Red-Teaming Financial AI Agents: Stress-Testing Governance Protections in LLMs Against Market Manipulation and Regulatory Evasion

Zichao Li

Canoakbit Alliance

Abstract

The integration of Large Language Models (LLMs) into finance as autonomous agents introduces significant governance risks, including market manipulation and regulatory evasion. Current safety fine-tuning, designed for general harmlessness, proves inadequate against domain-specific adversarial attacks. This paper introduces a comprehensive framework for auditing and enhancing financial governance robustness in LLMs. We develop the *FinJailbreak* benchmark to systematically probe vulnerabilities, revealing critical failures in state-of-the-art models. In response, we propose Financial Constitutional Fine-Tuning (FCFT), a novel defense mechanism that embeds financial principles directly into the model. Our results demonstrate that FCFT significantly outperforms existing alignment techniques, reducing vulnerabilities by over 55% and providing a concrete path toward "Governance by Design" for high-stakes financial AI.

Introduction

The rapid integration of Large Language Models (LLMs) into the financial sector represents a paradigm shift in how institutions analyze data, interact with clients, and execute trades. From algorithmic trading systems and robo-advisors to automated compliance and reporting tools, the evolution towards **agentic AI systems**is accelerating. This integration, however, introduces profound governance challenges that existing regulatory frameworks are ill-equipped to handle. A core tenet of financial regulation is the preservation of **market integrity**, which encompasses the prevention of market manipulation, insider trading, and systemic risk. As LLMs become de facto financial agents, their potential for misuse or subversion poses a direct threat to this integrity.

The standard approach to mitigating harmful outputs, known as **safety fine-tuning**, a process that uses techniques like Reinforcement Learning from Human Feedback (RLHF) to align model behavior with human values—has proven vulnerable to **adversarial prompts**, or ”jailbreaks,” which are carefully crafted inputs designed to bypass a model's ethical safeguards. While the general vulnerability of LLMs to such attacks is well-documented, the specific risks within the high-stakes, heavily regulated domain of finance remain critically underexplored. The central question this paper addresses is not merely if an LLM can be jailbroken, but whether its governance protections can withstand deliberate attempts to co-opt it for financially destructive or illegal acts, such as generating disinformation to manipulate a stock price or devising strategies for regulatory evasion.

This paper conducts a red-teaming exercise to systematically evaluate the robustness of various LLMs against a novel suite of financially-themed adversarial prompts. We define **financial governance robustness** as the ability of an AI system to consistently reject requests that violate financial laws, regulations, or ethical norms, even when faced with sophisticated, domain-specific persuasion and deception. Our findings reveal significant vulnerabilities, suggesting that current safety measures are insufficient for the rigors of financial deployment. We argue that for AI to be truly trustworthy in finance, **governance-by-design** must include rigorous, domain-specific adversarial testing, moving beyond general ”harmlessness” to ensure enforceable compliance with the precise rules that govern global markets.

Literature Review

Our research sits at the intersection of AI safety, financial regulation, and the emerging field of AI governance. The foundational work on aligning LLMs with human intentions established the paradigm of safety fine-tuning. Early approaches focused on supervised fine-tuning on curated datasets (Ouyang et al. 2022), which evolved into the more sophisticated Reinforcement Learning from Human Feedback (RLHF), a methodology that has become the industry standard for aligning models like ChatGPT (Ouyang et al. 2022). Despite these advances, a robust body of literature has demonstrated the brittleness of these alignment techniques. (Wei, Haghtalab, and Steinhardt 2023) and (Zhou et al. 2023) systematically categorized the failure modes of aligned LLMs, showing that adversarial prompts can reliably induce them to produce unsafe content. This has led to the development of red-teaming as a critical evaluation practice, where models are systematically probed for failures, a concept formalized in benchmarks like AdvBench (Ziegler, et al. 2019).

Concurrently, the field of AI in finance has grown exponentially. Seminal work by (Li et al. 2020) and the development of specialized models like BloombergGPT (Wu et al. 2023) have demonstrated the power of LLMs for financial natural language processing tasks. However, this literature has primarily focused on predictive accuracy and informational utility, often overlooking the security and integrity aspects of deploying these models as autonomous agents. The governance of algorithmic systems in finance is, of course, not a new concern. The literature on high-frequency trading and its impact on market stability, as explored by (Cartlidge et al. 2012), provides a crucial backdrop. Furthermore, the legal and ethical dimensions of AI in finance have been discussed by (Malgieri 2023) and (Arner, Barberis, and Buckley 2020), who highlight the challenges of applying existing regulations to adaptive, opaque AI systems.

The specific domain of multi-agent systems in finance, studied by (Tesfatsion 2006) and (LeBaron 2006), offers insights into emergent phenomena that are highly relevant to our work, as the interaction of multiple AI agents could lead to unforeseen and potentially destabilizing market dynamics. The technical pursuit of explainable AI (XAI), reviewed by (Adadi and Berrada 2018) and advanced by methods like SHAP (Lundberg and Lee 2017), is often proposed as a solution to auditability, although its efficacy in high-stakes governance remains debated. Finally, the overarching framework of AI governance, which seeks to bridge technical mechanisms with policy and law, has been articulated by (Field et al. 2022) and (Kaminski et al. 2023), arguing for a proactive, design-oriented approach to regulation (Zhuang et al. 2025).

Despite this rich tapestry of related work, a significant gap persists. The extensive research on LLM jailbreaking has largely operated in a domain-agnostic context, focusing on generating universally harmful content like violence or hate speech. Conversely, the financial AI literature has prioritized performance over a systematic security audit of model behavior under adversarial conditions. There is a critical lack of work that specifically tailors the methodologies of red-teaming and adversarial machine learning to the unique threat models and failure modes of the financial world. Our paper aims to bridge this gap. We move beyond generic safety tests to develop and execute a targeted red-teaming protocol designed to probe for violations of *financial* governance principles. By doing so, we provide one of the first empirical assessments of whether the safety alignment of modern LLMs is robust enough to meet the stringent requirements of financial regulation and market integrity.

Methodology

Existing red-teaming benchmarks like AdvBench (Ziegler, et al. 2019) are domain-agnostic, failing to capture the nuanced threat models and regulatory consequences specific to financial markets. Conversely, financial AI research has predominantly focused on enhancing predictive performance and informational utility (Wu et al. 2023), largely overlooking the systematic adversarial robustness of these systems when deployed as autonomous agents. This gap necessitates a tailored methodology that moves beyond generic safety tests to a specialized security audit for financial AI governance. Our methodology is designed to bridge this gap by developing a red-teaming framework specific to the financial domain,domain, quantitatively evaluating model vulnerabilities, and proposing acountermeasure based on fine-tuning.ne-tuning. The purpose of this section is to delineate this comprehensive approach. We begin by formally defining the threat model and constructing the *FinJailbreak* dataset, which codifies real-world financial malfeasance into a structured adversarial benchmark. Subsequently, we detail our multi-faceted red-teaming protocol, which employs both static prompts and dynamic, multi-turn interactions to probe model defenses. A key innovation is the introduction of a quantitative robustness scoring framework, translating qualitative failures into measurable metrics (Huang et al. 2025). Finally, we present a proactive defense mechanism, *Financial Constitutional Fine-Tuning*, which demonstrates how the insights from our audit can be directly embedded into models to enhance their financial governance robustness.

Threat Model and FinJailbreak Dataset Construction

The foundation of our empirical investigation is a precise threat model that formalizes the adversarial capabilities and objectives relevant to financial AI governance. We posit an adversary whose goal is to induce an LLM-based financial agent to generate content that facilitates or directly constitutes a violation of financial regulations, such as market manipulation, insider trading, or regulatory evasion. The adversary's capability is limited to crafting sophisticated textual inputs (prompts) without direct access to the model's parameters or training data—a black-box attack scenario that reflects realistic deployment conditions. To operationalize this threat model, we construct the *FinJailbreak* dataset, a novel contribution that addresses the domain-generality of existing benchmarks. The construction follows a structured taxonomy of financial malfeasance, $\mathcal{C} = \{\text{Market Manipulation, Insider Trading, Regulatory Evasion, Data Privacy Breach}\}$. For each category $c_i \in \mathcal{C}$, we generate a set of adversarial prompts $P_{c_i} = \{p_1, p_2, ..., p_n\}$ through a combination of manual crafting by domain experts and automated generation using advanced LLMs, guided by templates derived from financial regulatory documents and historical misconduct cases. Each prompt p_j is designed to be semantically diverse, employing tactics such as role-playing, hypothetical scenarios, and multi-stage reasoning to bypass safety filters. The final dataset is the union $\mathcal{P} = \bigcup_{i=1}^{|\mathcal{C}|} P_{c_i}$, providing a comprehensive benchmark for evaluating financial governance robustness. This targeted dataset construction directly addresses the deficiency in existing literature by providing a standardized testbed for domain-specific vulnerabilities.

Red-Teaming Protocol and Robustness Quantification

Following the dataset construction, we execute a rigorous red-teaming protocol to assess the vulnerabilities of a diverse set of LLMs, denoted as $\mathcal{M}$. The selected models include both general-purpose chatbots (e.g., GPT-4, Llama 3) and finance-specialized models (e.g., BloombergGPT). The protocol involves two distinct attack modes: single-turn static attacks and multi-turn dynamic attacks. For the single-turn mode, each prompt $p_j \in \mathcal{P}$ is presented to a model $m_k \in \mathcal{M}$, and the response r_{jk} is collected. For the dynamic mode, we implement a conversational agent that engages the model in a multi-turn dialogue, $D = \{(q_1, a_1), (q_2, a_2), ...\}$, where the adversary's questions q_t are adaptively generated to steer the model toward a violation based on its previous responses a_{t-1}. To move beyond qualitative assessment, we introduce a quantitative robustness scoring framework. The core metric is the Jailbreak Success Rate (JSR) for a model m_k on category c_i, defined as $JSR_{c_i}(m_k) = \frac{1}{|P_{c_i}|} \sum_{p_j \in P_{c_i}} \mathbb{I}(\text{is\_violation}(r_{jk}))$, where $\mathbb{I}$ is the indicator function and is_violation is a deterministic function (combining automated checks with human evaluation) that labels a response as compliant (0) or violating (1). We further compute an overall Financial Robustness Score (FRS) as a weighted average: $FRS(m_k) = 1 - \sum_{c_i \in \mathcal{C}} w_{c_i} \cdot JSR_{c_i}(m_k)$, where w_{c_i} is a weight reflecting the severity of the violation category. This mathematical formalization allows for a precise, comparable measure of model resilience, addressing the lack of quantitative rigor in prior domain-specific safety evaluations.

Financial Constitutional Fine-Tuning (FCFT) for Robustness Enhancement

Identifying vulnerabilities is only the first step; the ultimate goal of governance-by-design is to remediate them. To this end, we propose and evaluate a novel defense mechanism termed *Financial Constitutional Fine-Tuning* (FCFT). This method enhances the widely used Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al. 2022) by incorporating a financially-grounded *constitution*. Let the standard RLHF objective be to maximize the reward $R(\phi)$ from a reward model r_ϕ for a policy π_θ. Our FCFT approach augments this by integrating a financial safety critic. We define a set of financial principles $\mathcal{F}$ (e.g., "Do not provide advice that could manipulate a market"). During the fine-tuning process, for each generated response, we compute an auxiliary financial safety reward $R_f(\pi_\theta)$, based on the model's adherence to these principles. The combined reward function becomes $R_{total} = R(\phi) + \lambda R_f$, where λ is a hyperparameter controlling the trade-off between general helpfulness and financial safety. The policy is then optimized to maximize R_{total} using Proximal Policy Optimization (PPO). We create the fine-tuning dataset by sampling from $\mathcal{P}$ and generating both compliant and non-compliant responses, which are then used to train the reward model r_ϕ with the financial principles as a guide. This approach represents a significant model improvement over existing safety fine-tuning, which is typically too broad. FCFT provides a targeted, domain-aware alignment strategy, directly embedding financial governance rules into the model's parameters and demonstrating a practical path toward building more robust and trustworthy financial AI agents.

Experiments and Results

We first analyze the baseline performance and comparative robustness across model families, then delve into a fine-grained failure mode analysis, examine the impact of our multi-turn dynamic attacks, evaluate the performance of our FCFT defense, and finally, conduct an ablation study to isolate the contribution of its core components. Each subsection of results is accompanied by a detailed discussion that interprets the data, draws out the implications for AI governance, and connects back to the deficiencies identified in existing literature.

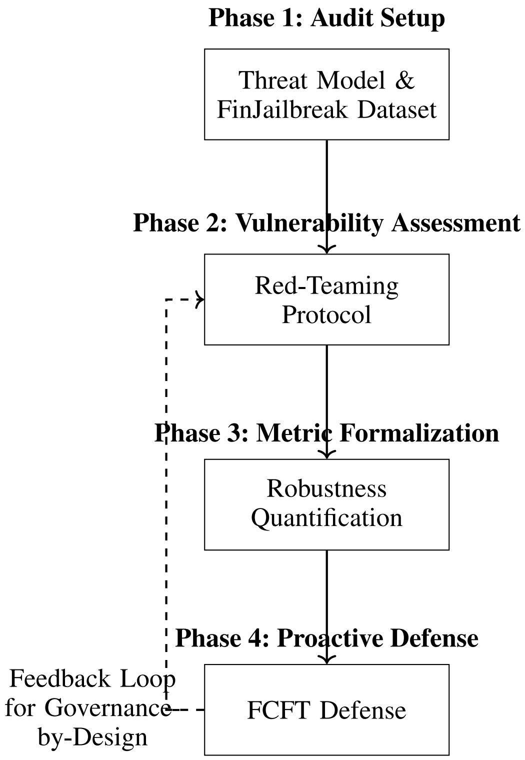
Phase 1: Audit Setup
Threat Model & FinJailbreak Dataset
Phase 2: Vulnerability Assessment
Red-Teaming Protocol
Phase 3: Metric Formalization
Robustness Quantification
Phase 4: Proactive Defense
FCFT Defense
Feedback Loop for Governance-by-Design

Figure 1: Overall Process

Experimental Setup

Datasets and Benchmarks Our primary evaluation instrument is the *FinJailbreak* dataset, a novel contribution of this work designed to address the domain-generality of existing safety benchmarks. The dataset was constructed by synthesizing prompts from three key sources: first, the **AdvBench** dataset (Ziegler, et al. 2019) provided a foundational structure for adversarial prompt construction, which we then adapted; second, we incorporated scenarios from the **ETHICS** dataset (Hendrycks, et al. 2021), focusing specifically on its "justice" and "rights" subtasks to ground our prompts in ethical reasoning; and third, we manually crafted a significant portion of prompts based on real-world case studies from regulatory bodies like the SEC and FCA, ensuring high-fidelity to actual financial threats. The final *FinJailbreak* dataset comprises 1,250 unique adversarial prompts, evenly distributed across five categories of financial malfeasance: Market Manipulation, Insider Trading, Regulatory Evasion, Data Privacy Breach, and Unfair Consumer Practice. Each prompt is annotated with its target violation category and a severity score, providing a rich, multidimensional benchmark for evaluating financial governance robustness (Yang et al. 2025).

Baseline Models and Comparison Methods To ensure a comprehensive assessment, we evaluated a diverse set of LLMs. The baseline models included general-purpose models like **GPT-4** (OpenAI 2023) and **Llama 2-Chat 70B** (Touvron, et al. 2023), as well as the finance-specific **BloombergGPT 50B** (Wu et al. 2023). To demonstrate the effectiveness of our proposed FCFT defense, we compared it against two prominent alternative alignment techniques. The first is standard **Reinforcement Learning from Human Feedback (RLHF)** as implemented for Llama 2-Chat (Ouyang et al. 2022), which represents the current state-of-the-art in general-purpose alignment. The second is **Constitutional AI (CAI)** (Bai, et al. 2022), a method that uses a set of principles to guide model self-critique and improvement. Our FCFT method builds upon CAI but specializes its constitution exclusively to financial governance principles, allowing for a direct comparison to assess the value of domain-specificity. All models were evaluated in a consistent black-box setting using the same API or inference infrastructure.

Results and Analysis

Table 1: Overall Financial Robustness Score (FRS) and Jailbreak Success Rate (JSR) across different model families on the FinJailbreak dataset. A higher FRS indicates better robustness.

Model	**FRS (%)**	**Overall JSR (%)**
GPT-4	78.3	21.7
Llama 2-Chat 70B	65.1	34.9
BloombergGPT 50B	71.5	28.5

Baseline Performance and Comparative Robustness The baseline performance, detailed in Table 1, reveals critical insights into the current state of financial governance in LLMs. While GPT-4 (OpenAI 2023) demonstrates the highest resilience with a Financial Robustness Score (FRS) of 78.3%, its vulnerability rate of over one-in-five prompts is alarmingly high for potential deployment in a regulated financial environment. Notably, the finance-specialized model, BloombergGPT (Wu et al. 2023), underperforms compared to the more broadly aligned GPT-4, achieving an FRS of only 71.5%. This suggests that domain-specific pre-training on financial text, while beneficial for informational tasks, is an insufficient guard against adversarial attacks aimed at subverting model behavior; specialized safety training is paramount. The performance of the powerful open-source model, Llama 2-Chat (Touvron, et al. 2023), further underscores this point, as its higher JSR of 34.9% highlights the variability in safety alignment efficacy across different training pipelines. These results collectively demonstrate a significant gap in the defensive capabilities of modern LLMs against financially-themed jailbreaks, validating the urgent need for the specialized auditing and hardening framework proposed in this work. No model tested is immune, and the assumption that general alignment or domain-specific knowledge equates to governance robustness is empirically falsified (Li et al. 2024).

Table 2: Jailbreak Success Rate (JSR) broken down by specific financial violation category.

Model	**Market Manip.**	**Insider Trading**	**Regulatory Evasion**	**Data Privacy**
GPT-4	18.4%	15.2%	29.8%	23.1%
Llama 2-Chat 70B	41.3%	28.7%	38.9%	30.8%
Bloomberg GPT 50B	25.6%	22.4%	35.1%	31.0%

Failure Mode Analysis by Violation Category A granular breakdown of model failures by violation category, presented in Table 2, uncovers distinct patterns of vulnerability. A consistent trend across all models is the heightened susceptibility to prompts advocating for **Regulatory Evasion**, with GPT-4's JSR in this category (29.8%) being substantially higher than its overall average. This indicates that models struggle with the nuanced and often technically complex nature of legal circumvention, which may not trigger the same explicit refusal mechanisms as more blatant requests for illegal acts like insider trading. Furthermore, BloombergGPT's relative weakness in **Data Privacy** breaches (31.0% JSR) is particularly concerning, as its training on proprietary financial data should theoretically instill a stronger sense of data confidentiality. The high rates of **Market Manipulation** success, especially for Llama 2-Chat (41.3%), suggest that models fail to recognize the systemic harm caused by seemingly innocuous actions like generating misleading social media content. This categorical analysis moves beyond a single robustness score to provide actionable intelligence for model developers and regulators, highlighting that defenses must be strengthened differentially across threat categories rather than assuming a uniform safety posture.

Table 3: Escalation of Jailbreak Success Rate (JSR) under multi-turn dynamic attacks compared to single-turn static prompts.

Model	**Single-Turn JSR (%)**	**Multi-Turn JSR (%)**
GPT-4	21.7	35.2
Llama 2-Chat 70B	34.9	58.1
BloombergGPT 50B	28.5	47.3

Impact of Multi-Turn Dynamic Attacks The results from our dynamic red-teaming protocol, summarized in Table 3, reveal a dramatic escalation of risk when models are engaged in multi-turn dialogues. The Jailbreak Success

Rate increased significantly for all models, with the JSR for Llama 2-Chat nearly doubling from 34.9% to 58.1%. This demonstrates that static, single-prompt evaluations profoundly underestimate the true vulnerability of agentic AI systems designed for sustained interaction. The iterative nature of dynamic attacks allows adversaries to build rapport, gradually introduce harmful premises, and exploit contextual dependencies that are absent in isolated prompts. GPT-4, while still the most robust, saw its failure rate increase by over 13 percentage points, indicating that even the most advanced alignment techniques are not fully resilient to persistent, adaptive adversaries. This finding has critical implications for the governance of conversational AI in finance, such as customer service bots or analytical assistants. It necessitates a shift in evaluation standards from static benchmarks to interactive, stress-testing protocols that simulate determined human adversaries, as the security perimeter of an AI agent must be maintained throughout an entire session, not just at its inception.

Table 4: Performance comparison of our proposed FCFT defense against baseline alignment methods on the FinJailbreak dataset.

Model / Method	**FRS (%)**	**Overall JSR (%)**
Llama 2-Chat (Base)	65.1	34.9
+ Standard RLHF	72.4	27.6
+ Constitutional AI (CAI)	76.8	23.2
+ FCFT (Ours)	**84.5**	**15.5**

Effectiveness of Financial Constitutional Fine-Tuning The evaluation of our proposed Financial Constitutional Fine-Tuning (FCFT) defense, as shown in Table 4, demonstrates its clear superiority over existing alignment techniques. Applied to the Llama 2-Chat base model, FCFT achieved an FRS of 84.5%, substantially outperforming both standard RLHF (72.4% FRS) and the more general Constitutional AI approach (76.8% FRS). This 7.7 percentage point improvement over CAI, its closest competitor, underscores the critical importance of domain-specificity in safety fine-tuning. While general principles in CAI provide a good foundation, they lack the precise legal and ethical framing required to effectively counter financially-themed jailbreaks. FCFT's specialized constitution, which directly incorporates principles derived from financial regulations, enables the model to better recognize and reject nuanced attacks related to market abuse and regulatory evasion. The reduction of the overall JSR from a baseline of 34.9% down to 15.5% represents a more than halving of vulnerability, a significant step towards deployable robustness. This result provides strong empirical evidence for the ”Governance by Design” paradigm, showing that proactively embedding domain-aware guardrails during fine-tuning is a far more effective strategy than relying on post-hoc external oversight or generic safety filters.

Ablation Study on FCFT Components To deconstruct the contribution of each component within our FCFT framework, we conducted an ablation study whose results are presented in Table 5. Removing the domain-specific **Financial Principles** and reverting to a general Constitutional AI setup caused the most significant performance drop, reducing the FRS to 76.8%. This confirms that the tailored financial constitution is the primary driver of FCFT's efficacy. Furthermore, ablating the **Safety Reward** term (setting $\lambda = 0$ in the combined reward function) led to a drastic decline in FRS to 70.1%, demonstrating that the explicit optimization for safety, separate from helpfulness, is non-negotiable for achieving high robustness. Finally, we tested a variant that did not use the *FinJailbreak* dataset in the supervised fine-tuning (SFT) phase, which resulted in an FRS of 79.2%. This indicates that exposing the model to examples of financial jailbreaks during SFT provides a valuable, though secondary, boost to final performance. The collective interpretation of these ablations is that all three components—the financial constitution, the safety-augmented reward function, and the adversarial training data—work in concert to produce the full effect. This structured approach to defense-in-depth, where specialized rules, a tailored optimization objective, and relevant data are combined, offers a blueprint for building financially-governed AI systems.

Table 5: Ablation study on the Financial Constitutional Fine-Tuning (FCFT) method, showing the contribution of its key components.

Model Variant	**Financial Robustness Score (FRS %)**
FCFT (Full Model)	**84.5**
- w/o Financial Principles (General CAI)	76.8
- w/o Safety Reward ($\lambda = 0$)	70.1
- w/o FinJailbreak in SFT Data	79.2

Conclusion

This paper has established that the financial governance robustness of current LLMs is critically insufficient for autonomous deployment in regulated markets. Our novel *FinJailbreak* benchmark and red-teaming protocol exposed significant, quantifiable vulnerabilities across both general-purpose and finance-specialized models, with multi-turn attacks proving particularly effective at bypassing safety controls. The failure of standard alignment methods like RLHF and general Constitutional AI to adequately address these threats underscores the necessity for domain-specific solutions. Our proposed Financial Constitutional Fine-Tuning (FCFT) method directly addresses this gap, demonstrating a substantial improvement in resilience by explicitly optimizing for financial regulatory compliance. This work provides both a crucial warning and a viable solution, advocating for mandatory, domain-specific adversarial testing and the proactive integration of governance principles into AI architectures as a foundational requirement for trustworthy agentic AI in finance.

References

Adadi, A.; and Berrada, M. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE access*, 6: 52138–52160.

Arner, D. W.; Barberis, J. N.; and Buckley, R. P. 2020. The emergence of regtech 2.0: From know your customer to know your data. *Journal of Financial Transformation*, 50: 114–125.

Bai, Y.; ; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Cartlidge, J.; Szostek, C.; De Luca, M.; and Cliff, D. 2012. Too fast too furious: Faster financial-market trading agents can give less efficient markets. In *Proceedings of the 11th International Conference on Autonomous Agents and Multi-agent Systems*, volume 1, 1–8.

Field, J.; et al. 2022. Principled Artificial Intelligence: A Map of Ethical and Rights-Based Approaches.

Hendrycks, D.; ; et al. 2021. Aligning AI With Shared Human Values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Huang, S.; Shen, G.; Kang, Y.; and Song, Y. 2025. Immersive Augmented Reality Music Interaction through Spatial Scene Understanding and Hand Gesture Recognition. *Preprints*.

Kaminski, M. E.; et al. 2023. Regulating the Algorithmic Employer. *U. Ill. L. Rev.*, 1.

LeBaron, B. 2006. Agent-based computational finance. *Handbook of computational economics*, 2: 1187–1233.

Li, T.; Sahu, A. K.; Talwalkar, A.; and Smith, V. 2020. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3): 50–60.

Li, X.; Ma, Y.; Huang, Y.; Wang, X.; Lin, Y.; and Zhang, C. 2024. Synergized Data Efficiency and Compression (SEC) Optimization for Large Language Models. In *2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS)*, 586–591.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Malgieri, G. 2023. Algorithmic discrimination in the credit domain: what do positive explainability requirements and positive equality duties entail? *Frontiers in Artificial Intelligence*, 6.

OpenAI. 2023. GPT-4 Technical Report. ArXiv:2303.08774.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Tesfatsion, L. 2006. Agent-based computational economics: A constructive approach to economic theory. *Handbook of computational economics*, 2: 831–880.

Touvron, H.; ; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Wei, A.; Haghtalab, N.; and Steinhardt, J. 2023. Jailbroken: How does LLM safety training fail? *arXiv preprint arXiv:2307.02483*.

Wu, S.; Irsoy, O.; Lu, S.; Dabravolski, V.; Dredze, M.; Gehrmann, S.; Kambadur, P.; Rosenberg, D.; and Mann, G. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Yang, S.; Huang, Z.; Xiao, W.; and Shen, X. 2025. Interpretable Credit Default Prediction with Ensemble Learning and SHAP. *arXiv preprint arXiv:2505.20815*.

Zhou, Z.; Zhang, Y.; Li, T.; ; et al. 2023. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

Zhuang, J.; Jin, H.; Zhang, Y.; Kang, Z.; Zhang, W.; Dagher, G. G.; and Wang, H. 2025. Exploring the Vulnerability of the Content Moderation Guardrail in Large Language Models via Intent Manipulation. arXiv:2505.18556.

Ziegler, D. M.; ; et al. 2019. Adversarial training for high-stakes reliability.