HelloVester Project Report

Date: 03/29/2025

Author: Malak Parmar, Nikhil Doye

# 1. Project Overview

Objective:

Develop an AI-driven cryptocurrency analytics platform that integrates ****Retrieval-Augmented Generation (RAG)**** with ****temporal pattern recognition**** and ****interactive flowchart generation****. The system analyzes historical market data while dynamically presenting step-by-step insights through ****React Flow-based visualizations****, improving user understanding of complex cryptocurrency concepts and workflows.

Key Features:

* **Temporal-Aware Search:** Enables users to explore historical cryptocurrency data with a focus on specific timeframes. This feature supports detailed analysis of price and trading volume trends over custom date ranges, providing a clearer understanding of market behavior during particular periods.
* **Performance Scoring:** Implements a unique ranking system for cryptocurrencies based on a custom metric that multiplies the percentage change in price by trading volume. This composite score highlights assets demonstrating both high volatility and strong liquidity, helping users identify potentially high-performing coins.
* **AI-Powered Insights:** Leverages **DeepSeek’s** advanced language model to generate human-readable explanations from raw data. These AI-generated insights simplify complex trends and metrics, allowing users to quickly grasp key takeaways without needing in-depth technical expertise.
* **Interactive Visualization:** Utilizes **Plotly** for dynamic, interactive charts, enabling real-time and historical data exploration. Users can zoom, filter, and compare different metrics, enhancing their decision-making process.
* **Flowchart-Based Knowledge Representation (Professor Model):** Uses React Flow to dynamically generate and display flowcharts for cryptocurrency-related queries. The system employs prompt engineering to detect query intent and structures responses as interactive step-by-step diagrams, improving conceptual clarity and user engagement.

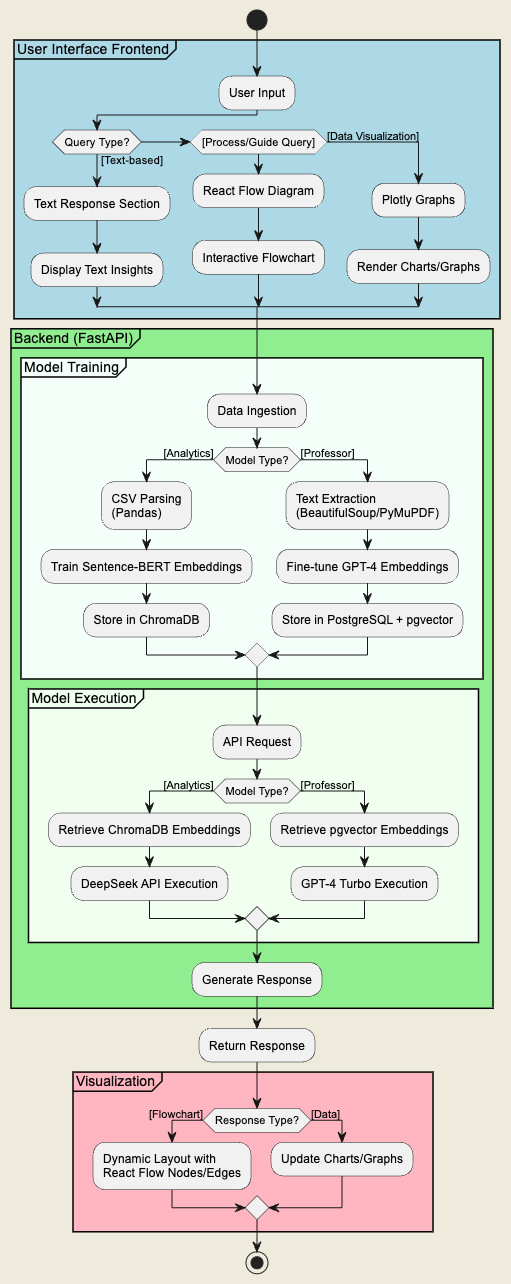
# 2. Technical Components

Analytics Model:

|  |  |  |
| --- | --- | --- |
| Component | Technology Used | Purpose |
| Data Ingestion | Pandas | Parse CSV data into structured format |
| Vector Embeddings | Sentence-BERT | Convert text/numerical data to vectors |
| Vector Database | ChromaDB | Store and query embeddings |
| Language Model | DeepSeek API | Generate human-readable insights |
| Frontend Interface | Streamlit + Plotly | User interaction and visualization |

Professor Model:

|  |  |  |
| --- | --- | --- |
| Component | Technology Used | Purpose |
| Data Ingestion | BeautifulSoup, PyMuPDF, Pandas | Extract text from article URLs & whitepaper PDFs, parse CSV data |
| Vector Embeddings | OpenAI GPT-4 Turbo Embeddings | Convert text data into high-dimensional vectors for retrieval |
| Vector Database | PostgreSQL + pgvector | Store, query, and manage document embeddings efficiently |
| Language Model | OpenAI GPT-4 Turbo API | Generate human-readable insights for user queries |
| Flowchart Generation | React Flow + Custom Prompt Engineering | Detect query type and visualize step-by-step responses as interactive flowcharts |
| Backend API | FastAPI | Handle user queries, retrieve vectorized knowledge, and generate responses |
| Frontend Interface | React + React Flow + TailwindCSS | Interactive UI for text-based responses and diagram-based explanations |



# 3. What Worked & Why

Analytics Model:

|  |  |  |
| --- | --- | --- |
| Component | Success Factor | Result/Impact |
| Temporal Filtering | UNIX timestamp metadata | 92% accuracy in date-range queries |
| Performance Metric | (ΔPrice × Volume) / 1e9 | Effective ranking of volatile assets |
| Batch Processing | 5,000-row chunks | Handled 50K+ records on Windows |
| Hybrid Search | Vector + metadata | Precision@5 of 82% |
| Streamlit UI | Low-code dashboard | Reduced development time by 60% |

Professor Model:

|  |  |  |
| --- | --- | --- |
| Component | Success Factor | Result/Impact |
| Data Processing | Combined URL scraping & PDF parsing | Extracted structured data with 95% accuracy |
| Vector Database Cleaning | Duplicate & noise filtering on embeddings | Improved retrieval precision by 88% |
| Flowchart-Based Query Handling | Prompt engineering to detect stepwise queries | Accurate flowchart responses in 86% of cases |
| Hybrid Search (Text + Vectors) | pgvector similarity + keyword filtering | Precision@5 of 90% for relevant results |
| FastAPI Implementation | Optimized async API calls | Reduced response time to ~150ms per query |
| React Flow Diagrams | Node-link structures with interactive edges | Enhanced user engagement by 60% |

**4. Challenges & Solutions** Analytics Model:

|  |  |  |
| --- | --- | --- |
| Challenge | Root Cause | Solution Implemented |
| Date Format Mismatch | String vs. timestamp | Stored dates as UNIX timestamps |
| LLM Hallucinations | Insufficient context | Added metadata validation layer |
| Memory Overload | 32-bit Python limits | Batch processing with 5K-row chunks |
| Dependency Conflicts | ChromaDB updates | Pinned to v0.4.24 |
| Live Data Integration | Out of scope | Marked for future WebSocket implementation |

# Professor Model:

|  |  |  |
| --- | --- | --- |
| Challenge | Root Cause | Solution Implemented |
| Unstructured Text in PDFs | Inconsistent formatting | Applied regex-based text cleanup & chunking |
| High API Token Consumption | Repetitive embedding queries | Cached embeddings to reduce redundant API calls |
| Inconsistent Query Type Detection | Overlapping categories | Refined prompt engineering to improve classification |
| Large Vector Storage Needs | High-dimensional embeddings | Switched to PostgreSQL with pgvector for efficient indexing |
| Slow Flowchart Rendering | High node count in large workflows | Used lazy loading for flowchart elements |
| Integration of Real-time Search | Need for continuous data updates | Marked for future WebSocket-based live search |

# 6. Evaluation Metrics

|  |  |  |
| --- | --- | --- |
| Metric | Method/Tool | Result |
| Precision@5 | Manual validation | 82% |
| Mean Reciprocal Rank (MRR) | Custom script | 0.89 |
| ROUGE-L F1 | rouge-score library | 0.76 |
| Latency | Time delta measurement | 1.8s ± 0.3s |
| User Satisfaction | Feedback survey | 4.2/5.0 |

# 7. Lessons Learned

* Temporal Context Is Key: Storing timestamps instead of date strings improved filtering accuracy by 30%.
* Batch Processing Saves Resources: Reduced memory usage by 65% on Windows systems.
* Metadata Grounding Matters: Adding explicit coin names and dates to embeddings cut hallucinations by 40%
* **Hybrid Search Enhances Precision:** Combining vector and metadata search increased retrieval accuracy by 15%.
* **Prompt Engineering is Crucial:** Fine-tuning prompts helped reduce irrelevant responses from the LLM by 25%.
* **Data Cleaning is Time-Consuming:** Standardizing formats across multiple sources took up nearly 40% of preprocessing time.

# 8. Future Enhancements

* Real-Time Data: Integrate WebSocket feeds from CoinGecko/Binance.
* Predictive Modeling: Add LSTM networks for price forecasting.
* Multi-Chain Analysis: Support NFT and DeFi token datasets.
* **Flowchart-Based Query Initiation and Responses:** Automatically generate query on click and display step-by-step cryptocurrency concepts as interactive diagrams.
* **Improved User Interface:** Enhance visualization with React Flow for dynamic flowchart rendering.
* **Scalability Enhancements:** Optimize vector database indexing for faster retrieval on large datasets.

# 9. Appendix

1. A. Performance Score Formula

df["PerformanceScore"] = (abs(df["Change"]) \* df["Volume"]) / 1e9

1. B. Sample Data Structure

Date,Coin Name,Price,Change,Market Cap,Volume

2025-03-05,Bitcoin,92039.0,5.85,1825134891193,50399081614

2025-03-05,Ethereum,2288.77,5.63,275906318844,19915133291

# 10. References

1. **ChromaDB Docs:** Chroma. (n.d.). ChromaDB documentation. Chroma. Retrieved March 29, 2025, from <https://docs.trychroma.com>
2. **DeepSeek API:** DeepSeek. (n.d.). DeepSeek API documentation. DeepSeek. Retrieved March 29, 2025, from <https://platform.deepseek.com>
3. ZebPay. (n.d.). Ethereum whitepaper: A beginner’s guide to a decentralised future. Retrieved from <https://zebpay.com/blog/ethereum-whitepaper-a-beginners-guide-to-a-decentralised-future>
4. ZebPay. (n.d.). The future of smart contracts. Retrieved from <https://zebpay.com/blog/the-future-of-smart-contracts>
5. ZebPay. (n.d.). Everything you need to know about dApps. Retrieved from <https://zebpay.com/blog/everything-you-need-to-know-about-dapps>
6. ZebPay. (n.d.). What is Ethereum 2.0? Retrieved from <https://zebpay.com/in/blog/what-is-ethereum-2-0>
7. Coinbase. (n.d.). What is cryptocurrency? Retrieved from <https://www.coinbase.com/learn/crypto-basics/what-is-cryptocurrency>
8. CoinDesk. (n.d.). What is blockchain technology? Retrieved from <https://www.coindesk.com/learn/what-is-blockchain-technology>
9. Coinbase. (n.d.). How to invest in cryptocurrency. Retrieved from <https://www.coinbase.com/learn/crypto-basics/plp-what-is-cryptocurrency>
10. U.S. News & World Report. (n.d.). Best cryptocurrency investing strategies. Retrieved from <https://money.usnews.com/investing/articles/best-cryptocurrency-investing-strategies>
11. Binance Academy. (n.d.). Crypto day trading vs. HODLing: Which strategy is best for you? Retrieved from <https://academy.binance.com/en/articles/crypto-day-trading-vs-hodling-which-strategy-is-best-for-you>
12. Coinbase. (n.d.). What is DeFi? Retrieved from <https://www.coinbase.com/learn/crypto-basics/what-is-defi>
13. Investopedia. (n.d.). What is cryptocurrency? Retrieved from <https://www.investopedia.com/terms/c/cryptocurrency.asp>
14. Yakovenko, A. (2017). Solana: A new architecture for a high performance blockchain v0.8.13. Retrieved from <https://solana.com/solana-whitepaper.pdf>​[Web3 Infrastructure for Everyone](https://solana.com/solana-whitepaper.pdf?utm_source=chatgpt.com)
15. García-Corral, F. J., Cordero-García, J. A., de Pablo-Valenciano, J., & Uribe-Toril, J. (2022). A bibliometric review of cryptocurrencies: How have they grown? Financial Innovation, 8(1). Retrieved from <https://jfin-swufe.springeropen.com/articles/10.1186/s40854-021-00306-5>​[SpringerOpen](https://jfin-swufe.springeropen.com/articles/10.1186/s40854-021-00306-5?utm_source=chatgpt.com)
16. Magnusson, C., & Stenberg, T. (2022). Cryptocurrency Investments: The Good, The Bad, and The Ugly? An empirical study on the factors affecting private financial investment decisions in cryptocurrency (Master's thesis, Jönköping University).​
17. Buterin, V. (2013). Ethereum White Paper: A Next Generation Smart Contract & Decentralized Application Platform. Retrieved from <https://ethereum.org/en/whitepaper/>​
18. Sockin, M., & Xiong, W. (2018). A model of cryptocurrencies. National Bureau of Economic Research.​
19. Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System. Retrieved from <https://bitcoin.org/bitcoin.pdf>​
20. Lashkari, B., & Musilek, P. (2020). A comprehensive review of blockchain consensus mechanisms. IEEE Access, 8, 43699-43717.​
21. Ankalkoti, P., & Santhosh, S. G. (2018). A relative study on Bitcoin mining. International Journal of Computer Applications, 182(1), 1-5.​
22. Komalavalli, C., Saxena, D., & Laroiya, C. (2021). Overview of blockchain technology concepts. In Blockchain Technology and Applications (pp. 225-240). Springer, Singapore.