Detection of bias in method comparison by regression analysis[J]. Analytica Chimica Acta, 1997, 338(1-2):19-40.DOI:10.1016/S0003-2670(96)00341-8.

[9] Chen Y, Wei W, Liu F, et al. Distributionally robust hydro-thermal-wind economic dispatch[J]. Applied Energy, 2016, 173(JUL.1):511-519.DOI:10.1016/j.apenergy.2016.04.060.

[10] Martin C .Is Systemic Risk Prevention the New Paradigm? A Proposal to Expand Investor Protection Principles to the Hedge Fund Industry[J].St Johns Law Review, 2014, 86. DOI:http://dx.doi.org/.

[11] Shahab S, Lades L K .Sludge and transaction costs[J]. Behavioural Public Policy, 2024, 8(2). DOI:10.1017/bpp.2021.12.

[12] Robert P, James S, Arun J, et al. Renal Replacement Therapy in Support of Operation Iraqi Freedom: A Tri-Service Perspective[J].Military Medicine, 2008(11):1115.DOI:10.7205/MILMED.173.11.1115.