LLM BASED CHATBOT FOR FINANCIAL ASSISTANCE

Aatheesh P V
Department of Artificial Intelligence and Data Science, Sri Ramakrishna Engineering College
Sri Ramakrishna Engineering College
Coimbatore, India

Mohammed Wasim B
Department of Artificial Intelligence and Data Science, Sri Ramakrishna Engineering College
Sri Ramakrishna Engineering College
Coimbatore, India

Adithya Sreedhar
Department of Artificial Intelligence and Data Science, Sri Ramakrishna Engineering College
Sri Ramakrishna Engineering College
Coimbatore, India

Mrs. K. Archana
Department of Artificial Intelligence and Data Science, Sri Ramakrishna Engineering College
Sri Ramakrishna Engineering College
Coimbatore, India

***Abstract*—** This paper introduces the development and deployment of a real-time chatbot tailored for financial news, leveraging three state-of-the-art technologies: Qdrant vector storage, the Gemini LLM, and the LLAMA index. In today's fast-paced financial markets, timely access to relevant news is crucial for informed decision-making. Traditional chatbots often struggle to analyze and understand data, leading to inaccuracies and delays in delivering insights and updates.

Moreover, we incorporate the advanced Gemini LLM model, specifically trained on financial text data, to enhance the chatbot's natural language understanding capabilities. By extensively leveraging financial news articles, Gemini undergoes refinement, enabling the chatbot to provide precise and intelligent responses to user queries. This is facilitated by the chatbot's thorough grasp of sentiment analysis, market trends, and financial terminology. The proposed chatbot architecture is deployed on a scalable and secure platform, ensuring reliable performance and data privacy.

***Keywords*—**LLM (Large Language Model), LLAMA, Qdrant Vector Storage.

I. INTRODUCTION

In the fast-paced realm of finance, staying abreast of the latest market developments is essential for traders, investors, and financial professionals. The challenge lies in swiftly accessing relevant information and leveraging it to make timely decisions amidst the continuous influx of reports, market analyses, and news updates. Traditional methods of monitoring financial news, such as relying on static news feeds or visiting multiple websites, are often deemed cumbersome and inefficient.

To address this issue, we propose the development of a real-time financial news chatbot—a sophisticated artificial intelligence (AI) system that utilizes natural language interaction to deliver timely updates, insights, and analysis to consumers. By harnessing real-time data sources and LLM models, the chatbot can sift through vast amounts of data, extract crucial insights, and present them to consumers in a clear and actionable manner.

II. LITERATURE SURVEY

According to Guendalina Caldarini, Sardar Jaf and Kenneth McGarry(2022) Chatbots are conversational computer programs adept at mimicking human speech, offer automated online assistance and guidance across various sectors. Their widespread adoption is driven by notable benefits they provide. Drawing from machine learning and natural language processing, both branches of artificial intelligence, chatbots utilize advanced techniques and algorithms. However, despite their advantages, there are challenges and limitations associated with their implementation. This overview explores recent advancements in chatbots leveraging artificial intelligence and natural language processing, highlighting key obstacles and constraints in existing research and proposing avenues for further investigation.

With Reference to Mikael tornwall, LMs lack statefulness, unless explicitly stated in the input prompt, an LM-based chatbot lacks awareness of prior interactions. Including the entire conversation history in the input prompt is often impractical due to length restrictions, demanding more financial and computational resources for larger prompts. This review introduces a short-term memory module designed to provide the chatbot with background information from previous exchanges, addressing this limitation. This review introduces two methods, the LimContext method and the Full Context method, aimed at generating an abstractive summary of conversation history. These methods condense a significant portion of relevant conversation history into a form that can be efficiently supplied with the input prompt, addressing resource constraints effectively.

According to the work of Inhwa Song, Sachin R. Pendse, Neha Kumar, Munmun De Choudhury (2024)**,** Large Language Model (LLM) chatbots are increasingly favored as mental health support tools for individuals experiencing extreme distress. Reports from social media highlight instances where such engagements have been credited with saving lives. However, data suggests that general-purpose

LLM chatbots pose substantial risks, potentially endangering user well-being if not developed responsibly. This study examines the lived experiences of individuals who have interacted with LLM chatbots for mental health support.

III. METHODOLOGY

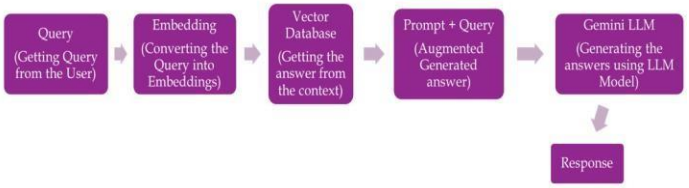


Fig 3.1 Block Diagram

The above block diagram shows the working of the proposed model. Data collection and Preprocessing is the first step to build an chatbot. An updated Financial News dataset is collected and preprocessed to clean the unnecessary data. These data are converted to vector embeddings and stored in Qdrant vector storage. Using ANN algorithm HNSW to retrieve the answer from the embeddings. Finally Gemini LLM model will be used to retrieve the answer based on the query given by the user.

A. Abbreviations and Acronyms

The Large Language Model (LLM) is a type of deep learning model that utilizes extensive training data to comprehend and generate language akin to human speech. Typically, LLMs are constructed using neural network architectures such as transformers and trained on vast collections of textual data, encompassing webpages, social media posts, books, and academic papers.The primary attribute of LLMs is their capability to capture the intricate patterns and structures inherent in natural language, enabling them to generate text that is coherent and contextually appropriate. This is achieved by their ability to predict, based on preceding context, the upcoming word or sequence of words in a sentence.

The abbreviation LLAMA, which stands for "Large Language Model Assessment," refers to a benchmarking package crafted to gauge the effectiveness of large language models (LLMs). It provides a standardized structure for appraising the language comprehension, generation, and reasoning abilities of LLMs, alongside other functionalities. Within the LLAMA suite, there exists a diverse array of tasks and datasets designed to evaluate LLM performance across various domains and linguistic phenomena. These tasks encompass language generation activities such as text summarization and completion, as well as language understanding tasks including question answering, text classification, and natural language inference.

B. Modules

Module 1- Data Preparation and Llama Index Embedding Module

- Pre processing the files containing natural language must be extracted, cleaned, and transformed into a format and an initial structure that an LLM can understand and also tend to perform better if one is able to isolate the most important natural language data and discard irrelevant data.
- Llama Index provides a simple wrapper over unstructured data in order to easily retrieve the parsed content and convert it into a format that Llama Index can ingest and manage the underlying data as an interface so that it can provide the correct inference during query-time.
- Llama Index integrates with FastEmbed for analyzing each news article and generate a numerical vector representing its meaning.
- The answer is indexing embeddings using different ANN algorithms such as HNSW. It is a graph-based algorithm that can efficiently handle billions of embeddings.

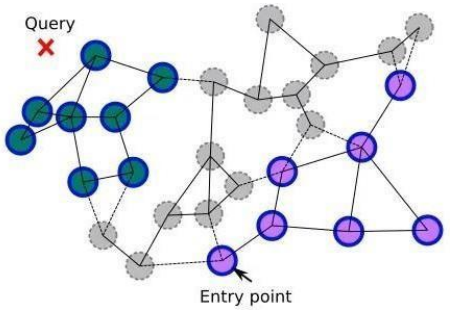


Fig 3.2 HNSW Algorithm

Module-2- Qdrant vector storage and retrieval module

- This module interacts with the vector store (Qdrant) to store the processed articles and their embeddings.
- Utilizing the LlamaIndex's storage context functionalities to connect with Qdrant and store the articles (potentially metadata) and their corresponding embeddings within Qdrant.
- It handles user queries and retrieves relevant news articles and leverage Llama Index's VectorIndexRetriever class to interact with the Qdrant vector store.
- When a user submits a query, LlamaIndex will Generate an embedding for the query using FastEmbed and the Gemini LLM model will search Qdrant for articles with embeddings closest to the query embedding based on a similarity metric (e.g., cosine similarity).
- Retrieve the top K most similar articles based on the chosen similarity threshold.

Module-3
Response generation and chat interface module

- This module utilizes Gemini LLM model and handles user interaction and chatbot responses.

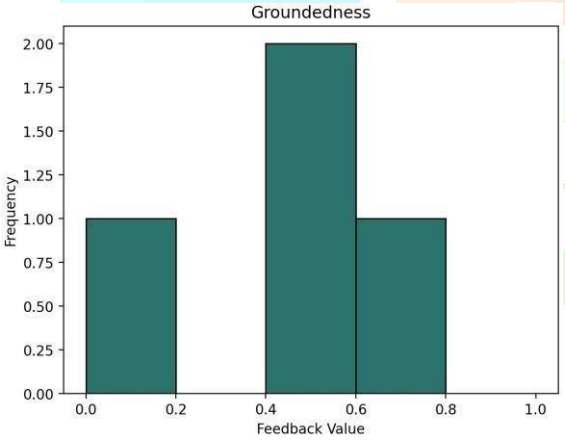
- The LLM model processes retrieved articles and generate informative responses and access the retrieved articles from the Retrieval Module.

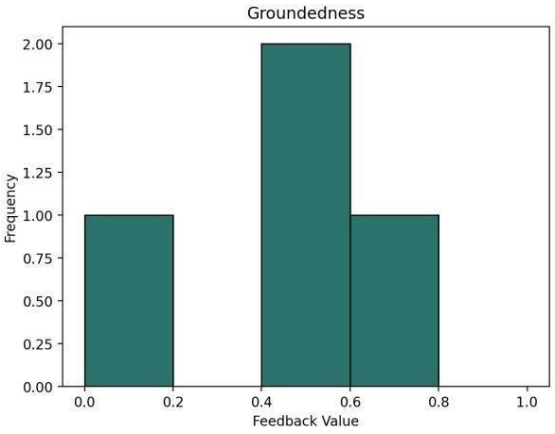
- To employ Gemini's LLM model capabilities to summarize the retrieved articles and analyzes the articles and identify relevant financial trends or insights based on the user's query to generate a response that combines retrieved snippets and Gemini-generated analysis.

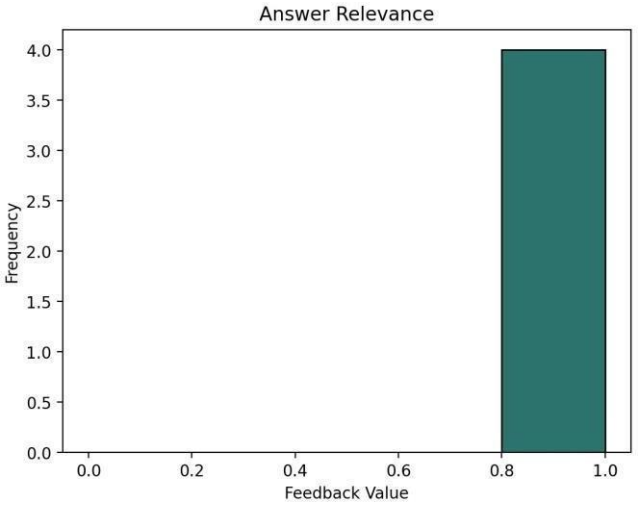
- Integrate the Retrieval and Response Generation modules to present retrieved articles and response-generated summaries/analyses to the user.

IV. RESULTS AND DISCUSSION

The objective of this initiative is to utilise the Gemini LLM model, a real-time financial news chatbot has been developed, demonstrating promising outcomes by delivering users in-depth analysis and timely updates. In this section, we delve into the key findings from our evaluation and the implications of our methodology. The chatbot adeptly furnishes users with current financial news updates sourced from diverse channels, ensuring they stay informed about market fluctuations.







The refinement of the Gemini LLM model on a vast corpus of financial news items has notably enhanced the chatbot's ability to comprehend and summarize complex financial data. Leveraging interaction with the LLAMA index and Qdrant vector storage, the chatbot efficiently retrieves and analyzes financial news articles in real-time, minimizing latency and ensuring a seamless user experience.

app_id	Groundedness	Context Relevance	Answer Relevance
App_1	0.7	0.5	1.0

Table 1- Performance of the models.

V. CONCLUSION

Throughout this project, we have utilized cutting-edge LLM models alongside the LLAMA index and Qdrant for real-time data processing. This integration has enabled the development of a comprehensive platform for retrieving, understanding, and analyzing financial news. Additionally, users benefit from access to the latest industry developments and trends facilitated by the chatbot's connection with real-time data sources such as the Qdrant vector store and the LLAMA index. In the fast-paced realm of finance, these real-time capabilities, coupled with personalized features like sentiment analysis and customized news alerts, enhance user engagement and facilitate informed decision-making.

VI. FUTURE SCOPE

Despite the overall success of the chatbot, challenges persist in optimizing the model for specific user preferences and specialized financial domains. Future investigations may focus on enhancing the chatbot's ability to handle multi-turn conversations, integrating additional data sources, and refining its understanding of financial terminology relevant to specific contexts.

Furthermore, advanced personalization options dependent on past interactions, user activity, and preferences should be

included. This could involve enhancing content recommendations, tailoring news alerts, and adjusting responses based on individual user profiles.

Implementing a feedback loop system to collect user input and interactions is crucial. Utilizing this data, continuous improvements can be made to enhance the accuracy and performance of the Gemini LLM model over time.

REFERENCES

[1] Huang, J., et al. "Financial Chatbot Evaluation: A Survey." Proceedings of the 15th International Conference on Web Information Systems and Technologies (2021).

[2] Zeng, J., et al. "FinBot: A Large-Scale Dataset for Financial Conversational AI." arXiv preprint arXiv:2110.07024 (2021).

[3] Roller, S., et al. "Recipes for building an open-domain chatbot." arXiv preprint arXiv:2004.13637 (2020).

[4] Lewis, M., et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020).

[5] Lewis, M., et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020).

[6] Liu, Y., et al. "Fine-Tuning Large Language Models for Retrieval: Challenges and Solutions." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020).

[7] Brown, T. B., et al. "Language Models are Few-Shot Learners." arXiv preprint arXiv:2005.14165 (2020).