Leveraging Language Models for Enhanced Stock Advisory: A Comprehensive Analysis

Shraddha Sundaresan
MPS Data Science
University of Maryland Baltimore County
Baltimore, USA
shradds1@umbc.edu

***Abstract*—This research introduces an innovative stock investment advising model using a Language Model, with a focus on Retrieval Augmented Generation. Real-time stock values are sourced from various platforms, including Yahoo Finance, and company-specific news is extracted from Bing News. The integration of Retrieval Augmented Generation methodology enhances the model's performance through retrieval system, significantly reducing hallucination risks. Prompt engineering optimizes the system's ability to generate insightful responses. The model evaluates the amalgamated data, offering suggestions on the advisability of investing in specific stocks. Beyond highlighting the technical intricacies of LLM implementation, the study explores synergies between financial data and sentiment analysis from news sources. The project's outcomes contribute to advancing natural language processing models in financial decision-making, offering a more reliable approach to stock advising through the effective mitigation of hallucination risks with the Retrieval Augmented Generation implementation.**

***Index Terms*—Financial Advising Automation, Large Language Models, LangChain, Web Scraping, Retrieval Augmented Generation**

I. Introduction

"In the dynamic landscape of financial markets, investors reign supreme. In the era of technological prowess, where the heartbeat of modern finance is dictated by advancements like LangChain, the realm of stock advisory question and answers emerges as the forefront of interaction between investors and cutting-edge intelligence. Here, questions find answers, concerns meet resolutions, and a foundation of trust is meticulously crafted."

This project is dedicated to enhancing investment decision-making for specific US companies' shares through a user-driven prompt-based system.

The key objective is to streamline the decision-making process by allowing users to input prompts that align with their specific queries or interests related to stock investments. This approach introduces a user-centric interaction model, providing a valuable tool for individuals seeking efficient and prompt-specific advice on the potential of investing in US company shares.

The structure of this paper is as follows: Section 2 presents the related work; Section 3 covers the methodologies and components used towards creating the application. Section 4 provides a discussion of the study, Section 5 concludes the paper by presenting the key findings, summarizing the conclusions. Lastly, Section 6 proposes potential future enhancements.

II. Background and Related Work

A. *Financial Sentiment Analysis*

In the early stages, financial sentiment analysis primarily involved the refinement of pre-trained models. The focus was on fine-tuning these models to cater specifically to the nuances of the financial domain. However, as financial news became increasingly nuanced, incorporating numerical information and lacking background context, the limitations of this approach became apparent.

As the financial landscape evolved, the need for a more sophisticated understanding of textual information, especially in the context of financial news, led to a realization that finetuning alone might not capture the intricacies and contextual dependencies required for accurate sentiment analysis.

B. *Instructional Fine-Tuning*

To overcome the challenges posed by the limitations of sentimental analysis, researchers transitioned towards instructional fine-tuning. Recent models, such as GPT-3 and LLaMA, embody this approach [1][2].

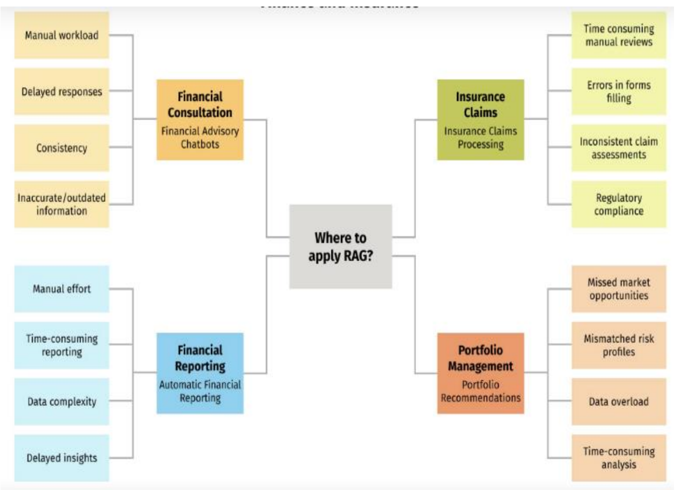


Fig. 1. Types of RAG implementations in financial domain

This phase involved a more nuanced strategy: refining Large Language Models (LLMs) through exposure to natural language task descriptions. The objective was to guide these models in adhering more closely to specific user instructions, thereby enhancing their capacity for contextual understanding.

Despite the advantages brought by instructional fine-tuning, including improved adaptability to user instructions, challenges persisted. The unpredictability in the outputs of these finely-tuned models became a focal point, prompting the exploration of alternative approaches that could mitigate these issues while maintaining enhanced contextual understanding.

C. Retrieval Augmented Generation

In response to the challenges encountered in the instructional fine-tuning phase, the financial domain witnessed a paradigm shift with the adoption of Retrieval Augmented Generation (RAG). RAG introduced a novel two-step process that departed from the conventional fine-tuning strategies. This approach combined context retrieval with LLMs, utilizing two distinct knowledge sources to augment the generation process.

The dual-knowledge approach of RAG involved retrieving relevant documents through a retrieval module based on the input prompt. These documents, sourced from external knowledge bases, provided additional context. The retrieved information, combined with the original input, was then fed into LLMs for the generation of contextually relevant outputs.

The adoption of RAG marked a significant evolution in addressing the limitations of prior methodologies. The synergistic use of parametric memory stored in LLMs' parameters and nonparametric memory obtained from the corpus of retrieved documents allowed for more accurate, contextually aware, and relevant generation [3]. This evolutionary step reflects an ongoing commitment to refining methodologies for financial sentiment analysis in response to the dynamic nature of financial information and user expectations.

D. Portfolio Management

Retrieval Augmented Generation (RAG) can play a pivotal role in portfolio management by leveraging its ability to retrieve relevant financial documents, market analyses, and historical data. The dual knowledge approach of RAG, utilizing both parametric memories stored in the model parameters and nonparametric memory from retrieved documents, enables it to provide contextually relevant information for making informed investment decisions. This application ensures that portfolio managers have access to a wealth of up-to-date and pertinent information when optimizing and adjusting investment portfolios.

E. Insurance Claims Processing:

In the context of insurance, RAG can streamline the claims processing workflow. By retrieving relevant information from external knowledge bases, such as legal documents, policy details, and historical claims data, RAG can assist in generating accurate and context-aware responses.

This can enhance the efficiency of claims assessment, reduce processing time, and improve overall customer satisfaction. The model's ability to amalgamate retrieved information with the user's input ensures that the responses are not only accurate but also tailored to the specific context of the insurance claim.

F. Financial Reporting

RAG can be employed in the generation of financial reports by retrieving and synthesizing information from various sources, including financial statements, market trends, and regulatory updates. This application ensures that financial reports are not only comprehensive but also reflective of the latest market conditions. The dual-knowledge approach allows RAG to consider both the inherent knowledge within its parameters and the dynamic information retrieved from external documents, contributing to the production of accurate and contextually relevant financial reports.

G. Financial Consultation:

In financial consultation, RAG can serve as an intelligent assistant, providing personalized and contextually relevant information to clients. By retrieving information from financial databases, market analyses, and investment strategies, RAG can assist financial consultants in offering tailored advice and insights. The model's ability to integrate retrieved knowledge with the client's queries ensures that the consultation process is not only informative but also adaptive to the specific needs and goals of the individual.

The website, inspired by these insights, is poised to adopt a multifaceted approach. It aims to harness the potential of generative AI models for diverse tasks, emphasizing user-friendly interactions through natural language prompts. Simultaneously, it recognizes the challenges of prompt engineering and intends to incorporate strategies proposed by researchers to enhance the alignment between user prompts and LLM understanding. Furthermore, the website seeks to contribute to the ongoing narrative on AI and employment, aligning with Professor Bengio’s perspective of digital transformation leading to role evolution rather than job displacement. The overarching goal is to present a comprehensive understanding of the capabilities and challenges associated with generative AI models while fostering a positive and nuanced discourse on AI’s impact on employment.

III. Methodology and Components

This section covers the data collection, details about the workflow, and integration with streamlit for a user-interface.

A. About the Data

In this project, comprehensive data is collected from diverse sources to ensure a robust foundation for the stock investment model. The primary data sources include Yahoo Finance for real-time and historical stock values and Bing for current and historical news related to specific companies. The rationale behind the choice of these platforms is rooted in the need for a holistic understanding of the companies’ current performance in both the stocks fluctuation and news domains.

Data from Yahoo Finance serves as a vital component, providing dynamic insights into the real-time performance of the companies’ stocks over a three-month period. By aggregating this financial data, the model gains a nuanced understanding of the companies' market trends, enabling more informed recommendations for potential investors.

Scraping current and historic news related to the selected companies from Bing News complements the financial data. This inclusion allows the model to factor in external influences, sentiments, and events that might impact stock performance. Integrating news data enhances the depth of analysis, providing a more comprehensive view of the companies' current standing in the market.

B. Data Collection

For this project, a dataset spanning three months of specific stock information from Yahoo Finance is complemented by a parallel collection of company-related news extracted over the same period from Bing. To facilitate analysis, this amalgamated dataset is exported as a text file, serving as the input for the final model.

C. Data Embedding and Vectorstore

Following data collection, the text file is segmented into smaller, semantically relevant chunks using the CharacterTextSplitter. Subsequently, these chunks undergo embedding using the OpenAI embedding and are stored in the vector database (Chroma db) [4]. This process is pivotal for transforming textual information into numerical vectors, facilitating efficient storage and retrieval. The use of embeddings enhances the model's ability to analyze and interpret the intricate relation- ships within the data, contributing to the robustness of the subsequent analysis and recommendation stages.

D. Prompt Template

To guide the model effectively, a purpose-specific prompt is crafted, communicating its role as a financial advisor tasked with assisting individuals in making sound financial decisions. The prompt is strategically designed to mitigate hallucination by incorporating contextual cues, ensuring responses align with the given financial context. This template serves as a crucial component in refining the model's understanding and aligning its responses with the intended objective of providing accurate and contextually appropriate stock investment advice.

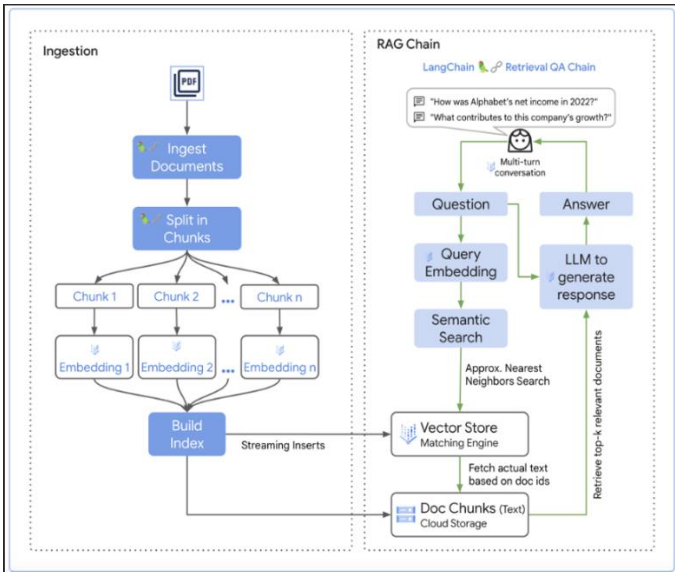


Fig.2. RAG workflow of Q and A engine

E. Building a RetrievalQA Chain for Question Answering

The construction of a RetrievalQA chain involves the integration of a language model and a vector database as a retriever, facilitating accurate question answering. To initiate this process, a language model, specifically the ChatOpenAI model, is initialized. The temperature is set to zero, optimizing the model for factual answers by minimizing variability and ensuring the highest fidelity.

The RetrievalQA chain leverages the "stuff" method by default, consolidating all relevant documents into the final prompt. This streamlined approach involves a singular call to the language model, simplifying the retrieval and answer-generation process.

F. Passing Query to the Chain

In the conclusive stage of the process, a query is introduced to the RetrievalQA chain, culminating in the generation of the ultimate suggestion. This pivotal step activates the integrated capabilities of the language model and the retrieval system, leveraging the established chain to of subjectivity and market volatility, contributing to the system's inclination toward suggestive responses. The integration of the RetrievalQA chain, encompassing the language model and vector database, plays a vital role in offering insights. However, the dynamic and speculative nature of investment decisions inherently introduces an element of uncertainty, influencing the nature of the suggestions provided.

G. Deploying Web Interface

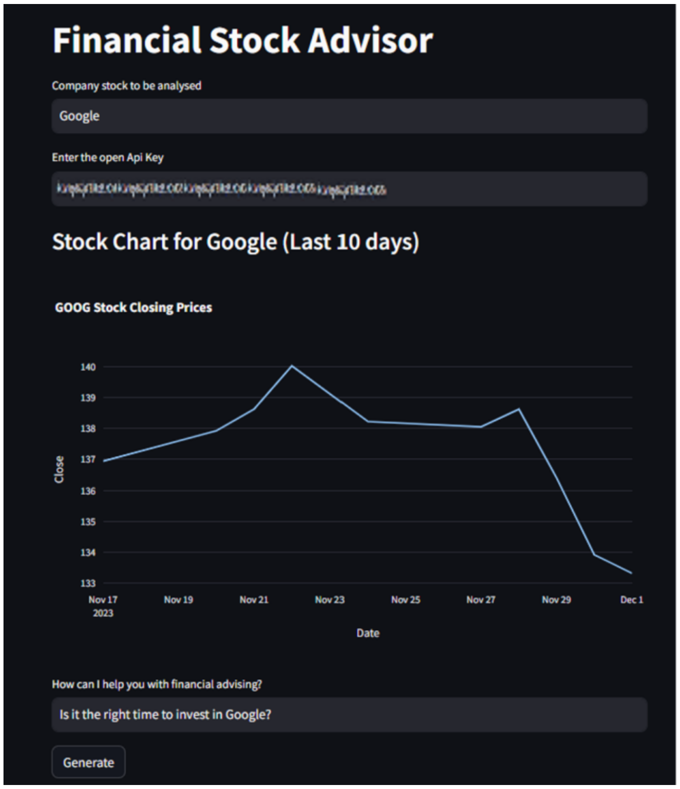
In the realm of financial decision-making, the deployment of a web application designed to offer investment suggestions requires consideration of various inputs, all of which are seamlessly integrated into an interface built with the sophisticated capabilities of Streamlit.

The company name acts as a centralizing element, directing the app's analysis to a specific entity. Leveraging BeautifulSoup along with the OpenAI API key, the application utilizes the company name to fetch and present detailed financial data and historical performance metrics.

In summary, this research underscores the significance of each input within the context of a Streamlit built web application designed for investment suggestions. The Open API key, the company name, and user-generated questions collectively contribute to a comprehensive, user-centric interface that seamlessly integrates external financial data and empowers users to make informed investment decisions.

IV. Results and Discussion

Upon querying, "Is it the right time to invest in Google shares?" the system generates a suggestion that leans more towards a suggestive nature rather than presenting a conclusive response. This outcome prompts a closer examination of the contextual factors influencing the suggestion generation process. The nature of financial queries often involves a degree of subjectivity and market volatility, contributing to the system's inclination toward suggestive responses. The integration of the RetrievalQA chain, encompassing the language model and vector database, plays a vital role in offering insights. However, the dynamic and speculative nature of investment decisions inherently introduces an element of uncertainty, influencing the nature of the suggestions provided.



Financial Stock Advisor

Company stock to be analysed

Google

Enter the open Api Key

[redacted]

Stock Chart for Google (Last 10 days)

How can I help you with financial advising?

Is it the right time to invest in Google?

Fig.3. Diagram of the web interface

V. Future Enhancements

Refinement of Prompt Templates: Explore and refine prompt templates to provide more specific and contextually rich queries, enabling the model to generate more precise and conclusive suggestions.

Diversification of Training Data: Expand and diversify the training dataset for the language model to expose it to a broader range of financial scenarios and market conditions, enhancing its understanding and adaptability.

User Preference Learning: Implement mechanisms for the model to learn and adapt to individual user preferences over time, creating a more personalized and user-centric advisory experience.

By focusing on these future enhancements, the goal is to create a more sophisticated, adaptable, and user-friendly system that significantly improves the quality and reliability of financial suggestions provided by the model.

VI. Conclusion

In concluding this project, the implementation of a Retrieval QA chain for question answering, integrating a language model and vector database, has shown promising results. The success of the system lies in its ability to leverage both the language model and retrieval steps, providing valuable insights for users seeking financial advice.

Moving forward, ongoing refinements in the prompt template, retrieval process, or language model training may further enhance the system's capacity to deliver more nuanced and conclusive suggestions. This project serves as a steppingstone, emphasizing the continual evolution of AI-driven financial advising models. As the system advances, addressing the complexities of financial decision-making remains paramount, and future iterations will aim to strike a balance between suggestion and certainty, fostering a more informed and reliable tool for users navigating the dynamic landscape of stock investments.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan,Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, AmandaAskell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan,Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jefrey Wu, Clemens Winter,Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, BenjaminChess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, IlyaSutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. InProceedings of the 34th International Conference on Neural Infor- mation ProcessingSystems (Vancouver, BC, Canada) (NIPS'20). Curran Associates Inc., Red Hook,NY, USA, Article 159, 25 pages.

[2] Nashipudimath, Madhu Shinde, Subhash Jain, Jayshree. (2020). An efficient integration and indexing method based on feature patterns and semantic analysis for big data. Array. 7. 100033. 10.1016/j.array.2020.100033. I.S. Jacobs and C.P. Bean, Fine particles, thin films and exchange anisotropy, in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[3] Su, Hongjin, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. " One Embedder, Any Task: Instruction-Finetuned Text Embeddings." ArXiv, (2022). /abs/2212.09741.

[4] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. InPars: Unsupervised Dataset Generation for Informa- tion Retrieval. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22). Association for Computing Machinery, New York, NY, USA, 2387–2392. https://doi.org/10.1145/3477495.3531863.