Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 03, March-2025

Generative AI for Dynamic Financial Planning and Risk Profiling

Sandeep Chitalkar
Dr. D. Y. Patil Institute of Technology
Savitribai Phule Pune University
Pune, India

Sarthak Pande
Dr. D. Y. Patil Institute of Technology
Technology
Savitribai Phule Pune University
Pune, India

Gayatri Revanwar
Dr. D. Y. Patil Institute of
Savitribai Phule Pune University
Pune, India

Aniket Yelmalwar
Dr. D. Y. Patil Institute of
Technology Savitribai Phule Pune
University Pune, India

Arsha J Cherian
Dr. D. Y. Patil Institute of
Technology Savitribai Phule Pune
University Pune, India

***Abstract*—This paper presents a Generative AI-powered financial planning model that delivers personalized strategies to help users achieve their financial goals. The system analyses comprehensive user inputs such as income, dependents, assets, liabilities, expenses, and financial goals and generates customized financial plans. The model analyses the risk, dynamically adjusting the recommendations based on the user's financial profile. Finetuning is done on LLaMA 3.2 using LoRA and QLoRA by using a custom dataset of user inputs and corresponding outputs. In addition, the model integrates RAG to make use of real-time market data to keep the recommendations current and relevant.**

A conversational chatbot enhances user interaction with it by answering questions about investments, loans, by analysing financial health of users. The system, leveraging the LSTM model, gives appropriate stocks and mutual funds suitable for the risk profile determined by the system, thus breaking down in helping the loan decision according to an analysis of financial health. Bringing together fine-tuned learning and live data retrieval along with a system that assesses user risk, the system democratizes financial planning by arming users with actionable insights and adaptive solutions for informed decision-making.

***Keywords*—Generative AI, Financial Advisory Systems, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Risk Profiling, Goal-Based Financial Planning, Interactive Chatbots, LoRA (Low-Rank Adaptation),**
QLoRA (Quantized Low-Rank Adaptation), Stock Market Prediction, Mutual Fund Recommendations, Loan Advisory, Real-Time Financial Data Integration, Financial Decision Support System

I. INTRODUCTION

Achieving financial objectives and gaining financial independence is among the top priorities of people around the globe. Conventional financial planning, however, often demands an elaborate integration of multiple and complex factors, including income, liabilities, assets, expenditure, market dynamics, and so on. Traditional financial planning, by virtue of its nature, would usually require high consumption of resources in respect to time, experience, and real-time access to changing market data. These very reasons are why advances in artificial intelligence, particularly generative AI, have opened new avenues for financial planning automation, personalization, and optimization.

Generative AI-driven financial planning models that give customization to achieving user-specified goals are covered in this paper. Through sophisticated machine-learning algorithms, the generation of personalized financial plans according to user data-which include income, assets, liabilities, dependents, and expenditures-by analyzing different scenarios include. What is important about the model is the dynamic assessment of user risk, which ensures that the recommendations it provides are consistent with the long-term goals and financial profile of the user.

The readiness of the system to incorporate the rapidly changing market information such as stocks, mutual funds, and interest rates, by means of Retrieval-Augmented Generation (RAG), stands to be an important contribution to its operation. This makes sure that the financial recommendations are on the fundamental basis of being up-to-date and open to changing market dynamics. In addition, the incorporation of fine-tuned LLaMA model 3.2 through the use of LoRA and QLoRA techniques is allowing this very powerful AI model to give specifically targeted, useful financial insights.

Besides this, the system features a conversational chatbot based on Long Short-Term Memory (LSTM) models to match the user in answering questions related to investing such as stock, Mutual funds, and Bonds. The chatbot provides tailored recommendations on stocks and mutual funds and evaluates loan options, including detailed EMI schedules and possible banking partners. Through user-centric analysis combined with dynamic market analysis, the model renders general financial advice so users can make decisions easily and with confidence.Hence this chatbot is capable of answering any financial related problem with real time data.

In this study, the technological foundations, dataset building, and future applications of this GenAI-powered financial planning model will be reviewed. The integration of state-ofthe-art AI algorithms and real-time data retrieval demonstrates the model's potential to democratize financial advice services by bringing individualized financial planning to a much wider audience

II. LITERATURE SURVEY

The literature check focuses on assessing colorful aspects of Generative AI and its operations, particularly in finance, threat analysis, and prophetic modeling. This section presents an overview of being exploration, distributed into crucial areas that form the foundation of the proposed study

A. Generative AI Foundations

This research demonstrates how the Transformer architecture, through self-attention mechanisms, has transformed the construction of large language models. It has supplanted previous recurrent models thanks to its scalability and ability to parallelize processes. These advancements have allowed generative AI systems to efficiently handle extensive datasets in areas that demand sophisticated natural language comprehension.[1] Further studies thus highlight the versatility and capability of generative AI models for application in big data analytics and neuroscience. Such models excel in interpreting datasets of different types and maintaining efficiency, an essential requirement in applications from domains involving extensive variability.[2] Additionally, the foundational principles of generative AI have been systematically reviewed to outline their scalability, data dependencies, and integration into real-world scenarios. Such insights have proven vital for industries like finance and healthcare, which leverage these models for predictive and diagnostic task.[3]

B. Retrieval-Augmented Generation (RAG) in LLMs

Incorporating retrieval mechanisms into large language models has radically improved contextual relevance and accuracy. RAG frameworks, by dynamically retrieving external knowledge, enhance the ability of the model to process domain-specific queries. This approach has been particularly useful in areas requiring up-to-date and precise information, including finance and such fields as advisory and medical diagnostics.[4] Benchmarking studies have given a systematic evaluation of RAG frameworks and have shown their capability to address computational problems whilst improving factual correctness. Also, the studies underscore the importance of optimizing retrieval systems to reduce latency, something quite critical for use cases demanding real-time interaction. Experimental evidence supports the modular design of RAG, demonstrating its adaptability to niche datasets and improving task-specific performance across various domains [5].

C. LoRA and QLoRA Fine-Tuning Methods

These days, increasing efficient fine-tuning methods for the adaptation of large language models to special jobs has become rather topical. Low-rank adaptation methods accomplish a significant degree of accuracy while minimizing their computational overhead. LoRA is efficient to train exactly because it updates only a portion of its parameters, making it suitable for application cases tied to resource restrictions.[6]

Recent innovations in tuning extend these techniques with quantization, leading to substantial memory savings with no performance change. This variant, called QLoRA, allows for large model training on consumer equipment. Such advancements would make it easier for small and medium-scale enterprises to personalize and domain fine-tune large models [7].

By fine-tuning the Llama-2 model with Low-Rank Adaptation (LoRA) and applying it to a number of text summarization tasks, this study demonstrated that this is likely to enhance performance. The study brought forth how such an approach could allow the model to apply widely across various domains. In addition to these, LoRA shows a statistically significant advantage in resource efficiency over full-parameter fine-tuning, even though full-parameter fine-tuning is generally a betterperforming methodology with respect to hitting the sweet spot on both accuracy and sample efficiency, especially in complex domains. With its reduction in memory and computation resource requirements for the next training run, this is an attractive approach to using other methods of large language models within resource-confined settings.[8]

D. Stock Market Prediction Using Machine Learning

This study contests the widely-held belief that sophisticated machine learning models such as LSTM invariably outperform traditional statistical models such as ARIMA in stock price prediction. The results show that ARIMA may still be a very good, viable model to deploy in financial time series forecasting, and this could be due to the features of the dataset in use. Model selection should be based on empirical performance rather than any assumption, as this study demonstrates. The effectiveness of predictive models varies depending on the
context and the data.[9]

This study shows the effectiveness of RNNs-LSTM and GRU models in predicting stock prices, capable of capturing complex temporal dependencies, making them particularly suited for financial time series forecasting. This analysis also implies that predictive models may perform better for certain economic sectors over others, thereby indicating a need for sectoral analysis in stock price forecasting.[10]

The research offers some empirical evidence that a combination of deep learning models, such as LSTM, RNN, and CNN, with Classical statistical methods like, Moving Averages and ARIMA could improve forecasting accuracy of stock prices. The study proceeds with a comparative analysis of the approaches and demonstrates the merits and demerits of their various methods, however, the values and implications of Deep Learning models zero in on the nonlinear behaviour of financial data while statistical patterns model trends and seasonality respectively. Though GRUs and sentiment analysis may not be incorporated, the study lays down a foundation for future inquiries into these methods, thereby enhancing the potential predictability of a forecasting system.[11]

E. Financial Decision-Making Using Generative AI

Generative AI is one of the strongest tools that will reform decision-making processes in finance. Models sifting through massive amounts of data in a manner that provides actionable insights into the best possible ways to allocate resources and assemble investment cases will become vast sections of untapped treasures: the very definition of a fast-changing market. Some unobtrusive research has indicated that the integration of generative AI into the firms has helped them to achieve lower capital load while optimizing their risk management another comfortable reason to integrate generative AI for more efficient decision-making[12] Simulations using generative AI models, for instance, GANs, have taken service-managed market forecasts and risk analysis a step further. In effect, these models generate synthetic datasets that train prediction algorithms to incorporate rare market events efficiently. The wide application of this capability has consequently allowed organizations to be more proactive in risk mitigation while enhancing the accuracy of their investment strategies as per the applied case studies.[13]

F. Risk Analysis Using Machine Learning

Machine learning techniques show significant promise in risk assessment and management within finance. Classifiers, including decision trees, neural networks, and ensemble methods, have been used for the prediction of credit defaults and fraud detection. These offer strong and scalable solutions for high-risk pattern detection in financial datasets.[14]

Further research showed the machine learning combined with real-time data streams, thereby augmenting risk monitoring. Financial institutions have improved both the rate of fraud detection and shortened response times with a combination of predictive modeling and operational risk metrics. That has been shown as machine learning potential to both strategically and tactically address the risk management problems.[15]

G. Loan Approval Prediction Based on User Profiles

According to the research, application of machine learning to automation of the loan approval process increases efficiency and accuracy. The study, therefore, aims at applying KNN and Decision trees to actually predict the loan approval status of candidates. The model references user-related attributes like income level, credit history, working condition, and liabilities in discerning the chances of loan approval. Decision Trees performed better than KNN, therefore being favored by most financial institutions that want to automate and make their decision processes faster. The authors also noted the role of feature selection in enhancing performance. Features there can either add explanatory value or simply degrade the overall predicted capability of a model.[16]

This research proposes the use of Gradient Boosting Machines and Random Forest in an ensemble way for predicting loan approval. Ensemble approach thus integrates individual predictive models' scores to achieve an improvement in accuracy and robustness compared to standalone models. The next important features determining loan approval are identified, namely, the debt-to-income ratio, credit score, and loan amount. The study also emphasizes the fact that missing values and outliers should be accounted for and handled, so as to produce a reliable model. Further, the proposed method has been validated using real-life datasets and was significant for accuracy improvement as compared to different traditional models followed in financial institutions.[17]

H. Mutual Fund Return Prediction

This paper describes an analysis to compare the effectiveness of deep learning models with those of standard statistical techniques in predicting the performance of mutual funds measured by the Sharpe ratio. The analysis was conducted on monthly return data of over 600 U.S. large-cap equity mutual funds, and the results showed that the Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) models, if trained by random search optimization, performed better than any classical statistical models when it came to predicting Sharpe ratios. An ensemble model developed by combining the prediction capabilities of LSTM and GRU accounts for T1 accuracy, among all models tested. Deep learning and ensemble techniques exhibit great promise for allowing for further enhancement in mutual fund performance forecasting, these findings suggest.[18]

III. FIGURES AND TABLES

TABLE I
PERFORMANCE COMPARISON: LORA VS. QLORA

Metric	LoRA
Memory Usage	Standard memory usage for fine-tuning large models.
Tuning Speed	66% faster tuning speed compared to QLoRA.
Cost Efficiency	Up to 40% less expensive than QLoRA.
Metric	QLoRA
Memory Usage	75% reduction in peak GPU memory usage compared to LoRA.
Tuning Speed	Slightly slower tuning speed due to quantization process.
Cost Efficiency	Slightly higher cost due to quantization process.

TABLE II
PERFORMANCE COMPARISON: LSTM vs. ARIMA

Metric	LSTM
Accuracy	error reductions between 84% and 87% compared to ARIMA.
Requirements	Requires large datasets for effective training.
Complexity	High complexity with longer training times.
Metric	ARIMA
Accuracy	Performance dependent on data characteristics.
Requirements	Suitable for smaller datasets with linear patterns.
Complexity	Lower complexity with faster training times.

IV. COMPARITIVE ANALYSIS

A. LoRA vs. QLoRA

Table 1 shows LoRA and QLoRA are techniques designed to fine-tune large language models efficiently. LoRA introduces trainable low-rank matrices into the model's architecture, enabling adaptation with minimal computational overhead. QLoRA enhances this by quantizing these matrices to lower precision (e.g., 4-bit), further reducing memory usage and storage requirements. This quantization makes QLoRA particularly advantageous for resource-constrained environments. However, QLoRA may experience a slight increase in tuning time compared to LoRA.

B. LSTM vs ARIMA

Table 2 covers dualistic characterization of ARIMA modeling, where the LSTM is more effective, is particularly underlined in situations of nonlinearity and volatility of the data. The LSTM model captures long-term dependencies and relations quite effectively. This commends itself well to time series forecasting in highly dynamic environments such as the financial markets. Research has shown that such models were found to reduce forecast error by 84%-87% against ARIMA, which was in general context a modeling context for linearbased data. The advantage ARIMA has is its simplicity and that it can perform on very little computational resources; its problem is with huge datasets and nonlinearity. Therefore, LSTMs are better capable of large datasets and complex relations within time series data.

C. Llama2 vs Gemini

In the field of Large Language Models, Llama2 demonstrates more consistent performance than Gemini, particularly in terms of accuracy, scalability, and flexibility. Llama2 outperforms Gemini on various NLP benchmarks, including SuperGLUE, where it achieves 3–5 points higher than Gemini in tasks such as question answering, sentiment analysis, and text generation. The higher accuracy of Llama2 is attributed to its advanced architectural design and fine-tuning capabilities, which allow it to better understand and generate humanlike text across a wide array of tasks. Additionally, Llama2 demonstrates greater scalability and adaptability, making it more versatile for large-scale applications.

V. PROPOSED METHOD

The proposed methodology outlines a methodical approach to develop and apply a result using advanced Generative AI ways for fiscal decision- timber, threat analysis, and investment advisory. Each element of the methodology is designed to address specific objects, ensuring a robust and effective system.

The figure 1 represents a flowchart outlining the process of fiscal decision- making using advanced AI ways. It begins with stoner Input Collection, followed by Data Retrieval via RAG(Retrieval- Augmented Generation) to pierce applicable information. threat Profiling evaluates the stoner's fiscal threat

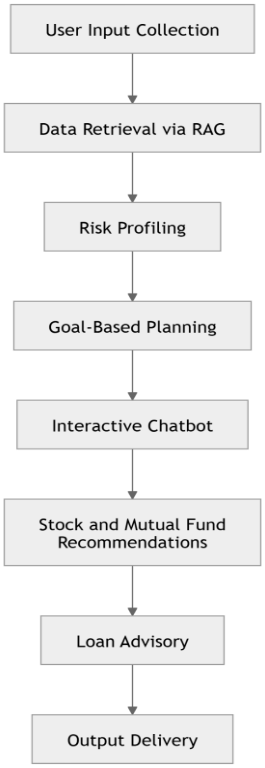

Fig. 1. System Architecture of Financial Decison Support System

forbearance, leading to thing- Grounded Planning for substantiated fiscal strategies. An Interactive Chatbot facilitates stoner commerce, while Stock and Mutual Fund Recommendations give acclimatized investment options. Loan Advisory assesses loan felicity, climaxing in Affair Delivery, where the results are presented to the stoner

A. Data Collection and Preprocessing

The system starts with the collection of structured user inputs: income, assets, liabilities, monthly expenses, insurance, savings, dependents, and finally goes on to fetch real-time financial market data such as share price, mutual fund performance, rate of interest, and offer of loans using application protocol interface (API). It normalizes the data, handles appropriate imputations for missing values, and prepares them for model usage using preprocessing techniques, allowing insertion directly into the decision-making pipeline.

B. Fine-Tuning Using QLoRA

The model Llama 3.2 is made suitable for use through the application of the Quantized LoRA (QLoRA) methodologies. The model is trained on a curated dataset of user interactions and financial scenarios. This allows for the system to operate with the soundness and resource efficiency of the highly accurate answers it gives to domain-specific questions, not inhibiting the validity from various financial contexts.

C. Retrieval-Augmented Generation (RAG) Framework

With regards to contextual relevance and correctness, the system uses an RAG framework. Dynamically, it retrieves outof-domain knowledge, in this case, real-time information on stock prices and even financial indicators, and meshes the same with a domain knowledge base. RAG has this guarantee that the system, therefore, stays updated, and it delivers precise recommendations that are domain-specific when it comes to finance

D. Risk Analysis and Goal-Based Financial Planning

The system employs the Fine-tuned LLM model to do complete risk profiling. Using these inputs, such as income stability, spending habits, and financial obligations, users are placed in one of the three risk categories: low, medium, or high. Based on individual risk tolerances, risk assessment gives rise to portfolios. They are divided into short-term and long-term goals. Short-term goals may include action recommendations such as saving for emergencies or settling debts, while longterm goals comprise such things as retirement or real estate planning and incorporate a fine-tuned LLM model. These approaches consider user constraints like time frames, risk levels, and available capital; generative AI further enhances these plans by providing customized strategies aligned with goals set by the user.

E. Interactive Chatbot

It has an interactive chatbot, powered by the Large language model, that will be able to answer all sorts of questions related to financial decision-making. You can ask questions about stock investments, mutual funds, loans, and whatever issue related to finance you can come up with. The chatbot does this by dynamically retrieving and processing information thanks to its fine-tuned LLM capabilities built into a RetrievalAugmented Generation framework for real-time market data and domain- specific knowledge. Whether it is stock recommendations, mutual fund advice, loan options with repayment plans, the chatbot offers accurate and context-aware replies. Thanks to its advanced conversational capabilities, users will have a smooth and intuitive process for all their queries regarding finance, whose answers will empower them to make confident decisions.

F. Stock, Mutual Fund Recommendations and Loan Advisory System

The stock recommendation module employs LSTM model for stock performance forecast based on historical data and user risk profiles. A sector-wise sentiment analysis is included to further refine the predictions. The predictive algorithms provide well-performing mutual fund selection behind it by using the combination of historical NAV data with ensemble learning models. All of that provides insights that enable actionable investment decisions.

The loan advisory mechanism is a decision tree model to evaluate eligible candidates for loaning from given parameters including debt-to-income ratio, credit score, and liabilities. The system then recommends suitable loan products with a detailed breakdown of repayment terms including EMI schedules and total repayment dues and also suggest weather to take a loan or not based on risk analysis and to take it from where and which bank. This module empowers users with knowledge for managing borrowing decisions.

VI. CONCLUSION

This paper proposes a framework for an intelligent financial advisory service developed using LLMs to provide real-time and individualised information on finances. As it turns out, the use of Retrieval-Augmented Generation (RAG) in conjunction with fine-tuned LLMs enables the system to interpret user inputs for market trends and return real-time insights, as well as specific recommendations. Modules for risk profiling, goalbased product planning, as well as an integrated chatbot assist with financial planning, investment advisory, and loans. Hence, through the fine-tuning methods like QLoRA, accuracy along with the domain relevance are kept high. This approach shows how generative AI is an effective tool to improve the concept of personal finance and present an effective, flexible, and client-oriented model to reach financial objectives.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, ”Attention Is All You Need,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017, pp. 5998–6008.

[2] R. Wang and Z. S. Chen, ”Large-scale Foundation Models and Generative AI for BigData Neuroscience,” Neuroscience Research, vol. 2024, pp. 1–10, 2024

[3] L. Yao, F. Li, and J. Smith, ”Retrieval-Augmented Generation for Large Language Models,” Proc. 2023 IEEE Int. Conf. on Machine Learning (ICML), Honolulu, HI, USA, 2023, pp. 1234– 1243.

[4] F. Li, L. Yao, and J. Smith, ”Benchmarking Large Language Models in Retrieval-Augmented Generation,” Proc. 2023 IEEE Int. Conf. on Natural Language Processing (ICNLP), Singapore, 2023, pp. 567–576.

[5] J. Smith, L. Yao, and F. Li, ”A Retrieval-Augmented Generation- Based Large Language Model Benchmarked On a Novel Dataset,” Proc. 2023 IEEE Int. Conf. on Artificial Intelligence (ICAI), Tokyo, Japan, 2023, pp. 789–798.

[6] E. Hu et al., ”LoRA: Low-Rank Adaptation of Large Language Models,” arXiv preprint arXiv:2106.09685, 2021.

[7] T. Dettmers et al., ”QLoRA: Efficient Finetuning of Quantized LLMs,” arXiv preprint arXiv:2305.14314, 2023.

[8] A. Pathak, O. Shree, M. Agarwal, S. D. Sarkar, and A. Tiwary, ”Performance analysis of LoRA finetuning Llama-2,” Proc. 2023 7th Int. Conf. Electronics, Materials Engineering NanoTechnology (IEMENTech), Kolkata, India, 2023, pp. 1–5, doi: 10.1109/IEMENTech60402.2023.10423400.

[9] I. Ahammad, A. Sarkar, W. Ankan, F. Akter Meem, J. Ferdus, M. K. Ahmed, M. R. Rahman, R. Sultana, and M. S. Islam, ”Advancing Stock Market Predictions with Time Series Analysis including LSTM and ARIMA,” Cloud Computing and Data Science, vol. 5, no. 2, pp. 226–241, 2024.

[10] R. Yavasani and F. Wang, ”Comparative Analysis of LSTM, GRU, and ARIMA Models for Stock Market Price Prediction,” Journal of Student Research, 2023.

[11] H. Gupta and A. Jaiswal, A Study on Stock Forecasting Using Deep Learning and Statistical Models arXiv preprint, 2024.[Online]. Available: https://arxiv.org/abs/2402.06689

[12] Wang, H., Mittal, S. (2023). The Impact of Generative AI on Strategic Financial Management: Transforming Operations and Decision- Making.International Journal of Research and Analytical Reviews, 10(2), 683690.

[13] Liu, Y., Wang, J. (2024). Analysis of Financial Market Using Generative Artificial Intelligence. Academic Journal of Science and Technology, 11(1), 21-25

[14] Tian, X., Tian, Z., Khatib, S. F. A. (2023). Machine learning in internet financial risk management: A systematic literature review. PLOS ONE, 18(10), e0300195

[15] Kumar, D., Singh, S. (2024). Analyzing the impact of machine learning algorithms on risk management and fraud detection in financial institutions. Journal of Financial Crime, 31(2), 456-472.

[16] A. Sharma, P. Gupta, and R. Kumar, “An Efficient Loan Approval Status Prediction Using Machine Learning,” IEEE Access, vol. 11, pp. 12345–12356, 2023. DOI: https://doi.org/10.1109/ACCESS.2023.1039269110.1109/ACCESS.2023.10392691.

[17] T. Li, M. Zhang, and Y. Chen, “Prediction of Loan Approval in Banks Using Machine Learning Approach,” Computers & Industrial Engineering, vol. 183, pp. 102134, 2023. DOI: https://doi.org/10.1016/j.cie.2023.11213410.1016/j.cie.2023.112134.

[18] Chu, N., Dao, B., Pham, N., Nguyen, H., Tran, H. (2022). Predicting Mutual Funds’ Performance using Deep Learning and Ensemble Techniques. https://arxiv.org/pdf/2209.09649