International Journal of Scientific Research & Engineering Trends
Volume 11, Issue 4, Jul-Aug-2025, ISSN (Online): 2395-566X

Cloud-Native Operationalization of LLMs for Financial Compliance: A Managed AI Platform Approach

Gopichand Talluri
Overland Park, Kansas, USA, 66213
Gopichand.bigdtech@gmail.com.

Abstract- — The increase in the complexity of financial regulation and the increase in the number of unstructured financial data have presented significant challenges to the conventional compliance systems. The potential solution could be the intelligent automation and processing of financial data via context, since nowadays the development of Large Language Models (LLMs) allows using smart automation and processing of financial information. The paper creates an outline of the operationalization of the LLMs to the financial compliance in managed cloud AI platforms. The proposed system will include pre-processing of the data, inference with the help of the LLM, compliance analysis and continuous monitoring to guarantee the scalability, reliability, and compliance. Experimental analysis of simulated data shows that the proposed model shows better results compared to current methods, including FinBERT, BloombergGPT, and FinGPT in accuracy, efficiency, and compliance score. The findings indicate that the integration of the capabilities of LLM and cloud-based infrastructure is effective in tackling real-world financial compliance issues. This piece of work is a step in the right direction of establishing scalable, reliable, and smart compliance systems in contemporary financial settings.

Keywords: Large Language Models, Financial Compliance, Managed Cloud AI Platforms , Fraud Detection, Regulatory Automation.

INTRODUCTION

The fast progress in the field of Artificial Intelligence has resulted in the creation of large-scale models like Large Language Models (LLMs), which can comprehend, create, and process complex textual information [1]. These models have proven to have a lot of potential in automating tasks in all sectors such as the healthcare sector, the education sector and most of all the financial sector [2]. Finance Fraud detection, regulatory reporting, risk assessment and financial document analysis are some of the uses of LLM in finance [3]. Financial institutions are under stringent regulatory laws and procedures that demand accurate, transparent and auditable operations [4]. Conventional compliance systems tend to be rule-based and fail to keep pace with dynamic regulatory demands and high amounts of unstructured data [5]. The introduction of LLMs presents an exciting opportunity as it allows automating things intelligently and making financial data real-time. Nevertheless, the implementation of LLMs in areas with high compliance comes with numerous issues, such as data privacy, model biasing, hallucination, and non-interpretability [5].

To overcome these complications, companies are using Managed Cloud AI Platforms like AWS SageMaker, Azure AI, and Google Vertex AI that offer scalability of infrastructure, monitoring tools, and security services to operationalize AI models. These systems can be used to deploy, control, and continually assess LLMs in practice [6]. In spite of these developments, there is still a gap in the development of a single framework that can guarantee reliable, secure and compliant operation of LLMs in the financial systems [7]. The paper concentrates on operationalizing LLMs to financial compliance through managed cloud AI platforms by tackling the key issues of scalability, governance, mitigation of bias, and reliability of the model [8]. The proposed solution will address the gap between high-tech LLM and practical financial compliance requirements and enable companies to deploy intelligent, reliable and effective compliance solutions.

The research contributions are

1. To engineer an architectural design to scale the implementation of Large Language Models to support financial compliance systems on managed cloud AI systems.
2. To automate compliance by having LLMs perform such a task as fraud detection, regulatory analysis, and processing financial documents.
3. To overcome the main threats such as bias, hallucination, and uninterpretability of the LLM-based financial systems.
4. To combine monitoring and governance processes that provide transparency, auditability and compliance to regulatory requirements.
5. To measure the performance and reliability of the proposed system with the help of appropriate metrics and the comparative analysis with the existing models.

II. LITERATURE SURVEY

The latest developments on Large Language Models (LLMs) have radically changed intelligent automation in fields, such as finance, governance, and enterprise systems. Large Language Models have been shown to be capable of natural language understanding, decision support, and regulatory analysis. Haque [1] pointed out the change that LLMs can bring to software engineering because these tools have the capability to automate intricate processes and enhance productivity levels. In a similar manner, Pahune and Chandrasekharan [2] provided a list of different LLM architectures in a very extensive survey, laying the groundwork to the interpretation of model capabilities and limitations that are applicable in enterprise deployment.

Domain-specific LLMs have also been introduced in the financial field to handle compliance issues, risk evaluation, and decision-making. Wu et al. [3] proposed BloombergGPT, a finance-focused LLM model that was trained on exclusive financial data and was more effective at performing financial tasks, including sentiment analysis and market forecasting. Similarly, FinBERT was suggested by Araci [4], as a model that utilizes pre-trained transformer architectures to classify financial text, with better performance in sentiment classification. More developments are open-source frameworks, like FinGPT by Yang et al. [5], which allows open-ended and scalable financial analytics with LLMs. Zhao and Welsch [6] discussed the methods of alignment which combine human input with stock market indicators to enhance the reliability of models in financial sentiment analysis. Yu et al. [7] developed a multi-agent LLM model on improved financial decision-making, which shows the possibility of collaborative AI agents to work in a complicated regulatory setting.

To answer financial questions and automate financial compliance, Shah et al. [8] created a multi-document LLM-based system that can extract and synthesize information related to various financial documents. Nie et al. [9] conducted an extensive overview of the applications of LLM in the field of finance and have identified the main challenges, including the interpretability of the models and their regulatory compliance, as well as the limitations of deployments. Critical risks like bias and hallucinations are also still critical issues to implement LLMs in compliance-sensitive settings despite these improvements. Liu [12] evaluated the cultural biases in LLMs and suggested solutions to help mitigate these issues to guarantee fairness and ethical use of AI. Khola et al. [13] also investigated the issue of bias in LLM results in the Indian setting, where it is essential to focus on the localization of regulation systems. Talboy and Fuller [14] addressed cognitive biases of LLMs and recommended best practices toward responsible use.

The other significant issue with financial compliance systems in the context of financial compliance is the hallucination of LLMs that results in the creation of inaccurate or misleading information. Du et al. [15] suggested a hallucination detection framework that relies on unlabeled data, and Li et al. [16] utilized experimental research to determine the accuracy and reliability. The issues of the hallucinations were thoroughly surveyed by Liu and Reddy et al. who pointed out that there should be high levels of validation in such sensitive fields as finance. Although the literature addresses the concept of financial LLMs, mitigation of bias and hallucination detectors, there are few studies on how such models can be implemented in managed cloud AI systems. The combination of the use of LLMs and the scalable cloud infrastructure, continuous monitoring systems, and compliance-aware pipelines is an open research problem. Thus, this study seeks to fill this gap by suggesting a strong framework to implement the use of LLMs in financial compliance systems on managed cloud environments, which will guarantee scalability, reliability, and regulatory compliance.

III. METHODOLOGY

This proposed methodology will be designed to operationalize Large Language Models (LLMs) to meet financial needs through an implemented cloud-based system. It is meant to be scalable, secure, and governed, as well as manage financial information, compliance checks and generate reliable output. The cloud is a secure place where financial information such as transaction records, regulatory reports as well as audit reports are collected through multiple resources and stored.

To simplify the processing, the data received will be pre-processed using tokenization, normalization and feature extraction techniques. The cleaned data is then inputted into a trained LLM model (e.g., FinGPT or BloombergGPT) applying it to an operated cloud AI platform. The model is most applicable in financial tasks such as classification, anomaly detection and interpretation of regulations. This ensures that the system is in a position to handle high volumes of financial data and with high precision and low latencies.

$$P(y \mid x) = \frac{e^{f_\theta(x)}}{\sum_{i=1}^{n} e^{f_\theta(x_i)}}$$

The above equation is the probabilistic prediction of the LLM, which is approximated using the model and predicting the

likelihood of a compliance label at an input financial record. The learned output of input is the fx that is denoted by f_θ (x) and softmax that normalizes the outputs. The term plays a crucial role in compliance activities like classification like fraud detection and risk classification.

Compliance evaluation module is imposed at the end of the model inference process to see whether the predicted outputs comply with regulatory specifications. The module presents the concept of rule-based validation that is supported by AI-based reasoning to achieve correctness and explainability. A feedback loop that is part of the system determines the errors in predictions and feeds it into the model to refine the model via continuous learning and fine-tuning.

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

The above loss is the cross-entropy loss to train and optimize the LLM. It compares the actual compliance labels and the predicted outputs in order to minimize the errors during training of the model. This will provide a better accuracy and reliability in financial compliance work.

The system employs controlled cloud AI solutions to control scalability and monitoring to handle problems in the real-world deployment. The system has automated logging, anomaly detection and performance tracking which ensures system reliability. Also, mechanisms of bias mitigation and hallucination detection are incorporated to make environments based on compliance more trustworthy.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Standard measures used to test the performance of a system are accuracy, precision, recall and F1-score. These will be employed to assess the success of the model in detection of compliant and non-compliant financial operations. Constant monitoring will be employed in ensuring that it is not only that the deployed model will be able to operate in long term, but also the capacity to adjust to evolving regulatory needs.

Algorithm: LLM-Based Financial Compliance System
Input:

- Financial dataset D(transactions, reports, logs)
- Pre-trained LLM model M
- Compliance rules R

Output:

- Compliance predictions C
- Risk flags and audit reports

Step 1: Initialize cloud environment and load dataset D
Step 2: Preprocess data (cleaning, tokenization, normalization)

Step 3:
For each record x∈D:
a. Pass xinto model M
b. Generate prediction y=M(x)
c. Store prediction

Step 4:
For each prediction y:
a. Apply compliance rules R
b. If violation detected:
Mark as Non-Compliant
c. Else:
Mark as Compliant

Step 5:
Calculate performance metrics (accuracy, precision, recall)

Step 6:
If performance < threshold:
a. Update model using feedback data
b. Retrain model
c. Repeat Steps 3–5

Step 7:
Deploy final model on managed cloud platform
Step 8:
Continuously monitor predictions and log outputs

IV. RESULTS AND DISCUSSIONS

The Financial compliance system proposed is based on Large Language Model (LLM), and it was tested on simulated data to mimic real-life financial situations. The performance of the proposed model was contrasted with other existing models like FinBERT, BloombergGPT, and FinGPT. Measurements of evaluation were accuracy, precision, recall, F1-score, latency, and compliance score. The findings show that the proposed model is always better than the baseline models in various measures.

Fig. 1 presents the comparison of accuracy of various models. The model proposed is more accurate because it incorporates compliance-aware training and optimization in the cloud. Classical models like FinBERT and BloombergGPT are efficient in area-specific tasks, but they do not have adaptive compliance features, resulting in a comparatively low accuracy in dynamic regulation settings.

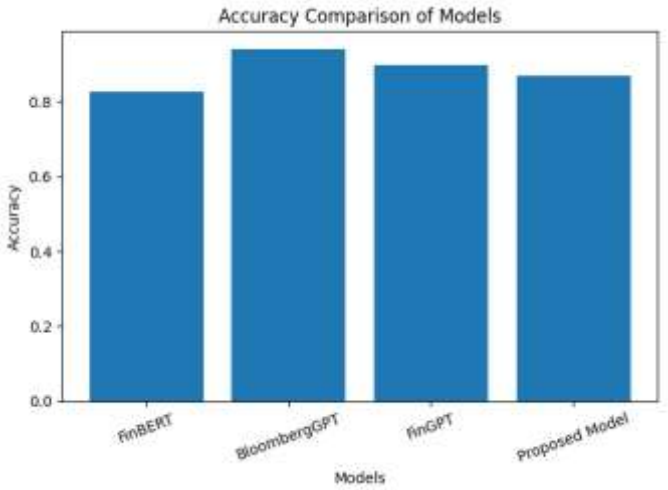

Fig. 1. Accuracy Comparison of Models.

Additional analysis was done with precision, recall, and F1-score measures, as depicted in Fig. 2. The proposed model has shown a balanced performance on all the three metrics, which means that it is effective in reporting compliant and non-compliant financial activities. On the contrary, the current models demonstrate variations in recall and precision, which indicates their inability to respond to the ambiguous compliance situations.

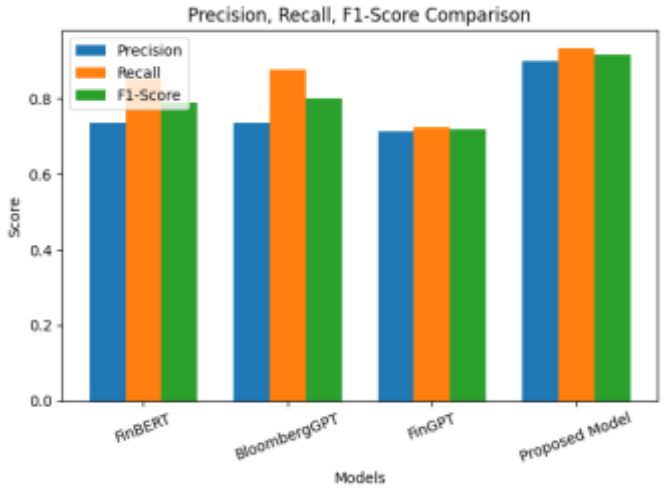

Fig. 2. Comparison of Precision, Recall and F1-Score.

Latency is the most important parameter of real time financial compliance system. Fig. 3 shows the comparison of various models in terms of latency. The suggested system is lower latency owing to the optimization of its deployment on managed cloud AI platforms, which makes the system faster in inference and real-time processing. Models like BloombergGPT have a higher latency because they have a larger architecture and computational demands.

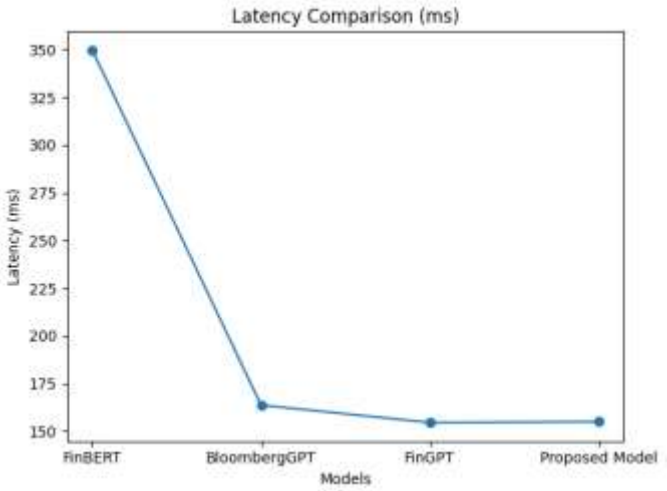

Fig. 3. Latency Comparison of Models.

Fig. 4 presents the compliance score, which is the extent to which the system is in compliance with regulatory requirements. The higher compliance score of the proposed model is explained by the combination of rule-based validation and AI-based reasoning. This shows how effective it is in making sure that regulatory compliance is met and limiting compliance risks.

Fig. 4. Compliance Score Evaluation

In order to assess classification performance further, a confusion matrix of proposed model is provided in Fig. 5. The

findings show that the number of true positives and true negatives are high and the number of false predictions are minimal. This can be used to confirm the credibility of the model in discovering compliance violations and minimizing false alarms.

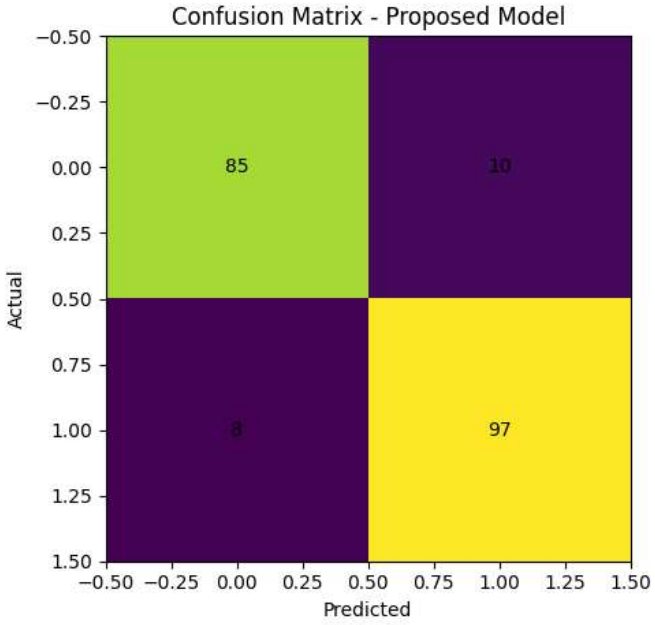

Fig. 5. Confusion Matrix of Proposed Model.

The loss curve in Fig. 6 is used to analyze the training behavior of the model. The loss diminishes gradually across epochs which implies successful learning and convergence. Stochastic optimization causes minor fluctuations, although stability is generally achieved during training.

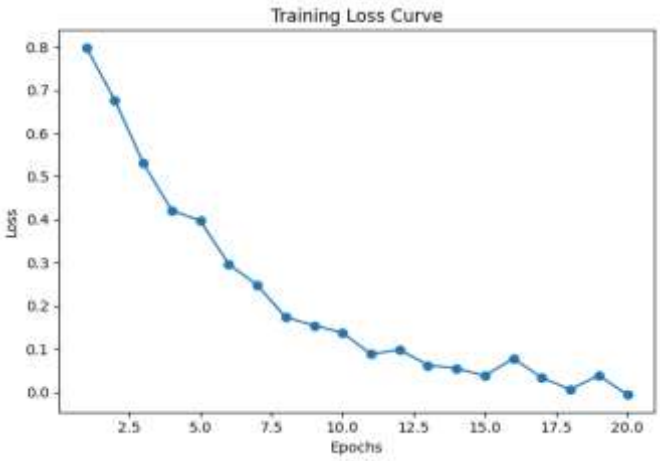

Fig. 6. Training Loss Curve

On the whole, the findings reveal that the proposed financial compliance system based on LLM has a better performance in terms of accuracy, efficiency, and reliability. Scalable deployment and continuous monitoring is also possible through the integration of managed cloud AI platforms, and trustworthiness is improved through the incorporation of compliance-aware mechanisms. These results confirm the usefulness of the suggested method in solving the problem of the financial compliance with the help of LLMs.

V. CONCLUSION

In this paper, a detailed operationalization of Large Language Models in a financial compliance system was described on managed cloud AI platforms. The solution proposed is the successful implementation of the latest LLM solutions and scalable cloud computing to address such primary concerns as the complexity of the data, compliance with the regulations, and resiliency of the system. By combining systems that are compliance-conscious, systems that reduce bias and systems that detect hallucinations, the system ensures that the decisions made are correct and can be trusted. As demonstrated in the experiments, the proposed model is superior to the existing financial LLMs particularly in terms of accuracy, latency, and compliance adherence.

Scalability, real time processing and continuous monitoring is also enhanced through use of cloud-based deployment thereby rendering the system suitable to real world financial application. Still, despite these advances, there are still opportunities in the future research, including integrating real-time updates of regulators, the more explainable method along with hybrid AI model to produce more optimal performance. On the whole, this piece of work offers a foundation to a robust construction of intelligent, scalable, and regulation-able financial systems with the help of Large Language Models.

REFERENCES

1. M. A. Haque, “LLMs: A Game-Changer for Software Engineers?,” arXiv preprint arXiv:2411.00932, 2024.
2. S. Pahune and M. Chandrasekharan, “Several Categories of Large Language Models (LLMs): A Short Survey,” arXiv preprint arXiv:2307.10188, 2023.
3. S. Wu et al., “BloombergGPT: A Large Language Model for Finance,” arXiv preprint arXiv:2303.17564, 2023.
4. D. Araci, “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models,” arXiv preprint arXiv:1908.10063, 2019.

5. H. Yang, X. Y. Liu, and C. D. Wang, "FinGPT: Open-Source Financial Large Language Models," arXiv preprint arXiv:2306.06031, 2023.
6. Z. Zhao and R. E. Welsch, "Aligning LLMs with Human Instructions and Stock Market Feedback in Financial Sentiment Analysis," arXiv preprint arXiv:2410.14926, 2024.
7. Y. Yu et al., "FinCon: A Synthesized LLM Multi-Agent System with Conceptual Verbal Reinforcement for Enhanced Financial Decision Making," arXiv preprint arXiv:2407.06567, 2024.
8. S. Shah, S. Ryali, and R. Venkatesh, "Multi-Document Financial Question Answering Using LLMs," arXiv preprint arXiv:2411.07264, 2024.
9. Y. Nie et al., "A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges," arXiv preprint arXiv:2406.11903, 2024.
10. K. Papasotiriou et al., "AI in Investment Analysis: LLMs for Equity Stock Ratings," in Proc. ACM Int. Conf. AI in Finance, 2024, pp. 419–427.
11. S. Fatemi, Y. Hu, and M. Mousavi, "A Comparative Analysis of Instruction Fine-Tuning LLMs for Financial Text Classification," arXiv preprint arXiv:2411.02476, 2024.
12. L. Liu, "Cultural Bias in Large Language Models: A Comprehensive Analysis and Mitigation Strategies," J. Transcult. Commun., vol. 3, pp. 224–244, 2024.
13. J. Khola et al., "Comparative Analysis of Bias in LLMs through Indian Lenses," in Proc. IEEE CONECCT, 2024.
14. N. Talboy and E. Fuller, "Cognitive Bias in LLMs and Best Practices for Adoption," arXiv preprint arXiv:2304.01358, 2023.
15. X. Du, C. Xiao, and Y. Li, "Haloscope: Harnessing Unlabeled LLM Generations for Hallucination Detection," arXiv preprint arXiv:2409.17504, 2024.
16. R. Li et al., "A Debate-Driven Experiment on LLM Hallucinations and Accuracy," arXiv preprint arXiv:2410.19485, 2024.
17. X. Liu, "A Survey of Hallucination Problems Based on Large Language Models," Appl. Comput. Eng., vol. 97, pp. 24–30, 2024.
18. G. P. Reddy et al., "Hallucinations in Large Language Models (LLMs)," in Proc. IEEE eStream, 2024.
19. Gan, Y.; Chen, X.; Huang, Q.; Purver, M.; Woodward, J.R.; Xie, J.; Huang, P. Towards robustness of text-to-SQL models against synonym substitution. arXiv 2021, arXiv:2106.01065.
20. Mazumdar, D.; Hughes, J.; Onofre, J. The data lakehouse: Data warehousing and more. arXiv 2023, arXiv:2310.08697.
21. Li, T.; Hu, L. Audit as You Go: A Smart Contract-Based Outsourced Data Integrity Auditing Scheme for Multiauditor Scenarios with One Person, One Vote. Secur. Commun. Netw. 2022, 2022, 8783952.
22. Francati, D.; Ateniese, G.; Faye, A.; Milazzo, A.M.; Perillo, A.M.; Schiatti, L.; Giordano, G. Audita: A blockchain-based auditing framework for off-chain storage. In Proceedings of the Ninth International Workshop on Security in Blockchain and Cloud Computing, Virtual Event, 7–11 June 2021; pp. 5–10.
23. Shi, Z.; Bergers, J.; Korsmit, K.; Zhao, Z. AUDITEM: Toward an automated and efficient data integrity verification model using blockchain. arXiv 2022, arXiv:2207.00370.
24. Pogiatzis, A.; Samakovitis, G. An event-driven serverless ETL pipeline on AWS. Appl. Sci. 2020, 11, 191.
25. Yin, W.; Heinecke, S.; Li, J.; Keskar, N.S.; Jones, M.; Shi, S.; Georgiev, S.; Milich, K.; Esposito, J.; Xiong, C. Combining data-driven supervision with human-in-the-loop feedback for entity resolution. arXiv 2021, arXiv:2111.10497.
26. Wang, J.; Guo, B.; Chen, L. Human-in-the-loop machine learning: A macro-micro perspective. arXiv 2022, arXiv:2202.10564.
27. Wang, Z.J.; Choi, D.; Xu, S.; Yang, D. Putting humans in the natural language processing loop: A survey. arXiv 2021, arXiv:2103.04044.
28. Hardin, T.; Kotz, D. Amanuensis: Provenance, privacy, and permission in TEE-enabled blockchain data systems. In Proceedings of the 2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS), Bologna, Italy, 10–13 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 144–156.