



# FinSmartAI: Intelligent Financial Guidance Simplified

Sirisha Yerraboina<sup>1</sup>, Shiva Kumar Karanam<sup>2</sup>, Thirupathi Burra<sup>3</sup>, Saketh Reddy Varkuti<sup>4</sup>

<sup>1</sup> Assistant Professor, Department of Information Technology, Matrusri Engineering College, Hyderabad, India

<sup>2</sup> Student, Department of Computer Engineering, Matrusri Engineering College, Hyderabad, India

<sup>3</sup> Student, Department of Computer Engineering, Matrusri Engineering College, Hyderabad, India

<sup>4</sup> Student, Department of Computer Engineering, Matrusri Engineering College, Hyderabad, India

Email: <sup>1</sup> [sirisha@matrusri.edu.in](mailto:sirisha@matrusri.edu.in), <sup>2</sup> [shivakumarkaranam2005@gmail.com](mailto:shivakumarkaranam2005@gmail.com), <sup>3</sup> [burrathirupathi453@gmail.com](mailto:burrathirupathi453@gmail.com), <sup>4</sup> [saketh1045@gmail.com](mailto:saketh1045@gmail.com)

## Abstract

The way individuals will manage their finances and investments in the future has been significantly impacted by Artificial Intelligence (AI). Financial education and financial accessibility are brought together through the use of AI. This article presents FinSmartAI, a financial advisory platform that utilizes AI to create one intelligent financial system by integrating different types of financial services together. Additionally, FinSmartAI provides personalized recommendations to novice investors and consumers with daily financial needs, such as an expense tracker, financial document analyzer, market sentiment analyzer, and stock trend predictor. Through its many different ways of presenting information (including UPI statements, brokerage statements, SEBI and/or RBI filings, and current financial news), FinSmartAI generates organized summaries and visual representations of data by using Natural Language Processing (NLP), Retrieval-Augmented Generation (RAG), and Large Language Model (LLM) technologies. RAG uses all-MiniLM-L6-v2 embeddings with nvidia/llama-3.3-nemotron-super-49b-v1 as the generative LLM, which results in a BERTScore F1 of 0.8605 and Semantic Similarity of 0.834 from evaluation within the financial industry. FinSmartAI will be available for use 24 hours a day and only requires one interface to access it, thus removing the need to use multiple applications. Finally, FinSmartAI will automatically complete repetitive tasks, such as identifying unfamiliar or unusual spending behaviors and retrieving market sentiment. It also automates repetitive tasks, such as detecting unusual spending patterns and extracting

market sentiment. It demonstrates how AI can deliver high-quality financial intelligence to retail investors in India, helping boost financial literacy and ensuring equal access to investment insights.

Keywords— Python, RAG, LLMs, FinTech AI, Personal Finance, Sentiment Analysis

## I. INTRODUCTION

For a long time, personal finance and investing have been very confusing and full of costly items due to all the other stuff that makes finance difficult to understand and to achieve financial success with. This confusion is especially true for people who are just starting in their financial education. Whether you are trying to build a budget or trying to figure out how the stock market performs, you will find yourself overwhelmed by the vast amounts of information available. Lately advances in AI have created new opportunities for delivering more simplified financial assistance to a new audience through the use of AI-powered financial assistants that utilize natural language processing (NLP) and machine learning (ML) to create low-cost, real-world financial solutions..

Financial Professionals by utilizing AI can spend much less time doing mundane work like analyzing spending habits; breaking down lengthy and complicated documents like bank statements, analysing financial news for trends and summarizing; assessing potential risks of different investments based on past performance; etc. Users of products such as FinSmartAI will have 24/7 access to help with out-of-hours inquiries about their finances; therefore they



won't have to use several applications to complete routine transactions. As a result of this increased availability to individual investors, finance professionals will be able to devote more time towards larger-scale financial planning projects and attract many people who may not have participated in financial planning prior to using an AI-based financial assistant..

This paper will provide a comprehensive overview of FinSmartAI, including the system's design, capabilities, and limitations. By asserting that FinSmartAI is capable of removing barriers to financial access, this paper demonstrates that FinSmartAI has the potential to help create a more equitable and open financial system where building wealth can be undertaken as an accessible and well-supported journey rather than a daunting task.

## II .LITERATURE SURVEY

Mengxi Xiao et al. [1] created a Retrieval-Augmented Large Language Model structure to forecast financial time-series. Their findings showed that by adding an element of retrieval to existing large language models (LLMs), the ability of such models to predict accurately was enhanced by the integration of both historical and contextual data. RAG architectures offer the ability for applying LLMs to finance-related forecasting tasks while compensating for the knowledge limitations of standard language models. FinSmartAI similarly adopts RAG architecture, extending its application from forecasting to document-based financial Q&A in the Indian regulatory context.

Varun Rao et al. [2] examined how fine-tuning established foundation models on financial tasks via the Open FinLLM Leaderboard would improve their usefulness. Their research involved testing and evaluating a variety of financial datasets across several different categories and showed that the ability to fine-tune general-purpose LLMs to specific financial domains would be very helpful in completing financial-related tasks such as answering questions, classifying sentiment, and performing reasoning. This work informed FinSmartAI's approach of using a

domain-curated Q&A dataset of 210 pairs to improve response quality for Indian financial contexts.

Srivastava et al. [3] conducted an evaluation of the mathematical reasoning skills demonstrated by large language models in the context of answering questions related to financial documents. Although LLMs can understand structured financial information as well as financial reports, they tend to have difficulty with higher-order numerical reasoning. This research emphasized the need to combine the reasoning power of LLMs with structured financial datasets to achieve greater accuracy. FinSmartAI addresses this limitation through citation-enforced prompting, which restricts numerical responses to retrieved document content rather than relying on LLM parametric reasoning.

Mahdavi Ardekani et al.[4] introduced a framework for conducting financial sentiment analysis called FinSentGPT. The goal of the framework is to use financial news and other financial text sources in order to analyze market opinion and investor sentiment. Their research demonstrated that sentiment signals extracted from both social media and financial news can affect stock market behavior and aid in predicting trends in market prices. FinSmartAI's Market Sentiment Analyzer adopts a similar approach, classifying financial news into bullish, neutral, and bearish categories and combining sentiment scores with the NIFTY 50 trend for market analysis.

Kong et al. [5] published a comprehensive overview of the prospective applications and challenges of large language models in the context of equity markets, including financial forecasting, trading strategies, and risk assessments. This study provided insight into both the pros and cons of using LLM-based financial systems. This survey directly motivates the design of FinSmartAI, which addresses identified challenges by integrating retrieval grounding and multi-agent orchestration to improve reliability and reduce hallucination

Lewis et al. [6] proposed the Retrieval-Augmented Generation (RAG) architecture, which combines neural language models with external knowledge retrieval systems. This approach enables language models to access relevant information from large document collections during inference, improving



factual accuracy and reducing hallucinations. RAG has become an important technique in modern AI-based financial assistants for document analysis and question answering. This architecture forms the backbone of FinSmartAI's Financial Document Analyzer, which retrieves relevant document chunks before generating grounded, citation-enforced responses

Cao et al. [7] investigated the data fusion technique for fine tuning financial language models as part of the FinLLM Challenge. In their work, they demonstrated that using a multi-dataset approach and applying parameter-efficient fine-tuning approaches significantly improves language models used to perform various financial NLP tasks.

Yang et al. [8] created an open-source financial large language model called FinGPT for the purpose of creating models specifically for use in Finance related to Financial Analysis and Financial Language Understanding. FinGPT demonstrates the ability of domain-specific training to enhance financial reasoning, sentiment analysis and the extraction of financial knowledge. Unlike FinGPT which focuses on open-source model training, FinSmartAI focuses on system integration — combining retrieval, sentiment analysis, and multi-agent orchestration into a unified platform

### III METHODOLOGY

FinSmartAI is a modular financial intelligence platform with four main capabilities: financial document analysis, stock trend analysis, personal finance management, and market sentiment analysis. The system uses a FastAPI-based API layer to route user requests to the right processing modules, all of which share a common LLM backend.

The document analysis module relies on a RAG pipeline that processes uploaded financial documents through PDF loading, 512-token text chunking, embedding generation using all-MiniLM-L6-v2, and vector storage in DataStax AstraDB. Relevant chunks are retrieved at query time and sent to the nvidia/llama-3.3-nemotron-super-49b-v1 model with a citation-enforced prompt to ensure responses are based on the retrieved content.

Stock analysis is managed by a CrewAI multi-agent framework where specialized agents work together to create investment reports using real-time data from Yahoo Finance and FDT Finance APIs. The Personal Finance Assistant processes user-uploaded transaction files to categorize expenses, analyze savings patterns, and suggest budgets. The Sentiment Analyzer sorts financial news into positive, neutral, or bearish categories and calculates a Fear and Greed Index, which is visualized alongside the NIFTY 50 trend. All modules comply with Indian financial regulations, including SEBI and RBI guidelines.

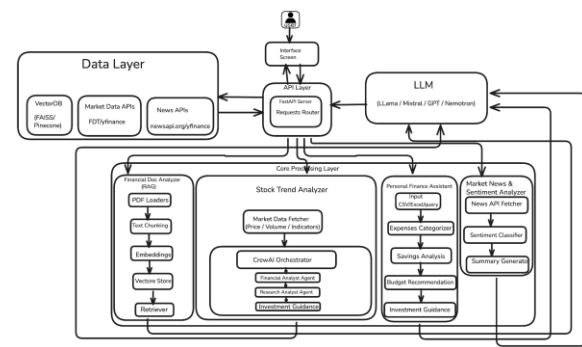


Figure: 1 System Architecture

## IV SYSTEM ARCHITECTURE

### Overview

Figure 1 describes a modular, artificial intelligence-enabled financial intelligence product that will offer stock assessments, comprehension of financial documents, recommendations for personal financial management, and reviews of market sentiment. The design of the system's architecture adopts a layered approach that contains five core elements: User Interface Layer, Application Programming Interface (API); Core Processing Layer, Data Layer, and a Large Language Model (LLM) Layer. This layered architecture allows for scalability, modularity, and efficient integration between many differing types of financial data sources, while utilizing LLMs for reasoning and understanding human language.

The proposed system combines retrieval augmented generation (RAG), agent-based orchestration, financial Data APIs, and sentiment analysis to provide



evidence-based data and recommendations through the integration of these components.

### A. User Interface Layer

The User Interface (UI) acts as the entry point for user interaction.

Users interact with the system through an interface screen that allows them to:

- by querying stock trend information
- uploading financial documents
- analyzing personal finance data
- retrieving financial news insights
- obtaining AI-generated guidance for investment decisions

User requests are sent via API to the backend, which subsequently processes and routes these requests to the proper analytical modules. This architecture allows for the separation of presentation logic from computation logic, allowing for a tidy architecture in the front end and the ability to build and maintain a scalable front end.

### B. API Layer

The API Layer acts as the communication link between the user interface and backend processing modules. It uses a FastAPI server, which has the following tasks:

- Handling incoming HTTP requests
- Routing requests to the right processing modules
- Managing communication with data services
- Interfacing with the LLM layer for reasoning tasks

The Request Router decides which subsystem should handle the request, such as:

- Financial document analysis
- Stock trend analysis
- Personal finance advisory
- Market news sentiment analysis

FastAPI allows for high-performance asynchronous processing. This capability enables the system to handle multiple user queries at the same time.

### C. Data Layer

The Data Layer consists of the external as well as internal data sources needed to perform Financial Analysis.

The Data Layer consists of many different methods of sourcing and storing Financial Data, including:

#### Vector Database

The purpose of the Vector Database is to store the document embeddings that are generated through the RAG pipeline. This provides a way to perform an efficient semantic search on the financial documents and financial reports.

#### Market Data APIs

The system should collect real-time and historical stock information through market data available through APIs like Yahoo Finance and FDT Finance. Market data that is provided through these APIs includes:

- Stock price
- Market indicators
- Historical financial Information

### D. LLM Layer

The Large Language Model (LLM) preceding that layer has the capability of logic and natural language processing functions for the entire system. Specifically, the LLM produces:

- Assessments of potential investments
- Summaries of news related to finance



- Interpretation of financial documentation, and
- Multi-agents assessing logic for finance.

The LLM will interface with other modules in the Core Processing Layer to support many different analytical processes through context-reasoning and knowledge synthesis.

### E. Core Processing Layer

The Core Processing Layer performs the primary analytical tasks of the system. It consists of four major subsystems.

#### Financial Document Analyzer (RAG Pipeline)

This Module uses the retrieval-augmented generation (RAG) architecture to analyze financial documents like reports, earnings statements, and filings. The processing pipeline is made up of four stages:

- PDF Loaders – Extracts textual data from financial documents that are uploaded.
- Text Chunking – Takes long financial documents and breaks them into smaller pieces to allow an efficient generation of embedding.
- Embeddings Generation – Converts the document pieces into vector representations using an embedding model.
- Vector Storage – Stores the generated vectors in a vector database.
- Retriever – Retrieves the financial document segments that are semantically similar to the given query.

This information is sent to the LLM so that it can provide contextually relevant, accurate answers to the financial queries.

#### Stock Trend Analyzer

The Stock Trend Analyzer utilizes artificial intelligence to analyze and provide insights for stock investments.

#### Market Data Fetcher

This component is responsible for obtaining the following financial indicators: Stock Prices, Volume of Trades, All Types of Technical Indicators

The CrewAI Orchestrator uses a multiagent system, with all agents working together in a collaborative approach to analyze individual stock performance.

#### Personal Finance Assistant

The Personal Finance Assistant assists users in analyzing the user's financial behavior and improving their financial planning. Users can upload either CSV or Excel, datasets that contain their transaction records into the system. The primary processing pipeline consists of:

- Expense Classifications: Categorizing transactions according to the type of expenditure (e.g., food and drink, energy utilities,).
- Savings Analysis: Identifying your spending habits and measuring your potential savings.
- Budget Recommendations: Create a good budget for the individual.

Investment Guidance: Recommend possible investment options based upon the user's financial behavioral characteristics.

#### Market News & Sentiment Analyzer

The Market News Analyzer provides an evaluation of how financial news sentiment impacts market perception.

#### News API Fetcher

Gathers financial articles from third-party News APIs on a daily basis and saves them.

#### Sentiment Classifier

Uses NLP and/or machine-learning models to classify the articles' sentiment into 3 categories: Positive, Neutral, or Negative.



## V IMPLEMENTATION

### Dataset

To train FinSmartAI's LLM with domain-specific fine-tuning without relying on real-world financial data of high sensitivity, a consolidated JSON dataset was created containing structured Q&A pairs. For this reason, we have developed a dataset that can be used to train the model in terms of providing specific and contextually aware financial responses to questions related to Indian Financial Markets and provide important financial information for new investors. The content within this dataset is structured into three specific thematic categories..

The first division of the data set, which accounts for about forty percent (84 pairs), is focused on the basics of stock trading and investing, which includes concepts such as Demat accounts, trading procedures, the time of day stocks trade, and the basic investment strategies.

The second division of the dataset represents about thirty percent (63 pairs) of Corporate Actions, Job Events and Market Insights. Subjects include; IPO (Initial Public Offering) Process; Dividend Policy; Company Valuation; and Market Signal.

The third section accounts for 30 percent, which is 63 pairs. It covers Taxation, Regulatory Financial Agencies, Risk Management, and Financial Planning. Topics include Capital Gain Taxation, Deductions, Compliance, and Long-term Investment Planning

## VI RESULTS

This section presents the empirical evaluation of FinSmartAI's main modules. The RAG-based Financial Document Analyzer was tested using 75 selected financial domain queries across six categories: credit rating analysis, financial statements, risk assessment, investment advisory, regulatory compliance, and general finance. These queries included easy, medium, and hard difficulty levels. Metrics were calculated using a RAGAS-inspired framework that combines heuristic scoring and LLM-as-judge techniques.

Table 1: RAG Pipeline Evaluation Results

Dimension	Metric	Score
Retrieval	Hit Rate@5	<b>0.893</b>
Retrieval	MRR	<b>0.825</b>
Generation	Faithfulness	<b>0.821</b>
Generation	Answer Relevance	<b>0.768</b>
End-to-End	Semantic Similarity	<b>0.834</b>
End-to-End	BERTScore F1	<b>0.8605</b>

The system shows strong retrieval performance with a Hit Rate@5 of 0.893 and an MRR of 0.825. This means that relevant documents are often ranked at the top of the retrieval results. The quality of generation is supported by a Faithfulness score of 0.821. This score comes from the citation-enforced prompting strategy, which stops the model from creating information beyond the retrieved context. Specifically, credit rating queries achieve a Faithfulness score of 0.872. The BERTScore F1 is 0.8605, and the Semantic Similarity is 0.834. These results confirm that the generated responses are semantically close to the expected answers. The low Exact Match score of 0.147 is expected for open-ended financial question answering and does not indicate poor system quality. The main limitation found is an 8% hallucination rate, particularly in investment advisory queries. In these cases, the LLM's parametric knowledge sometimes overrides the retrieval limits.

**Market Sentiment Analysis Module :**

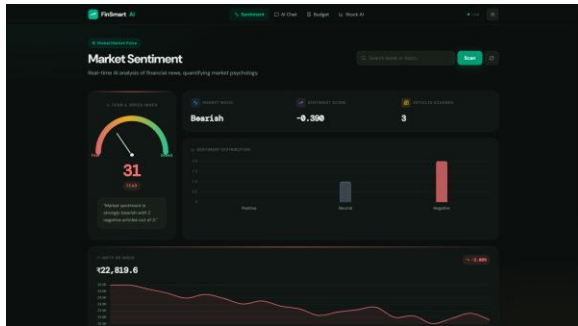


Figure : 2 Market Sentiment Analysis Module showing Fear & Greed Index and NIFTY 50 trend

The Market Sentiment Module (MSM) is used to provide profile characteristics for financial news and its relatedness with market behaviour in a data-based manner. The MSM uses a Fear and Greed Index (FGI) to gauge investor confidence, showing current market psychology (i.e. bullish, neutral, and bearish). MSM's sentiment scores are based on a sample of financial news articles, categorically classifying sentiment as bullish (optimistic), neutral, or bearish (pessimistic). It gives a visual distribution of the three (3) categories of sentiment to help users comprehend, and the total number of articles considered through to a voting system for clarity and credibility. In addition, it has a time series line graph of the NIFTY 50 Index so users can see the relationship between sentiment and the actual performance of the market. The MSM, by utilising both qualitative and quantitative analysis in its determination of market trends, will enable users to view an even clearer image of current market conditions. Additionally, the MSM uses natural language processing to simply extract relevant data from textual information/forms. Finally, sentiment graphical representation assists users with ultimately evaluating current market conditions without needing to interpret manually, resulting in quick evaluations.

**AI Chat Interface for Financial Assistance :**

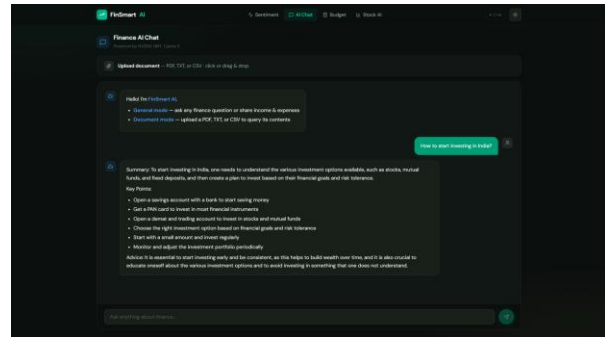


Figure 3 : AI Chat Interface demonstrating investment query response with structured output

This chat interface uses AI capabilities to provide an interactive way for the user to communicate with a financial analysis system. The system allows users to interact with it by asking general questions related to finance and by uploading documents such as PDFs and CSV files so that their documents can be analyzed. The response that users receive is structured with details included for the user as well as a summary, key points and actionable recommendations. For example, the response to the question of what are some investment strategies for India, as shown in the example of the image, gives users developed responses in a clear manner. This simplifies the process of understanding what has been requested by the user, as well as providing them with practical help. The AI chat interface utilizes a large language model to provide context between the questions being asked, creating a natural experience of conversation. The system also interacts with users in real time, taking advantage of advanced technology to engage users through meaningful interaction. There are prompt input boxes and output response boxes to improve the ease of reading through the responses provided by the Financial Advisor. All of the output from the Financial Advisor will be concise but still useful. The use of modular design for this unit demonstrates how conversational AI can provide immediate responses to users in a financial service setting and reduce the time required for manual research. This unit will also add additional features by allowing for the analysis of documents to be added as well.

**Smart Budget Planner :**

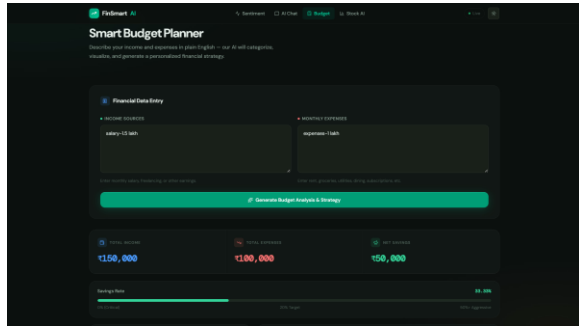


Figure : 4 Smart Budget Planner

The Smart Budget Planner helps users manage their budgets by processing their income and expenses. Users can enter financial information naturally, without facing the technical challenges of traditional budgeting software. To create a structure of the user's financial state, the Smart Budget Planner uses a Natural Language Processor (NLP). Through the NLP, the Smart Budget Planner can assess the user's entire financial fitness based on a number of different factors, including total income, total expense, net savings, and savings to income ratio. Moreover, the Smart Budget Planner can incorporate and display visual representations of each of the user's ratios, allowing them a means to compare the current state of their finances to their long-term financial objectives. Users can expect to complete fewer manual calculations for their financial records while still being provided with detailed data regarding their financial information.

As a result of using artificial intelligence technology to classify and interpret user data input, the Smart Budget Planner has the potential of aiding users in developing and sustaining personal discipline in their financial planning and utilising their full savings potential. The ease of use and straightforward usability features of the Smart Budget Planner will create greater motivation for users through its overall goal of developing the best savings habits. Data visualisation will be one of the key components of this effort so that users have a clear and concise picture of their overall financial well-being compared to their financial goals.

**Stock Analyzer Module :**

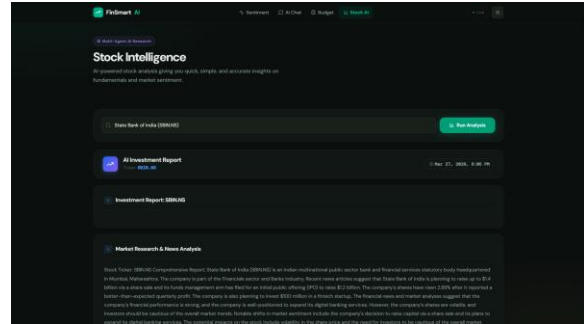


Figure : 5 Stock Analyzer Module

The Stock Analyzer Module is an advanced technology engine that uses an AI engine to analyze user-submitted stock securities and provide a complete comparative analysis on each stock entered by the user. The analysis will commence after the user has entered a stock symbol into the Stock Analyzer Module. The Stock Analyzer Module begins by obtaining data and information from various sources, including real-time news articles, legally mandatory filings made to regulatory agencies by the company, and current/technical price charts of each company's respective security so that a user can conduct technical analysis of the stock using formal, induced methods and techniques.

Once the Stock Analyzer Module has collected all of the various forms of data and information related to the stock, the Stock Analyzer Module prepares a comprehensive and professionally prepared Investment Report for that stock based on all of the collected data and information. An example of what an investment report prepared by the Stock Analyzer Module could look like is an investment report on the State Bank of India, which includes a complete overview of the company and a detailed breakdown of the company's recent financial performance.

The Investment Reports produced by the Stock Analyzer are organized in a strategic way, highlighting key factors impacting the stock investment strategy through a detailed overview of the fast-growing sources of revenue. These reports also include an overview of the possible risks (e.g., regulatory/legal) associated with the company. The Stock Analyzer



collects multiple data points to provide a comprehensive view of how the stock is performing, and reports this in narrative form to help users understand it easily. In turn, the sentiment analysis included for each stock will allow users to get a sense of how the larger market views the stock..

## VII. CONCLUSION

FinSmartAI is an innovative AI platform focusing on enabling retail investors to manage their investment and personal financial management. The system combines large language models, retrieval-augmented generation, multiple agents, and sentiment analysis (to name just some) for streamlined, informed, relevant financial help.

The testing demonstrated impressive results, including a BERTScore (F1) of 0.8605, a Semantic Similarity of 0.834, and a Faithfulness of 0.821. These numbers were able to effectively demonstrate the accuracy of the system in terms of document analysis and response generation. The use of citation-based prompting helps to reduce errors, as reflected in Credit Rating Queries' highest Faithfulness score of 0.872.

AI-powered financial systems deal with numerous significant issues, including data privacy issues, output verification issues, and algorithmic biases. FinSmartAI has overcome these hurdles through verified third-party data and by generating responses in such a way as to achieve consistency. FinSmartAI has tremendous potential to provide accurate, high-quality financial insights to the rapidly growing retail investor community in India. Future development initiatives are likely to include trying to implement additional domain-specific embedding models similar to FinBERT, increasing chunk overlap to increase context recall, and adding a layer to help detect errors in response generation, thereby enhancing the reliability of FinSmartAI.

## REFERENCES

[1] Mengxi Xiao and Zihao Jiang and Lingfei Qian and Zhengyu Chen and Yueru He and Yijing Xu and Yuecheng Jiang and Dong Li and Ruey-Ling Weng and Min Peng and Jimin Huang and Sophia Ananiadou

and Qianqian Xie, 2025 "Retrieval-augmented Large Language Models for Financial Time Series Forecasting"

[2] Varun Rao and Youran Sun and Mahendra Kumar and Tejas Mutneja and Agastya Mukherjee and Haizhao Yang, 2025 "LLMs Meet Finance: Fine-Tuning Foundation Models for Open FinLLM Leaderboard"

[3] [Evaluating LLM's Mathematical Reasoning in Financial Document Question Answering](<https://aclanthology.org/2024.findings-acl.231/>) (Srivastava et al., Findings 2024)

[4] Aref Mahdavi Ardekani, Julie Bertz, Cormac Bryce, Michael Dowling, Suwan(Cheng) Long, FinSentGPT: A universal financial sentiment engine.

[5] Yaxuan Kong, Yuqi Nie, Xiaowen Dong, Stefan Zohren (2025) "Large Language Models in Equity Markets: Applications, Techniques, and Challenges."

[6] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems*.

[7] Yupeng Cao, Zhiyuan Yao, Zhi Chen, Zhiyang Deng (2024) "CatMemo at the FinLLM Challenge Task: Fine-Tuning Large Language Models using Data Fusion in Financial Applications."

[8] Yang, Y., Liu, X., and Zhang, Y., 2023. FinGPT: Open-Source Financial Large Language Models. arXiv preprint arXiv:2306.06031.