



Large Language Models in FinTech: Applications in Financial Analysis, Customer Service, and Risk Management

Kennedy Ayanga

Department of Computer Science, Maseno University, P.O. Box 333 40105, Maseno, Kenya

*Corresponding author: kennedy.ayangal@maseno.ac.ke

Abstract

Large language models (LLMs) are changing how financial institutions and FinTech firms work with text, from earnings reports and regulatory filings to customer conversations and complaint records. Their promise is considerable: faster analysis, more responsive service, and earlier detection of emerging risks. Their risks are equally significant: fabricated facts, weak numerical reasoning, bias, data leakage, and actions no one can explain. This article examines LLM applications across three FinTech domains: financial analysis, customer service, and risk management. It proposes an architecture for deploying them responsibly. The architecture combines finance-adapted language models, retrieval-augmented generation (RAG) over institutional knowledge, guardrails and governance, scalable service delivery, configurable risk and decision workflows, and autonomous operations. Using a design-oriented approach grounded in ten studies published between 2019 and 2025, the article maps the evidence base, presents the architecture and the flow of a customer request, and classifies applications by LLM task, risk level, and required controls. It also sets out the main risks with their mitigations. It illustrates the architecture with a digital bank that uses LLMs in credit analysis, customer disputes, and complaint-based risk monitoring. It argues that LLMs create lasting value in FinTech when they are grounded in verified data, matched to the risk of each task, and embedded in governed workflows.

Keywords: large language models, FinTech, financial analysis, customer service, risk management, retrieval-augmented generation, financial NLP, AI governance, enterprise workflow, autonomous agents



1. Introduction

Finance runs on text as much as on numbers. Analysts read annual reports, earnings call transcripts, and news. Customer service teams handle millions of questions and complaints. Risk and compliance teams review policies, regulatory updates, incident reports, and communications. Until recently, most of this text could be analyzed only by people, or by narrow natural language processing (NLP) models built for single tasks such as sentiment classification (Araci, 2019).

LLMs have changed this. Models trained on large volumes of text can summarize, answer questions, extract information, classify, and draft across a wide range of tasks. Finance-specific models show that adapting LLMs to financial language improves performance on financial tasks. Examples include the proprietary BloombergGPT (Wu et al., 2023) and the open-source FinGPT (Yang et al., 2023). Research has also found that LLM assessments of news can predict stock returns (Lopez-Lira & Tang, 2023).

The limits are real. Numerical reasoning over financial reports remains difficult for language models (Chen et al., 2021). LLMs can produce confident but incorrect statements and raise wider ethical and social risks (Weidinger et al., 2021). In a regulated industry, these weaknesses matter. Mitigating them requires grounding outputs in verified institutional knowledge (Lewis et al., 2020), delivering services through scalable, reliable architecture (Sannidhanam, 2020), operating systems with bounded autonomy (Sannidhanam, 2025), and routing consequential decisions through governed workflows (Basireddy, 2022).

This article asks: How can LLMs be applied across financial analysis, customer service, and risk management in FinTech while controlling their risks? It contributes a mapping of the evidence base (Figures 1 and 2), a deployment architecture (Figure 3), a customer request flow (Figure 4), a classification of applications by risk and control, a risk and mitigation framework, and an illustrative application.

2. Related Work

2.1 Financial language models

Araci (2019) introduced FinBERT, a BERT model further trained on financial text. It improved financial sentiment classification over general models and showed the value of domain adaptation. Wu et al. (2023) presented BloombergGPT, a 50-billion-parameter model trained on a mix of financial and general-purpose data. It outperformed comparable general models on



financial NLP tasks while remaining competitive on general benchmarks. Yang et al. (2023) presented FinGPT, an open-source framework that emphasizes automated collection of financial data and lightweight adaptation of existing models. It offers a lower-cost alternative to training large proprietary models.

2.2 Financial analysis and reasoning

Chen et al. (2021) introduced FinQA, a dataset of questions that require numerical reasoning over financial reports combining text and tables. They found that the language models of the time fell well short of human experts. Calculating ratios, comparing periods, and chaining operations over tables remain hard for language models. Lopez-Lira and Tang (2023) used ChatGPT to assess whether news headlines were good or bad for companies and found that these assessments predicted subsequent daily stock returns. Predictive power was stronger for more capable models. Together, these studies show that LLMs can add value to financial analysis, especially in interpreting text, but need support for precise calculation.

2.3 Retrieval-augmented generation

Lewis et al. (2020) introduced RAG, which retrieves relevant documents and conditions generation on them. Outputs become more accurate and can be traced to their sources, and knowledge can be updated without retraining. In finance, retrieval can ground LLM outputs in the institution's own product terms, policies, customer records, and approved market data. This reduces fabrication and supports audit.

2.4 Risks of language models

Weidinger et al. (2021) set out a taxonomy of ethical and social risks from language models. It covers discrimination and toxicity, information hazards such as leaking private data, misinformation from false or misleading outputs, malicious uses, harms from human-like interaction, and environmental and socioeconomic harms. Many of these risks are acute in finance. Misinformation can mislead customers or investors, private data leakage can breach confidentiality, and discriminatory outputs can lead to unfair treatment.

2.5 Scalable service delivery

Sannidhanam (2020) proposes a scalable web services architecture for healthcare data



processing, using modular services, load distribution, and horizontal scaling to handle growing data and request volumes. Healthcare and finance share requirements for scale, reliability, and protection of sensitive data. The same approach can deliver LLM capabilities as scalable services that customer channels, analyst tools, and risk systems call through defined interfaces.

2.6 Autonomous operations

Sannidhanam (2025) examines autonomous AI agents for cloud infrastructure operations, which monitor systems, diagnose problems, and carry out corrective actions within defined limits under human oversight. LLM services have variable, bursty demand and depend on several components, including models, retrieval indexes, and guardrail services. Autonomous agents can keep these services responsive and healthy.

2.7 Risk and decision workflows

Basireddy (2022) proposes configurable enterprise workflow architecture for digitizing risk management, in which identification, assessment, approval, escalation, and remediation are defined as configurable workflows with audit trails. LLM outputs that affect customers, credit, or compliance should enter such workflows, so that people with the right authority review consequential decisions and every step is recorded.

2.8 Research gap

These strands have developed separately, leaving limited guidance on deploying LLMs across FinTech domains with controls matched to the risk of each application.

3. Research Methodology

The study follows a design-oriented research approach. The problem, that LLMs offer major benefits in FinTech but carry risks that current deployments often do not manage systematically, was identified from the literature. Requirements were derived from the reviewed studies: finance-appropriate models, grounding in verified knowledge, support for numerical accuracy, risk-matched controls, scalable delivery, reliable operations, and governed decisions. The architecture was designed from ten studies and demonstrated through an illustrative digital bank case.

To make the evidence base transparent, each study was assigned to the theme it primarily supports. Figure 1 shows the distribution across themes, and Figure 2 shows the studies by publication year. These figures describe the literature used in this article. They are not performance measurements. The architecture has not been evaluated empirically, which is left to future work.

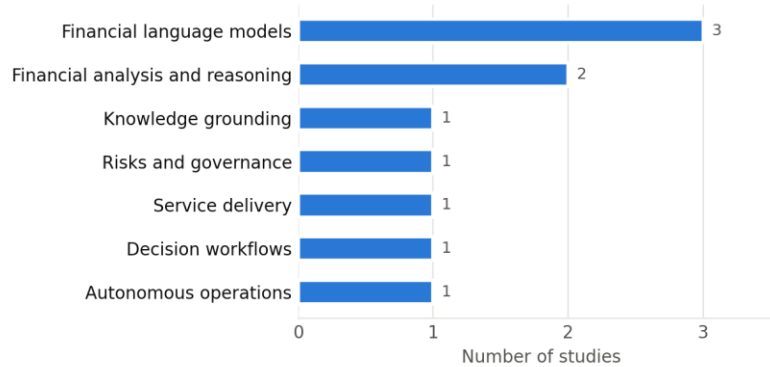


Figure 1. Number of reviewed studies primarily supporting each theme (n = 10)

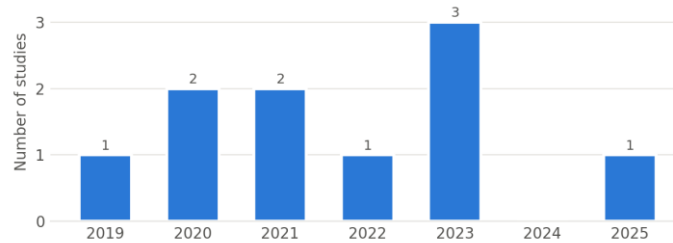


Figure 2. Reviewed studies by publication year (n = 10)

Financial language models are the most developed theme, and most studies date from 2019 onward, with a cluster in 2023 following the rapid growth of LLMs. The deployment themes, including grounding, governance, service delivery, workflows, and operations, each rest on a single study. This suggests that research on building finance-specific models has advanced faster than research on deploying them responsibly.

4. Architecture

The architecture has three application domains built on a shared LLM platform and supported by delivery, workflow, and operations layers (Figure 3 and Table 1).

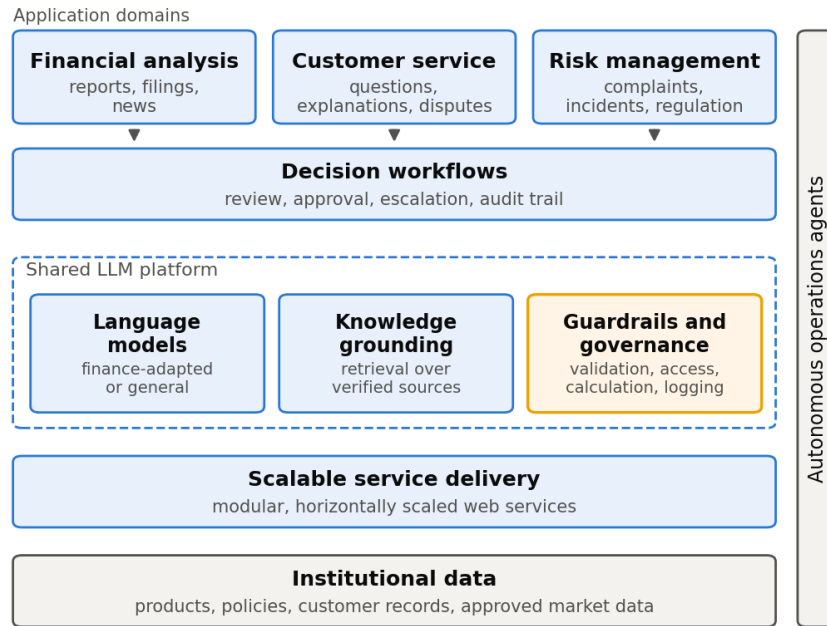


Figure 3. Architecture for deploying LLMs across FinTech domains

Table 1. Layers of the FinTech LLM architecture

Layer	Function	Key mechanisms	Supporting studies
1. Application domains	Deliver LLM capabilities to users	Financial analysis, customer service, and risk management applications	Lopez-Lira & Tang (2023); Chen et al. (2021)
2. Language models	Understand and generate financial text	Finance-adapted or general LLMs chosen per task	Araci (2019); Wu et al. (2023); Yang et al. (2023)
3. Knowledge grounding	Ground outputs in verified information	RAG over policies, products, customer records, approved data	Lewis et al. (2020)
4. Guardrails and governance	Control risks	Access control, output validation, calculation checks, bias testing, logging	Weidinger et al. (2021)
5. Service delivery	Serve LLM capabilities at scale	Modular, horizontally scalable web services	Sannidhanam (2020)
6. Decision workflows	Route consequential outputs for approval	Configurable workflows, escalation, audit	Basireddy (2022)
7. Autonomous operations	Keep services healthy	Bounded monitoring and scaling agents	Sannidhanam (2025)

Language models. Different tasks call for different models. A finance-adapted model may be used where financial language matters most (Wu et al., 2023; Yang et al., 2023). Smaller specialized models may handle high-volume classification such as sentiment or intent (Araci, 2019), and general models may handle open-ended conversation.

Knowledge grounding. For any output that states facts about products, customers, policies, or

markets, the model is given retrieved information from verified institutional sources and must cite them (Lewis et al., 2020).

Guardrails and governance. Guardrails address the risks identified by Weidinger et al. (2021). They enforce access permissions on retrieved data, validate outputs, route numerical work to calculation tools rather than relying on the model's arithmetic (Chen et al., 2021), test for bias, and log every interaction.

Service delivery. LLM capabilities are exposed as modular services, such as summarization, question answering, classification, and drafting, that scale horizontally with demand (Sannidhanam, 2020).

Decision workflows. Outputs that affect customers, credit, or compliance enter configurable workflows for review, approval, or escalation (Basireddy, 2022).

Autonomous operations. Agents monitor latency, error rates, and demand across model, retrieval, and guardrail services. They scale and restart services within approved limits and escalate other issues (Sannidhanam, 2025).

5. Applications

5.1 Financial analysis

LLMs can summarize earnings calls and annual reports, extract key figures and risk disclosures, compare companies, and assess sentiment in news and filings (Araci, 2019; Wu et al., 2023; Lopez-Lira & Tang, 2023). Because numerical reasoning over reports is error-prone (Chen et al., 2021), the architecture has the model extract figures and describe the calculation required. A calculation tool performs the arithmetic, and the results are shown with links to the source pages so analysts can verify them.

5.2 Customer service

LLMs can answer customer questions, explain products and transactions, help with disputes, and draft responses for human agents. Figure 4 shows the flow of a customer request.

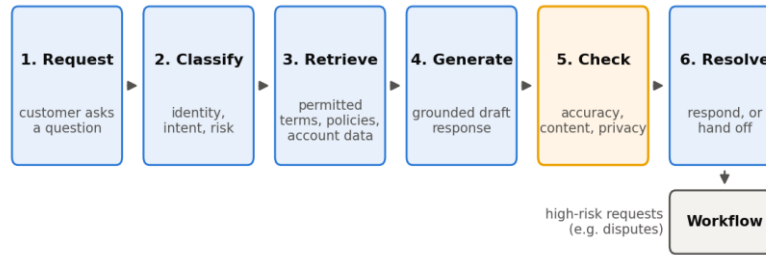


Figure 4. Flow of a customer request through the FinTech LLM architecture

The request is first authenticated and classified by intent and risk. Relevant product terms, policies, and the customer's own account information are then retrieved, limited to what that customer is permitted to see (Lewis et al., 2020). The model generates a response grounded in this information. Guardrails check the response for accuracy against the retrieved data, prohibited content such as unauthorized financial advice, and exposure of sensitive data (Weidinger et al., 2021). The system then either responds directly or hands the case to a human agent or a workflow, for example to open a dispute (Basireddy, 2022).

5.3 Risk management

LLMs can help risk and compliance teams monitor complaints and communications for emerging issues, summarize incidents, map new regulations to internal policies and controls, and draft risk reports. Classified complaints and flagged communications enter risk workflows (Basireddy, 2022), where analysts confirm issues and assign remediation.

5.4 Matching controls to risk

Applications differ greatly in their potential for harm. Table 2 classifies representative applications by LLM task, risk level, and required control.

Table 2. FinTech LLM applications by task, risk level, and required control

Domain	Application	LLM task	Risk level	Required control
Financial analysis	Earnings call summary for analysts	Summarization	Moderate	Source links, analyst review
Financial analysis	Ratio and trend extraction	Extraction and calculation	High	Calculation tools, verification against source
Customer service	Product and fee questions	Grounded question answering	Moderate	Retrieval from approved terms, output validation



Customer service	Dispute handling	Classification and drafting	High	Human agent or workflow for decisions
Risk management	Complaint theme detection	Classification and clustering	Moderate	Analyst confirmation of themes
Risk management	Regulation-to-policy mapping	Retrieval and comparison	High	Compliance officer sign-off

Moderate-risk applications can run with automated checks and sampled review. High-risk applications, where errors could harm customers or breach obligations, require human decision-making through workflows.

6. Risks and Mitigations

Drawing on Weidinger et al. (2021) and the financial context, Table 3 sets out the main risks and their mitigations.

Table 3. Main LLM risks in FinTech and their mitigations

Risk	Example in FinTech	Mitigation	Supporting studies
Fabricated information	Incorrect fee or interest rate stated to a customer	Retrieval from approved sources, citation, output validation	Lewis et al. (2020); Weidinger et al. (2021)
Numerical errors	Wrong ratio in a credit memo	Calculation tools, verification against filings	Chen et al. (2021)
Data leakage	Another customer's details in a response	Permission-aware retrieval, sensitive data filtering	Weidinger et al. (2021)
Discrimination	Different treatment of customers by group	Bias testing, monitoring of outcomes by segment	Weidinger et al. (2021)
Unreviewed consequential actions	Dispute resolved or account changed without authority	Workflow approvals, escalation, audit trail	Basireddy (2022)
Service failure under load	Assistant unavailable during a market event	Scalable services, autonomous scaling within limits	Sannidhanam (2020, 2025)
Overreliance on market signals	Trading on LLM sentiment without validation	Backtesting, human oversight of investment decisions	Lopez-Lira & Tang (2023)

7. Illustrative Case: A Digital Bank

Consider a digital bank that deploys LLMs in its credit, customer service, and risk teams on a shared platform.

Financial analysis. A credit analyst reviewing a small business loan asks the analysis assistant



to summarize the applicant's latest annual report and compute its debt service coverage. The assistant extracts the relevant figures and states the formula. A calculation tool computes the ratio, and each figure links back to its page in the report (Chen et al., 2021; Wu et al., 2023). The analyst spots that one figure came from a restated section, corrects it, and recalculates.

Customer service. A customer asks the chat assistant why a card payment was declined and says a charge on their statement is not theirs. The request is classified as an explanation plus a potential dispute (Figure 4). The assistant retrieves the decline reason and the dispute policy, limited to this customer's own data (Lewis et al., 2020). It explains that the payment was declined because a daily limit had been reached. Guardrails confirm that the explanation matches the account records. Because disputes carry high risk (Table 2), the assistant does not resolve the disputed charge itself. It opens a dispute case in the workflow, with a summary of the conversation attached for a human agent (Basireddy, 2022).

Risk management. Each night, the complaint monitoring service classifies new complaints by theme. It detects a sharp rise in complaints about a recently changed overdraft fee, many saying the change was not clearly communicated. The system opens a conduct risk case in the risk workflow with representative examples and a draft summary (Basireddy, 2022). A risk analyst confirms the issue, and the bank reviews its communication and considers refunds.

Operations. On the day the bank announces results, customer inquiries and analyst queries both surge. Operations agents scale the model and retrieval services within approved limits and keep response times stable (Sannidhanam, 2020, 2025).

Governance. Every interaction is logged with retrieved sources and guardrail results. Quarterly bias testing compares assistant outcomes across customer segments (Weidinger et al., 2021).

The case shows how one platform serves three domains, with controls matched to the risk of each task and consequential decisions kept with accountable people.

8. Discussion

8.1 Implications

For research, the article brings together work on financial language models, financial reasoning, retrieval, LLM risks, and enterprise architecture into a single deployment view. The evidence mapping shows that model development has advanced faster than research on responsible deployment, pointing to a need for empirical work on grounding, governance, and workflow

integration in finance. For practitioners, the priorities are to ground factual outputs in verified sources, route calculations to tools, and match controls to the risk of each application.

8.2 Limitations and future work

The architecture was demonstrated through an illustrative case and draws on a focused set of ten studies, and the evidence mapping reflects only these studies. LLM capabilities are also changing quickly, so specific performance findings may date. Future work should evaluate the architecture in practice, measuring factual accuracy, calculation errors, resolution rates and handoff rates in customer service, time to detect emerging risks, and fairness across customer groups. Further research is needed on evaluation benchmarks for financial LLM applications, on regulatory expectations for LLM use in consumer finance, and on how customers and staff calibrate trust in LLM assistance.

9. Conclusion

LLMs offer FinTech firms and financial institutions new ways to analyze information, serve customers, and manage risk. This article proposed an architecture that combines finance-appropriate models, retrieval grounding, guardrails, scalable service delivery, governed workflows, and autonomous operations. It showed how applications in each domain can be classified by risk and given matching controls. The digital bank case showed LLMs speeding up credit analysis, explaining customer issues while escalating disputes, and detecting a conduct risk early. The central conclusion is that LLMs create durable value in FinTech when their outputs are grounded in verified data, their controls are matched to the risk of each task, and consequential decisions stay within governed, accountable workflows.

9. References

1. Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models (arXiv:1908.10063). arXiv.
2. Basireddy, S. R. (2022). Digitizing risk management through configurable enterprise workflow architecture. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 14(4), 231–239.
3. Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., ... Wang, W. Y. (2021). FinQA: A dataset of numerical reasoning over financial data. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 3697–3711).



4. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
5. Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models (arXiv:2304.07619). arXiv.
6. Sannidhanam, A. H. (2020). Scalable web services architecture for healthcare data processing. *International Journal of Recent Advances in Science and Technology*, 7(01), 1–14. <https://ijrast.in/index.php/ijrast/article/view/65>
7. Sannidhanam, A. H. (2025). Autonomous AI agents for cloud infrastructure operations. *Journal of Contemporary Science and Technology Management*, 1(01), 83–100. <https://jcstm.info/index.php/jcstm/article/view/20>
8. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., ... Gabriel, I. (2021). Ethical and social risks of harm from language models (arXiv:2112.04359). arXiv.
9. Wu, S., Irsoy, O., Lu, S., Dabrovolski, V., Dredze, M., Gehrmann, S., ... Mann, G. (2023). BloombergGPT: A large language model for finance (arXiv:2303.17564). arXiv.
10. Yang, H., Liu, X.-Y., & Wang, C. D. (2023). FinGPT: Open-source financial large language models (arXiv:2306.06031). arXiv.