



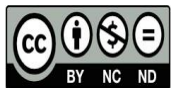
Auto Trader X: An Agentic AI and Retrieval-Augmented Generation Framework for Real-Time Explainable Financial Decision Support

Swayam¹, Mayank Gahlawat², Anshuman Lochav³, Dr. Vivek Mehta⁴

^{1,2,3} Department of Computer Science and Engineering, Netaji Subhas University of Technology (NSUT), Delhi, India.

⁴ Supervisor, Department of Computer Science and Engineering, NSUT, Delhi, India.

To Cite this Article: Swayam¹, Mayank Gahlawat², Anshuman Lochav³, Dr. Vivek Mehta⁴, "Auto Trader X: An Agentic AI and Retrieval-Augmented Generation Framework for Real-Time Explainable Financial Decision Support", Indian Journal of Computer Science and Technology, Volume 05, Issue 02 (May-August 2026), PP: 101-103.



Copyright: ©2026 This is an open access journal, and articles are distributed under the terms of the [Creative Commons Attribution License](#); Which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract: Financial markets are increasingly driven by a combination of quantitative trends and qualitative market sentiment originating from news, macroeconomic events, and social media discussions. Traditional financial forecasting systems that rely solely on historical time-series information often fail to capture sudden volatility caused by real-world events. This research presents AutoTraderX, a hybrid cloud-native financial intelligence framework integrating Agentic AI workflows, Retrieval-Augmented Generation (RAG), and Fine-Tuned Large Language Models for real-time explainable trading support.

The proposed framework combines real-time market data ingestion, vector database retrieval, and contextual reasoning using a Quantized Llama-3 Large Language Model fine-tuned using QLoRA techniques. The system generates explainable BUY/SELL/HOLD recommendations while maintaining low inference latency and high semantic precision. Experimental evaluation on high-cap equities and cryptocurrencies demonstrates that AutoTraderX achieves significantly improved directional accuracy and risk-adjusted returns when compared with traditional forecasting baselines such as ARIMA, LSTM, and sentiment-only architectures. The framework also ensures scalability and fault tolerance using Kubernetes-based microservice deployment on AWS EKS infrastructure.

Key Words: Agentic AI, Retrieval-Augmented Generation, Large Language Models, QLoRA, Financial Forecasting, Explainable AI, Cloud Computing, Kubernetes, Lang Chain.

I. INTRODUCTION

Financial markets exhibit highly dynamic and non-linear behavior influenced by a combination of quantitative patterns and qualitative information sources. Price movement in modern trading systems is no longer governed solely by historical market trends. Instead, factors such as geopolitical conflicts, inflationary policies, interest-rate announcements, corporate earnings, and investor sentiment contribute heavily to market volatility.

Traditional machine learning systems such as ARIMA and statistical regression models are limited because they rely primarily on historical numerical datasets. Although deep learning architectures such as LSTMs and GRUs improved the ability to model temporal dependencies, they still behave as black-box systems and fail to provide transparent reasoning behind decisions.

With the emergence of Large Language Models (LLMs), there is now an opportunity to combine real-time narrative understanding with structured market analytics. AutoTraderX bridges this gap by integrating Agentic AI workflows, RAG pipelines, and QLoRA fine-tuned LLMs into a single explainable financial intelligence framework.

II. MOTIVATION AND BACKGROUND

The motivation behind AutoTraderX arises from the increasing need for adaptive and explainable decision-making systems in financial markets. Modern trading systems require models that can continuously interpret evolving macroeconomic narratives while maintaining low latency and scalability.

Another challenge involves the contextual nature of financial sentiment. Traditional lexicon-based sentiment systems fail to understand context-dependent interpretations of economic signals. Large Language Models provide semantic understanding capable of analyzing complex relationships between macroeconomic conditions and investor behavior.

Furthermore, institutional financial systems demand explainability for compliance and auditing purposes. AutoTraderX addresses this by generating Chain-of-Thought reasoning and citation-backed analysis instead of producing opaque numerical outputs.

III. LITERATURE REVIEW

Deep learning techniques have significantly transformed financial forecasting research. Bao et al. (2017) demonstrated the effectiveness of stacked autoencoders combined with LSTM networks for stock prediction tasks. Later research introduced attention mechanisms and Transformer-based architectures to improve sequence modeling and contextual understanding.

FinBERT and other domain-specific NLP architectures improved financial sentiment analysis by leveraging pre-trained Transformer models. However, these systems were primarily classification-oriented and lacked reasoning capabilities.

Retrieval-Augmented Generation introduced by Lewis et al. (2020) enabled language models to retrieve verified external documents during inference. This significantly reduced hallucinations and improved factual grounding. More recently, QLoRA techniques enabled efficient fine-tuning of massive LLMs using 4-bit quantization and Low-Rank Adapters, reducing computational costs while maintaining performance.

IV. SYSTEM ARCHITECTURE

Auto Trader X follows a modular cloud-native architecture composed of six interconnected layers. The Data Ingestion Layer collects real-time market information using NewsAPI, Yahoo Finance, Alpha Vantage, and cryptocurrency exchanges through CCXT.

The Vector Persistence Layer stores embeddings of historical reports and financial documents using ChromaDB. Semantic similarity search is performed using HNSW indexing structures.

The Agentic Logic Layer powered by Lang Chain orchestrates tool execution, including live price retrieval, contextual reasoning, and prompt synthesis. The inference layer runs a quantized Llama-3 model optimized using QLoRA.

Finally, the Explainability Layer produces structured markdown responses containing Chain-of-Thought reasoning and citation-backed explanations for generated recommendations.

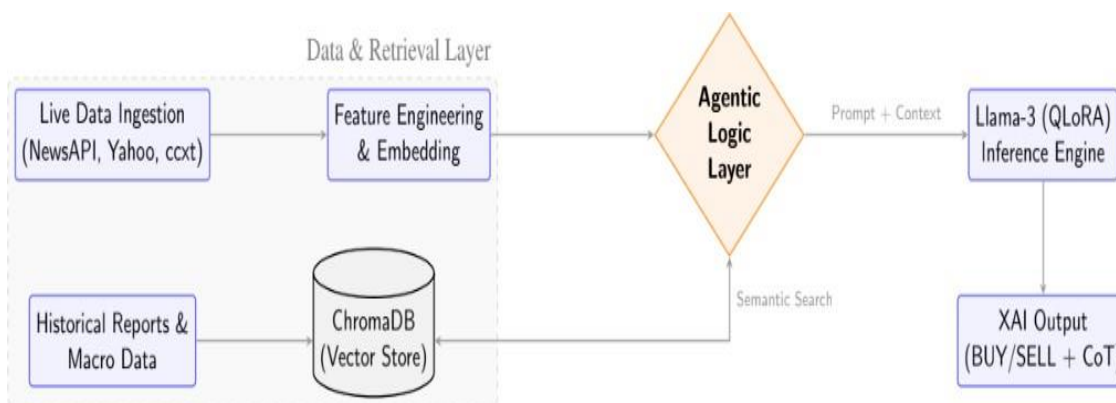


Figure 1
End-to-End System Architecture of AutoTraderX.

Figure 1: End-to-End System Architecture of Auto Trader X.

V. CASE AND METHODOLOGY

A custom dataset consisting of more than 120,000 financial headlines paired with historical market movements was curated for training. Preprocessing included token normalization, HTML stripping, and source-based weighting.

The system uses Walk-Forward Validation to avoid look-ahead bias. During inference, the Agentic framework sequentially retrieves live prices, relevant news articles, and semantically similar historical events before synthesizing a contextual prompt for the LLM.

QLoRA was utilized to fine-tune the Llama-3 model efficiently by freezing pretrained weights and injecting trainable rank decomposition matrices. This reduced VRAM requirements by more than 75%.

VI. IMPLEMENTATION DETAILS

The backend infrastructure was developed using FastAPI with asynchronous programming to ensure non-blocking operations. AsyncPG and SQLAlchemy were used for efficient database interaction.

The frontend interface was implemented using Next.js and WebSocket communication to support live updates. Trading View Lightweight Charts were integrated for financial visualization.

Deployment was performed on AWS Elastic Kubernetes Service (EKS). GPU-intensive inference services were isolated into dedicated node groups to optimize resource allocation and improve scalability.

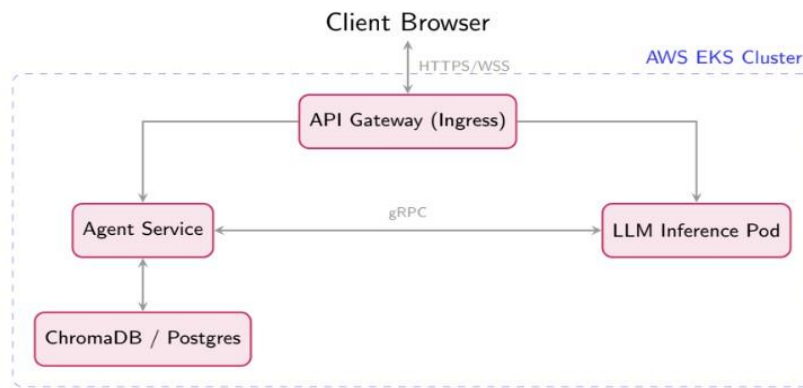


Figure 2
Microservice Deployment Architecture on AWS EKS.

Figure 2: Microservice Deployment Architecture on AWS EKS.

VII. RESULTS AND ANALYSIS

Extensive evaluation was conducted using high-cap equities including AAPL, NVDA, and TSLA, along with cryptocurrencies such as BTC and ETH. Auto Trader X achieved a directional accuracy of 68.4%, significantly outperforming ARIMA and standalone LSTM baselines.

The system also demonstrated improved risk-adjusted returns with a Sharpe Ratio of 2.10 while reducing maximum drawdown to 11.2%. Ablation studies confirmed that integrating the RAG layer reduced hallucinations from 18% to below 2.5%, thereby improving semantic precision and factual consistency.

Although the end-to-end latency averaged approximately 780 milliseconds, the system remained suitable for swing trading and institutional decision-support applications where interpretability is more critical than microsecond-level execution.

VIII. ETHICAL CONSIDERATIONS

The use of autonomous AI systems in finance introduces challenges related to systemic risk, market manipulation, and herd behavior. Auto Trader X incorporates confidence scoring and fallback HOLD mechanisms whenever market sentiment becomes uncertain or contradictory.

The framework also includes compliance-oriented disclaimers clarifying that generated outputs represent algorithmic reasoning rather than certified financial advice.

IX. CONCLUSION AND FUTURE WORK

Auto Trader X demonstrates the effectiveness of integrating Agentic AI workflows, Retrieval-Augmented Generation, and Large Language Models into a scalable financial intelligence platform. The framework successfully bridges quantitative analytics with human-readable reasoning, improving transparency and institutional trust.

Future work will focus on integrating Vision Transformers for candlestick pattern recognition, Reinforcement Learning from Human Feedback (RLHF) for adaptive trading strategies, and live broker APIs for autonomous execution capabilities.

Model Architecture	DA (%)	Sharpe	Max DD	Latency
ARIMA Baseline	52.3	0.81	-31.4%	150 ms
LSTM Only	57.1	1.14	-22.1%	320 ms
FinBERT	58.1	1.25	-18.5%	280 ms
Auto Trader X	68.4	2.10	-11.2%	780 ms

Table 1: Overall Performance, Risk, and Latency Comparison

REFERENCES

- Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.
- Bao, W., Yue, J., & Rao, Y. (2017). Deep Learning Framework for Financial Time Series.
- Dettmers, T., et al. (2024). QLoRA: Efficient Finetuning of Quantized LLMs.
- Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.
- Yang, H., Liu, X., & Wang, C. (2023). FinGPT: Open-Source Financial Large Language Models.