

Large Language Model Agents for Automated Financial Risk Analysis and Decision Support

Pim Meeres

Department of Computer Science, University of Central Florida, Orlando, FL, USA.
meeres1985@ucf.edu

Duilio Canales

Department of Computer Science and Engineering, University of Nevada, Reno, NV, USA.
duiliocanales674@unr.edu

Cahal Guerin

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
cahal.guerin250@colostate.edu

Abstract

Large language model agents are increasingly proposed as instruments for automating financial risk analysis and decision support, yet their deployment in regulated financial environments raises structural questions that cannot be reduced to model accuracy. This paper examines the system-level architecture, governance, infrastructure, and policy implications of using large language model agents in financial risk workflows. It argues that effective automation depends less on a single generative model than on an integrated socio-technical stack that combines data ingestion, retrieval, tool use, multi-agent coordination, human oversight, auditability, and regulatory alignment. The paper analyzes trade-offs among latency, explainability, robustness, fairness, and sustainability in high-stakes financial settings. It further considers how regime shifts, liquidity stress, operational disruptions, and adversarial information environments challenge the reliability of automated agents. Through cross-domain comparisons with conventional machine learning systems, operational risk management, and human-centered artificial intelligence, the paper identifies recurring failure modes such as automation bias, hidden technical debt, opaque escalation paths, and uneven accountability. It concludes that responsible deployment requires trust calibration, continuous evaluation, model risk governance, and clear boundaries for autonomous action. Rather than treating large language model agents as substitutes for human judgment, the paper positions them as bounded decision-support infrastructures whose value depends on institutional controls, data quality, and transparent escalation. The analysis offers a forward-looking perspective for banks, regulators, and technology providers seeking to reconcile innovation with financial stability.

Keywords

large language model agents, financial risk analysis, decision support systems, multi-agent systems, governance, fairness, robustness, systemic risk, human oversight, regulatory technology.

1. Introduction

Financial risk analysis has become a computationally intensive and institutionally distributed activity. Banks, asset managers, insurers, and regulators must interpret vast quantities of market data, filings, news, internal exposures, and counterparty information under severe time constraints. Large language models have entered this space because they can summarize documents, extract entities, generate narratives, and interact with software tools through natural language interfaces [1], [2]. Their capacity to generalize across tasks has encouraged the development of autonomous or semi-autonomous agents that retrieve data, call analytical functions, and coordinate with other agents in pursuit of a risk assessment objective [3]. The promise is substantial: faster monitoring, broader coverage of unstructured information, and more adaptive decision support during market stress.

The promise is accompanied by structural risks. Generative systems can produce plausible but unsupported claims, inherit biases from training corpora, and behave unpredictably when deployed outside their original distribution [4]. Financial institutions also face longstanding challenges in operational risk, model risk, and technology governance that become more acute when automated agents participate in consequential decisions [5], [6]. Machine learning systems accrue hidden technical debt through entanglement, undeclared consumers, and unstable dependencies, and these problems magnify when agents are composed into multi-step workflows [7]. Financial regulation further imposes requirements for auditability, explainability, and accountability that are not satisfied merely by high predictive performance [8]. The central question is therefore not whether large language model agents can produce useful financial analysis, but under what architectural and governance conditions they can do so without undermining institutional resilience.

This paper adopts a system-level perspective. It examines how large language model agents should be designed, deployed, monitored, and governed when they are embedded in financial risk analysis and decision support. Section two reviews foundational work on language models, agentic systems, and financial risk management. Section three proposes a layered architecture for automated risk analysis. Section four discusses data, infrastructure, and sustainability. Section five analyzes decision support, governance, and fairness. Section six considers robustness under regime shifts and systemic stress. Section seven addresses evaluation, deployment, and policy implications. The conclusion synthesizes the argument and identifies priorities for future research.

2. Foundations and Related Work

The technical foundation of modern large language model agents lies in transformer architectures and large-scale pretraining [1]. These models learn statistical regularities from broad corpora and can perform tasks through prompting, few-shot examples, and instruction tuning [2]. Foundation models extend this logic by providing general-purpose representations that can be adapted to many downstream applications, including text classification, summarization, question answering, and code generation [3]. In financial risk analysis, such capabilities are attractive because much relevant information is textual, fragmented, and time-sensitive. Earnings calls, regulatory filings, news reports, analyst notes, and internal memos all contain signals that traditional numerical models may miss.

However, the literature on large language models also emphasizes their limitations. They can reproduce social stereotypes, generate false information, and obscure the provenance of their outputs [4]. Safety research has identified specification gaming, reward hacking, negative side effects, and distributional shift as recurring problems in artificial intelligence systems [6]. These issues are not incidental to financial deployment; they affect the reliability of risk

assessments, the fairness of credit or compliance decisions, and the credibility of explanations offered to regulators. Human-centered approaches argue that automation should be designed to enhance human control, encourage oversight, and support contestability rather than replace judgment [5]. In finance, this aligns with the principle that risk decisions must remain accountable to boards, regulators, and affected stakeholders.

Financial risk management has its own intellectual traditions. Operational risk frameworks emphasize process controls, incident reporting, business continuity, and the sound management of technology and human error [8]. Market risk and credit risk rely on quantitative models, stress testing, and scenario analysis, but these models are often fragile during regime shifts. Machine learning systems add technical debt through complex dependencies, data drift, and feedback loops that are difficult to detect [7]. The result is a hybrid domain in which generative agents must operate alongside deterministic models, legacy reporting systems, and human committees. The most promising research direction is therefore not isolated model improvement but architectural integration that makes agent behavior observable, bounded, and contestable.

3. System Architecture for Large Language Model Agents in Financial Risk

A practical architecture for large language model agents in financial risk analysis should be layered. At the base is a data and retrieval layer that connects structured market data, internal exposure systems, document repositories, and external news feeds. Above this sits a knowledge and memory layer that maintains entity definitions, temporal context, previous assessments, and institutional policies. The reasoning layer contains one or more large language model agents that interpret tasks, generate hypotheses, and select analytical tools. The execution layer invokes calculators, database queries, simulation engines, and reporting services. Finally, an oversight layer monitors outputs, enforces permissions, records provenance, and escalates uncertain or high-impact cases to human reviewers. This layered view prevents the agent from becoming an opaque monolith and creates insertion points for governance.

Within this architecture, multi-agent coordination can improve coverage but also introduces new failure modes. One agent may specialize in market news, another in counterparty exposure, another in regulatory compliance, and another in scenario construction. Trust-calibrated multi-agent decision intelligence under market regime shifts offers a relevant design principle: agents should not be treated as equally reliable across contexts, and their influence should be adjusted according to demonstrated competence, data freshness, and uncertainty [9]. Without such calibration, a fluent but poorly grounded agent may dominate a committee-style deliberation. The architecture must therefore distinguish between agents that generate candidate explanations and agents that verify, challenge, or constrain those explanations.

Tool use is another defining feature. Financial agents require access to pricing libraries, risk engines, macroeconomic databases, and document management systems. They may also need to write reports, open tickets, or trigger alerts, but such actions should be permissioned and reversible where possible. The design challenge is to balance autonomy with containment. An agent that can query a database is useful; an agent that can alter positions or submit regulatory filings without human approval is dangerous. Consequently, the architecture should support graduated autonomy, with low-risk summarization tasks operating with minimal oversight and high-impact decisions requiring multi-person review, audit trails, and explicit sign-off.

4. Data, Infrastructure, and Operational Sustainability

Data quality is the first determinant of agent reliability. Financial institutions already struggle with inconsistent identifiers, missing timestamps, stale reference data, and incompatible reporting standards. Large language model agents can amplify these problems because they may blend sources of different reliability without clearly signaling provenance. Regulatory and supervisory bodies have noted that artificial intelligence and machine learning can improve efficiency but also create new dependencies, concentration risks, and opacity in financial services [10]. A robust data layer must therefore include lineage tracking, versioning, conflict detection, and source authority rules. The agent should know not only what information exists but also how recent it is, who produced it, and what transformations it has undergone.

Unstructured data presents particular challenges. Sentiment extracted from news or filings can be informative, but financial markets are shaped by behavioral biases and narrative dynamics that are not captured by simple polarity scores [11]. Research on investor sentiment shows that media content can affect prices and trading activity, sometimes in ways that diverge from fundamentals [12]. Textual analysis of corporate disclosures has also demonstrated that domain-specific dictionaries and language models outperform generic approaches [13]. Surveys of textual sentiment in finance emphasize methodological choices, evaluation problems, and the difficulty of separating correlation from causation [14]. Large language model agents can incorporate these insights by combining sentiment measures with event detection, uncertainty estimates, and human review, but they must avoid presenting narrative summaries as objective risk truth.

Infrastructure and sustainability are often underestimated. Running large models at scale requires significant computation, energy, and specialized hardware. Financial institutions may choose between on-premise deployment, private cloud, or third-party application programming interfaces, each with different trade-offs in cost, latency, security, and vendor concentration. On-premise systems offer greater control but require substantial capital and expertise. Cloud services provide elasticity but may complicate data residency and regulatory reporting. Third-party model providers can accelerate innovation but introduce dependency on external update cycles and unclear data retention practices. Sustainability considerations include energy consumption, cooling requirements, and the opportunity cost of computational resources. A prudent institution will treat agent infrastructure as part of operational risk, not merely as an information technology expense.

5. Decision Support, Governance, and Fairness

Large language model agents should be positioned primarily as decision-support systems rather than autonomous decision-makers. Decision support requires that outputs be interpretable, comparable, and linked to evidence. An agent that produces a risk rating without explaining which indicators drove the rating, how confident it is, and what would change its conclusion is of limited value in a regulated setting. Governance frameworks for artificial intelligence emphasize transparency, accountability, fairness, and contestability [15]. These principles translate into concrete requirements: logging of prompts and retrieved evidence, explanations in domain language, mechanisms for appeal, and clear assignment of responsibility when an agent contributes to a decision.

Fairness is especially complex in financial risk. Credit, fraud detection, insurance, and compliance systems can produce disparate impacts across protected groups even when

sensitive attributes are excluded, because proxies and historical data carry structural inequalities. Large language models may introduce additional concerns through biased language, uneven performance across languages or dialects, and inconsistent treatment of similar cases. A governance program should therefore combine statistical fairness testing with qualitative review, stakeholder consultation, and domain-specific policy analysis. It is not enough to state that the model is fair in general; the institution must specify which fairness criteria matter for each use case and how trade-offs among them are resolved.

The distribution of responsibility is another central issue. If an agent summarizes a counterparty's filings, recommends an exposure limit, and drafts a committee memo, who is accountable when the recommendation proves harmful? A human-centered approach suggests that responsibility should remain with identifiable roles and committees, supported by agent documentation and audit trails [16]. Regulatory proposals increasingly require risk management systems, data governance, human oversight, and post-market monitoring for high-risk artificial intelligence applications [17]. In finance, these requirements intersect with existing model risk management, operational risk, and conduct rules. The most defensible arrangement is one in which the agent has clearly bounded authority, its recommendations are advisory unless explicitly approved, and its performance is continuously monitored against both technical and institutional metrics.

6. Robustness, Regime Shifts, and Systemic Risk

Robustness is the capacity of a system to maintain acceptable performance under perturbation, distributional change, and adversarial pressure. Financial markets are particularly challenging because they are adaptive, reflexive, and prone to regime shifts. A model trained during low-volatility conditions may fail during liquidity crises, and an agent optimized for historical patterns may misinterpret novel policy interventions. Multi-agent frameworks that enable conversation and role specialization can improve adaptability, but they can also create cascading errors if agents reinforce one another's mistaken assumptions [19]. Reasoning-and-acting architectures that interleave deliberation with tool use offer more grounded behavior, yet they still depend on the quality of the tools and the stability of the environment [20]. Robustness must therefore be tested through scenario analysis, red-teaming, stress testing, and controlled simulations.

Systemic risk adds a macroprudential dimension. If many institutions adopt similar agents trained on similar data and using similar risk signals, their responses may become correlated. Herding, crowded trades, and simultaneous deleveraging can amplify shocks even when each individual agent appears rational. Supervisory attention to artificial intelligence in finance has highlighted the possibility of new channels of contagion through common vendors, shared models, and automated execution [10]. A single institution may manage its own model risk, but systemic resilience requires coordination across firms, regulators, and infrastructure providers. This includes standards for incident reporting, disclosure of material model dependencies, and circuit breakers for automated actions during extreme stress.

Adversarial robustness is also essential. Financial agents may be targeted through data poisoning, prompt injection, document manipulation, or fake news designed to influence retrieval. An agent that treats all retrieved text as authoritative can be misled by strategically placed content. Defenses include provenance verification, source reputation scoring, anomaly detection, and separation between trusted internal data and open-source intelligence. Human escalation should be mandatory when evidence conflicts, when uncertainty exceeds thresholds,

or when the potential impact crosses predefined limits. In this sense, robustness is not only a technical property but an institutional capability to detect, contain, and learn from failure.

7. Evaluation, Deployment, and Policy Implications

Evaluation of large language model agents in finance cannot rely solely on benchmark accuracy. It must assess factual grounding, calibration, consistency, latency, cost, fairness, and auditability. Evaluation should be continuous because markets and data distributions change. Institutions need versioned test sets, shadow deployments, challenger models, and post-deployment monitoring. Metrics should distinguish between errors that are inconvenient and errors that are materially harmful. For example, a minor summarization error in an internal briefing may be tolerable, while a missed covenant breach or misclassified exposure may have regulatory and financial consequences. Evaluation criteria should therefore be tied to use-case criticality.

Deployment should proceed through staged autonomy. Initial applications might include document triage, entity extraction, and analyst assistance. More advanced applications might include scenario generation, portfolio commentary, and early-warning monitoring. Fully autonomous execution of trades or credit decisions should require extraordinary justification and extensive safeguards. A staged approach allows institutions to learn about failure modes, build trust, and adjust governance before expanding scope. It also supports regulatory dialogue, because supervisors can examine evidence from controlled deployments rather than debating hypothetical capabilities.

Policy implications extend beyond individual firms. Regulators may need to clarify how existing model risk, operational risk, and conduct rules apply to generative agents. They may also require documentation of training data, evaluation results, human oversight, and incident response. International coordination is important because financial markets are global and model providers may operate across jurisdictions. At the same time, overly rigid rules could freeze useful innovation or entrench incumbent vendors. A principles-based approach that focuses on outcomes such as accountability, transparency, fairness, and resilience may be more adaptable than prescriptive requirements tied to specific architectures. The goal should be to make agent behavior legible to institutions and supervisors without pretending that all risks can be eliminated.

8. Conclusion

Large language model agents offer a credible path toward more comprehensive and responsive financial risk analysis, but their value depends on system design rather than model novelty. The most important architectural choices concern retrieval, memory, tool permissions, multi-agent coordination, human escalation, and auditability. The most important governance choices concern accountability, fairness, transparency, and the boundaries of autonomous action. The most important infrastructural choices concern data lineage, vendor concentration, energy use, and operational resilience. These choices determine whether agents remain useful decision-support instruments or become new sources of opacity and systemic fragility.

The analysis suggests that financial institutions should treat large language model agents as bounded socio-technical systems. They should be embedded in workflows with clear roles, measurable performance criteria, and reversible actions. They should be calibrated for trust, meaning that influence is adjusted according to evidence quality, context, and demonstrated reliability. They should be evaluated continuously under stress, with attention to regime shifts

and adversarial manipulation. They should be governed by human accountability structures that preserve contestability and regulatory compliance. Under these conditions, large language model agents can improve the speed and breadth of risk analysis while supporting, rather than displacing, expert judgment.

Future research should examine cross-institutional contagion from shared agent infrastructure, methods for verifying provenance in agentic retrieval, and governance mechanisms for multi-agent deliberation. It should also explore how regulators can audit generative systems without requiring disclosure of proprietary model weights. Finally, it should investigate the labor and organizational implications of automated risk work, including changes in skill requirements, oversight burdens, and professional responsibility. The central challenge is not to build the most capable agent, but to build an institutional environment in which capable agents can be used safely, fairly, and accountably.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
3. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
4. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery.
5. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6), 495–504.
6. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
7. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28, 2503–2511.
8. Basel Committee on Banking Supervision. (2011). Principles for the sound management of operational risk. Bank for International Settlements.
9. Liu, T. (2026). Trust-Calibrated Multi-Agent Decision Intelligence for Financial Risk Management under Market Regime Shifts. Available at SSRN 7307503.
10. Financial Stability Board. (2017). Artificial intelligence and machine learning in financial services: Market developments and potential financial stability implications. Financial Stability Board.

11. Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–291.
12. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3), 1139–1168.
13. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1), 35–65.
14. Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33, 171–185.
15. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1).
16. Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
17. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final.
18. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0). U.S. Department of Commerce.
19. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2023). AutoGen: Enabling next-gen LLM applications via multi-agent conversation. arXiv preprint arXiv:2308.08155.
20. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.