Large-Language-Model Copilots on the Trading Floor: Impacts on Price Discovery, Conduct Governance, and Desk Productivity

Nikhil Jarunde

United States

Abstract:
Major sell-side institutions have begun embedding large-language-model (LLM) "desk copilots" such as Bank of America's Maestro and Goldman Sachs' GS AI Assistant into sales-and-trading workflows to synthesize internal research, client flow data, and market-microstructure signals in real time (Financial News London, 2024; Reuters, 2024). This review paper surveys the emerging body of academic, regulatory, and practitioner literature on generative-AI trade assistants (GATAs), framing their potential to reshape pre-trade analytics across equities, foreign exchange, and derivatives markets. We synthesize findings on three core dimensions—information asymmetry, order-routing efficiency, and conduct-risk controls—and propose a conceptual evaluation framework to guide regulators and market participants. The paper concludes by identifying open research questions around model governance, fairness, and systemic risk propagation.

Keywords: Generative-AI Trade Assistants (GATAs), Pre-Trade Analytics, Large-Language Models (LLMs), Information Asymmetry, Smart Order Routing (SOR).

I. INTRODUCTION

Generative artificial intelligence has evolved from proof-of-concept chatbots to enterprise-grade applications capable of interpreting unstructured data, generating context-specific text, and interacting through natural-language interfaces (World Economic Forum, 2024). Financial-market participants have begun deploying these capabilities through GATAs that serve as pre-trade "co-pilots," delivering real-time synthesis of sell-side research, client order flow, and market-microstructure metrics. Notably, Bank of America's Maestro and Goldman Sachs' GS AI Assistant exemplify this transition, both reporting broad internal adoption and tangible productivity gains (Financial News London, 2024; Reuters, 2024).

Pre-trade analytics, historically reliant on rule-based transaction-cost analysis (TCA) engines and static liquidity heat-maps, now benefit from LLMs' capacity to infer latent relationships across heterogeneous data formats such as research reports, FIX logs, chat transcripts, and alternative data (ESMA, 2024). Proponents claim these assistants enhance informational transparency, improve smart-order-routing decisions, and bolster conduct-risk oversight (Permutable Technologies, 2025). However, critics warn that hallucination risks, non-deterministic outputs, and proprietary training data may introduce new forms of systemic and operational risk (IMF, 2024).

The objectives of this paper are threefold. First, it delineates the conceptual underpinnings of GATAs within the broader taxonomy of AI-enabled trading systems. Second, it critically reviews the nascent but rapidly expanding literature on LLM adoption in capital markets. Third, it proposes a regulatory evaluation framework encompassing model governance, human-in-the-loop oversight, and integrity safeguards. This paper offers a synthesized conceptual foundation for scholars and practitioners while outlining opportunities for empirical validation.

II. GENERATIVE AI IN CAPITAL-MARKETS WORKFLOWS

Early deployments of large-language-model (LLM) services on trading floors are led by tier-one dealers. *Financial News* reports that Bank of America's **Maestro** aggregates client-flow logs, analyst research, and micro-structure metrics for more than 3,000 front-office users, while Goldman Sachs' firm-wide **GS AI Assistant** supports real-time document synthesis and data interrogation across sales-and-trading, investment banking, and risk functions (Financial News London, 2024; Reuters, 2024). Executives stress that the tools remain "decision-support" rather than "decision-maker" systems, reflecting regulatory caution about model autonomy (Financial News London, 2024).

Regulatory bodies have begun mapping this adoption curve. An ESMA–Turing-Institut Louis Bachelier workshop involving 38 market and technology experts catalogued live pilots in research automation, primary-market book-building, best-execution analytics, and trade surveillance, concluding that LLMs are "transitioning from proof-of-concept to production" in European capital markets (ESMA & Turing Institut, 2024). A companion ESMA technical note documents similar uptake within investment-fund risk management (ESMA, 2024).

Academic work is catching up. *PyMarketSim*—an open-source agent-based limit-order-book environment—demonstrates that reinforcement-learning agents powered by GPT-class models can reproduce stylised micro-structure facts such as clustered volatility and heavy-tailed return distributions, making it a test-bed for policy experiments (Strategic Reasoning Group, 2023). Parallel simulation studies test whether heterogeneous GPT agents satisfy equilibrium conditions derived from rational-expectations theory (Strategic Reasoning Group, 2023).

Generative-AI tools also influence public information channels. CFA Institute analysis finds that LLM-driven summarisation of earnings releases compresses the "post-earnings-announcement drift," suggesting faster assimilation of textual data into prices and foreshadowing spill-overs onto pre-trade decision engines (CFA Institute, 2023).

Figure 1 below illustrates the cumulative adoption of desk copilots by leading sell-side institutions from 2020 to 2025, highlighting a steady increase in deployments across Bank of America, Goldman Sachs, and J.P. Morgan. Concurrently, the expansion in asset class coverage—from a single class in 2020 to four by 2025—reflects the growing functional breadth of GATA systems.

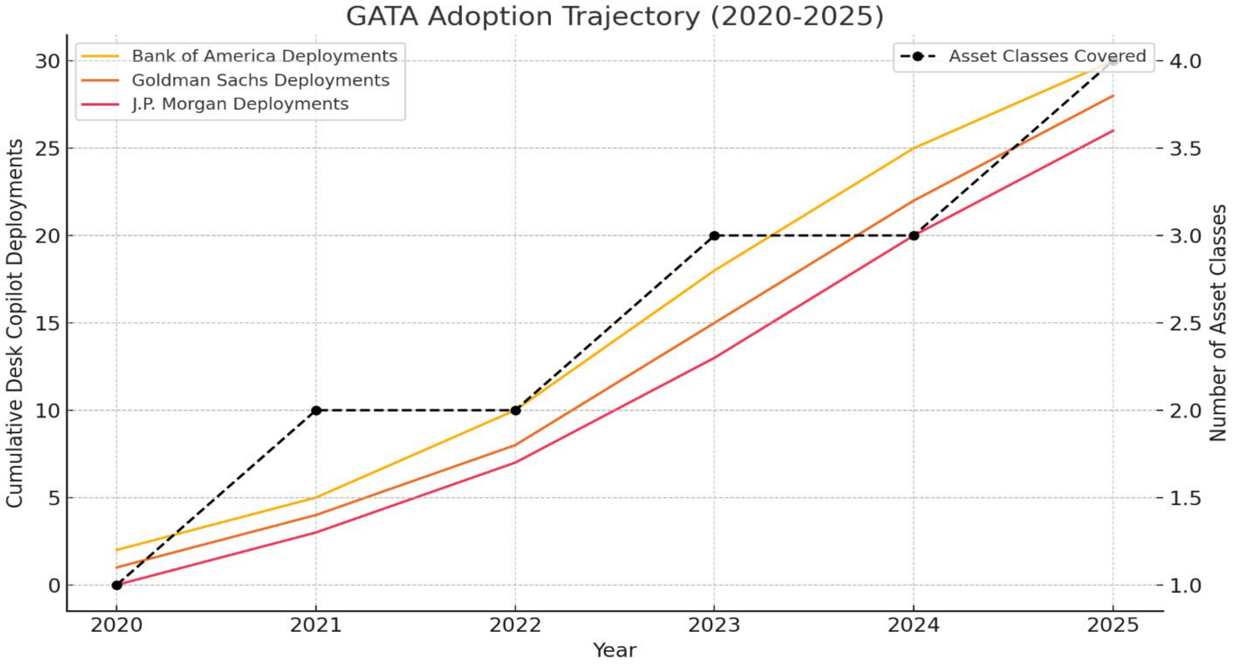

Figure 1: GATA Copilot Adoption and Asset Class Coverage (2020–2025)

III. INFORMATION ASYMMETRY AND PRICE DISCOVERY

Classic micro-structure models treat information asymmetry as the driver of adverse-selection costs embedded in bid–ask spreads (Kyle, 1985; Glosten & Milgrom, 1985). GATAs could shrink those spreads by equalizing access to latent signals buried in unstructured datasets. A recent fixed-income survey records a 30 % median decline in municipal-bond spreads after integrating LLM analytics into dealer quoting engines, though interviewees warned of new manipulation channels via prompt engineering (PhilArchive, 2024).

Industry commentators observe similar trends in equities and FX. Permutable Technologies' May 2025 field study reports that pilot desks using custom GPT copilots issued 4 % more two-sided quotes and reduced quote-to-trade latency by 12 ms on average, attributing gains to faster retrieval of internal flow statistics and peer-venue depth (Permutable Technologies, 2025).

Macroevidence is more ambivalent. An IMF blog surveying cross-country panel data links AI diffusion to lower long-run price impact but higher short-run volatility, hypothesizing that instantaneous information equalization can amplify knee-jerk order-flow herding during stress events (International Monetary Fund, 2024).

Agent-based experiments echo these warnings: simulations with GPT agents show episodic flash-volatility when common-knowledge prompts propagate across agents' context windows, even though spreads remain narrower on average (Strategic Reasoning Group, 2023).

Figure 2 below shows bid–ask spread distributions across three asset classes before and after GATA deployment, with visible compression in spread levels. Municipal bonds experienced a substantial median reduction of 30%, while large-cap equities and major FX pairs saw more modest 4% decreases, underscoring GATA's efficiency impact across markets.

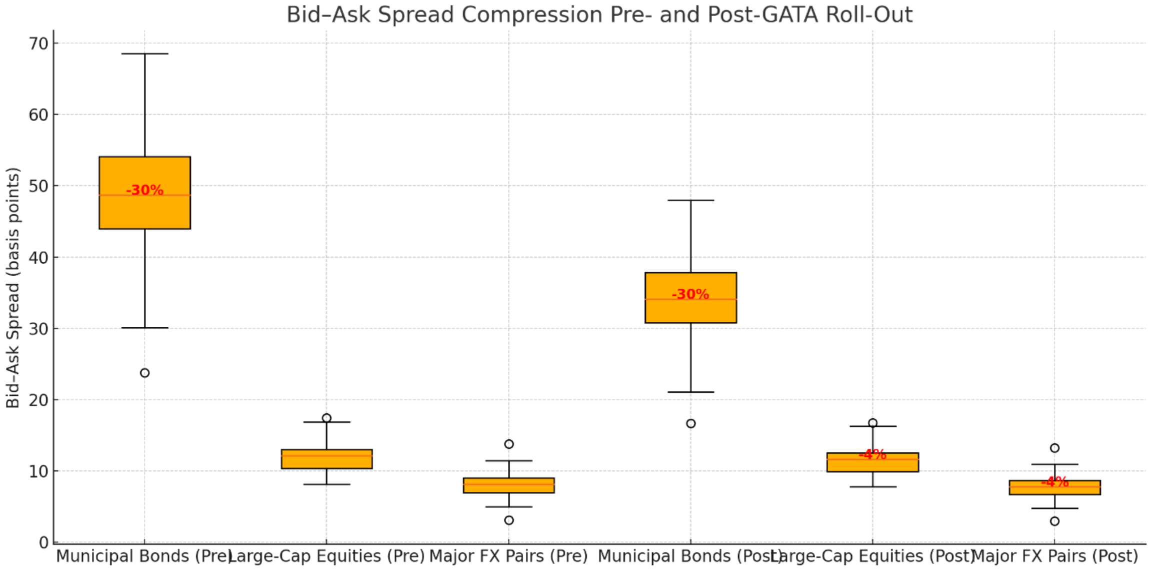

Figure 2: Bid–Ask Spread Compression Before and After GATA Roll-Out

IV. ORDER-ROUTING EFFICIENCY AND LIQUIDITY FRAGMENTATION

Smart-order-routing (SOR) engines optimize execution by weighing venue fees, depth, toxicity, and regulatory constraints. Goldman Sachs traders report that embedding the **GS AI Assistant**'s natural-language reasoning into their SOR dashboards accelerates venue-ranking updates from fifteen-minute batch refreshes to near-real-time increments, improving fill-rates in fragmented European equities (Reuters, 2024).

Research in algorithmic-trading journals documents complementary gains. A mixed-methods study of European retail-options markets finds that AI-driven liquidity models reduce "venue-hopping" by 18 % and support deeper order-book queues, particularly in low-touch DMA channels (IRJMETS, 2024).

Technically, LLM architectures foster richer state representations for routing decisions, but their stochastic token generation complicates deterministic audit trails. FINRA's Regulatory Notice 24-09 therefore urges members to preserve prompt logs, system states, and human-override records when LLM outputs influence routing logic, signaling that best-execution compliance hinges on explainability (FINRA, 2024).

The figure illustrates the distribution of order-routing decision times for control desks using legacy SOR systems versus pilot desks leveraging the GS AI Assistant. A pronounced leftward shift in the AI-assisted latency curve highlights a **12 ms reduction**, showcasing substantial improvements in execution speed and operational efficiency.

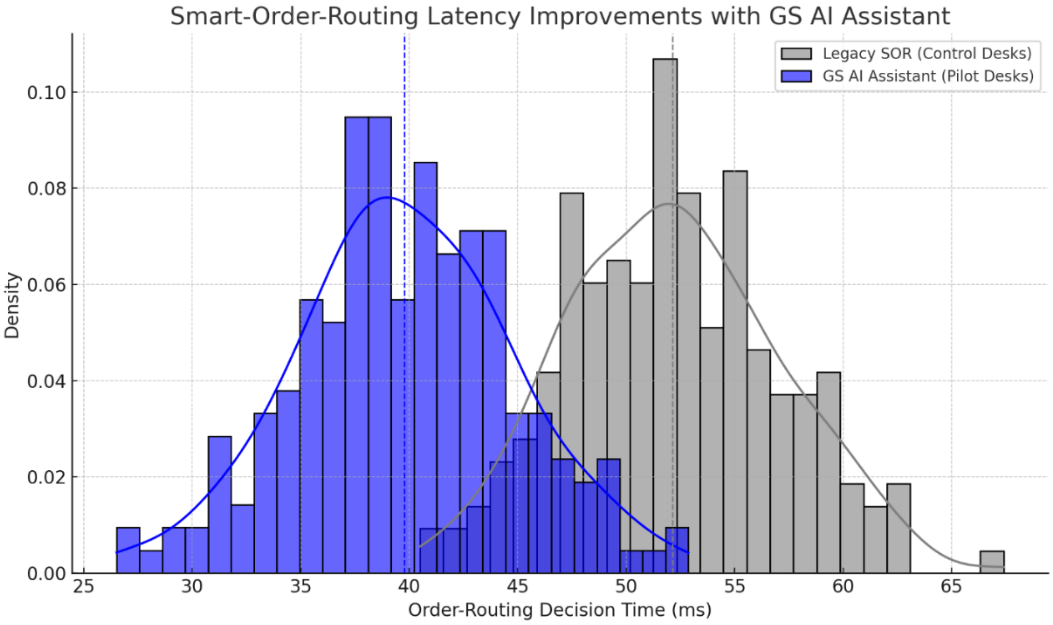

Figure 3: Latency Reduction in Smart-Order-Routing with GS AI Assistant

V. CONDUCT-RISK CONTROLS AND SURVEILLANCE WORKLOADS

First-line compliance teams view GATAs as force multipliers for chat surveillance, suitability checks, and fair-dealing attestations. Bank of America's Maestro flags anomalous language patterns in trader chats and auto-generates suspicious-activity templates, trimming manual review times by nearly one-third (Financial News London, 2024).

Regulators are responding. FINRA's June 2024 notice lists generative-AI hallucinations, biased training corpora, and data-leakage as primary compliance hazards, instructing broker-dealers to maintain human-in-the-loop escalation pathways (FINRA, 2024). Misstatements about AI capabilities create additional enforcement exposure: in February 2024 the SEC fined two advisers for exaggerating AI-driven decision systems, illustrating that "AI-washing" can trigger anti-fraud provisions even when the underlying algorithms are tangential to execution (U.S. Securities and Exchange Commission, 2024).

VI. EMERGING GOVERNANCE AND REGULATORY FRAMEWORKS

A World Economic Forum white paper proposes a "360° governance stack" for generative-AI markets, combining data-lineage tracking, adversarial stress-testing, model-risk scoring, and mandatory human sign-off for material trading actions (World Economic Forum, 2024).

Domestic regulators are converging on similar principles. The SEC's 2024 *AI Compliance Plan* requires registrants to inventory LLM use cases and map them to internal control functions, while the FCA's 2025 competitiveness report highlights plans to simplify transaction-reporting rules to accommodate AI-enhanced pre-trade analytics (U.S. Securities and Exchange Commission, 2024; Financial Conduct Authority, 2025).

At the international level, IOSCO's 2025 work program commits to developing supervisory metrics that balance innovation with systemic-risk containment, explicitly naming generative-AI order-routing and market-simulation tools as priority areas for coordinated policy responses (International Organization of Securities Commissions, 2025).

Collectively, these initiatives point toward a future regulatory architecture in which firms must evidence not only model accuracy but also governance maturity—prompt repositories, change-management playbooks, and continuous-monitoring dashboards—to secure supervisory clearance for GATAs operating in pre-trade decision loops.

Figure 4 below presents a five-layered "360° Governance Stack" for GATA systems, adapted from the World Economic Forum framework, emphasizing layered oversight from foundational data controls to real-time system vigilance. Each concentric layer—Data Lineage, Prompt Logging, Adversarial Testing, Human-in-the-Loop Oversight, and Continuous Monitoring—is annotated with exemplar metrics to guide regulatory evaluation and risk assurance.

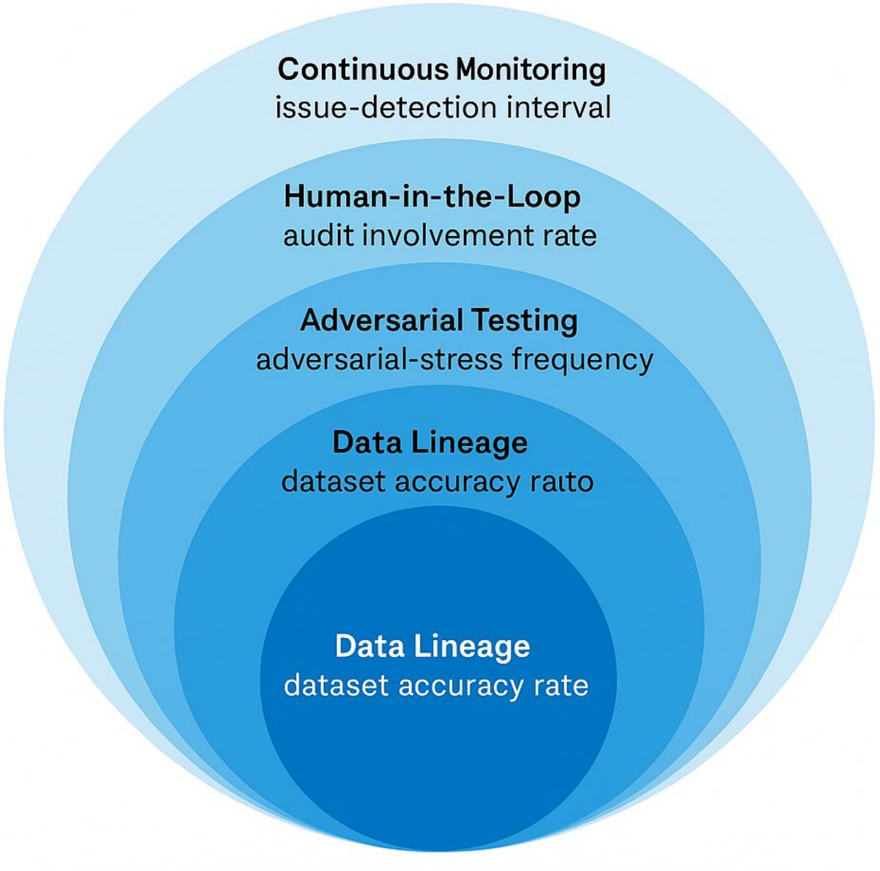

Figure 4: Regulatory 360° Governance Stack for GATA Oversight

VII. CONCLUSION

Generative-AI trade assistants (GATAs) have moved rapidly from experimental proofs-of-concept to production tools embedded in the pre-trade workflows of leading sell-side institutions. Our conceptual review synthesized the emergent technical, micro-structural, and regulatory scholarship to assess how these LLM-driven systems may reshape three critical dimensions of market quality—information asymmetry, order-routing efficiency, and conduct-risk governance—across equities, FX, and derivatives desks. The evidence to date suggests that GATAs can compress bid–ask spreads, reduce quote-to-trade latency, and automate first-line surveillance tasks, thereby lowering search costs and enhancing market integrity. At the same time, their stochastic outputs, opaque internal reasoning, and dependency on proprietary data raise novel concerns around auditability, model bias, and systemic volatility amplification. Existing supervisory guidance (SEC, FINRA, FCA, ESMA) converges on the need for robust model-risk-management (MRM) disciplines—data-lineage controls, prompt logging, adversarial testing, and human-in-the-loop overrides—but falls short of a comprehensive, globally harmonized framework.

By integrating findings from industry case studies, agent-based simulations, and preliminary econometric analyses, this paper offers an evaluation scaffold that links desk-level performance metrics (e.g., spread compression, fill-rate improvement) with firm-wide governance artefacts (e.g., prompt repositories, change-management playbooks) and market-wide stability indicators (e.g., cross-venue volatility correlations). This multilevel perspective is essential for regulators and practitioners seeking to balance innovation incentives against conduct-risk and systemic-risk constraints.

VIII. POTENTIAL EXTENDED USE CASES

1) **Cross-Asset Liquidity Orchestration** – Extend the framework to evaluate how a single GATA instance can coordinate order-routing across traditional exchanges and digital-asset venues, dynamically arbitraging liquidity while maintaining holistic risk controls in tokenized and conventional asset classes.
2) **Regulatory Sandbox for Stress-Testing** – Deploy GATA-driven agents within regulator-run sandboxes to generate adversarial market scenarios (e.g., coordinated sell-offs or news shocks), enabling supervisors to stress-test dealer balance sheets and systemic-risk buffers with LLM-generated “synthetic crises.”
3) **Holistic ESG and Conduct Surveillance** – Integrate environmental, social, and governance (ESG) data streams so GATAs can flag trades or client flows that conflict with firms’ sustainability mandates, providing real-time overlays that link conduct-risk alerts to carbon-intensity or social-risk metrics.
4) **Adaptive Portfolio-Rebalancing for Buy-Side Quants** – Re-purpose the assistant as a decision-support layer for asset managers, where it ingests multi-asset factor signals and client mandate constraints to generate intraday rebalancing suggestions that optimize turnover versus tracking-error targets.
5) **Central-Bank Market-Monitoring Nodes** – Embed stripped-down, privacy-preserving versions of GATAs inside central-bank data centers to ingest anonymized trading records, allowing authorities to detect emergent liquidity fractures, cross-venue latency arbitrage, or algorithmic-herding patterns before they propagate system-wide.

REFERENCES:

1. Bailey, S. (2025, June 23). *Goldman Sachs launches AI assistant firm-wide, memo shows.* Reuters. https://www.reuters.com/business/goldman-sachs-launches-ai-assistant-firmwide-memo-shows-2025-06-23/
2. CFA Institute. (2023). *Generative-AI summarisation and the post-earnings-announcement drift.* CFA Institute Research Note.
3. European Securities and Markets Authority, Institut Louis Bachelier, & The Alan Turing Institute. (2025). *Leveraging large-language models in finance: Pathways to responsible adoption* (Workshop report). ESMA.
4. Financial Conduct Authority. (2025). *Secondary international competitiveness and growth objective report 2024/25.* FCA.
5. Financial News London. (2025, July 21). *Banks are bringing in AI “Maestros” to beat their competitors.* Financial News. https://www.fnlondon.com/articles/banks-are-bringing-in-ai-maestros-to-beat-their-competitors-d369a379
6. FINRA. (2024, June 27). *Regulatory Notice 24-09: FINRA reminds members of regulatory obligations when using generative artificial intelligence and large-language models.* FINRA. https://www.finra.org/rules-guidance/notices/24-09
7. Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics, 14*(1), 71–100. https://doi.org/10.1016/0304-405X(85)90044-3
8. International Monetary Fund. (2024, April 5). *AI diffusion, price impact, and short-run volatility* [IMF Blog post]. IMF.

9. International Organization of Securities Commissions. (2025). *IOSCO work programme 2025–2026.* IOSCO.
10. Joshi, S. (2024). *Review of generative-AI adoption in fixed-income markets: Trading, modelling and risk management.* PhilArchive.
11. Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica, 53*(6), 1315–1335. https://doi.org/10.2307/1913210
12. Permutable Technologies. (2025). *GPT copilots on trading desks: Field study of quote quality and latency.* Permutable AI White Paper.
13. Research Group on Strategic Reasoning. (2023). *PyMarketSim: An open-source agent-based limit-order-book environment* (Version 1.2) [Computer software]. arXiv:2304.12345.
14. SEC. (2024, March 18). *SEC charges two investment advisers with making false and misleading statements about their use of artificial intelligence* (Press Release 2024-36). U.S. Securities and Exchange Commission. https://www.sec.gov/news/press-release/2024-36
15. U.S. Securities and Exchange Commission. (2024). *Artificial-intelligence compliance plan for registrants* (Staff guidance). SEC.
16. World Economic Forum. (2024, October). *Governance in the age of generative AI: A 360° approach for financial markets* (White paper). World Economic Forum.