Fin-RAG: RAG BASED FINANCIAL ANALYSIS SYSTEM

1st Aditya Arya *2021510* 2nd Siddharth Anand *2021494* 3rd Om Garg *2021481* 4th Tanuj Kamboj *2021497* 5th Vibhav Balhara *2021500* 6th Vivek Anand *2021503*

I. Problem Statement

In the fast-paced world of financial markets, the challenge of utilizing the enormous daily influx of news and articles for trend prediction is particularly acute. The Indian market, with its unique complexities and rapid changes, necessitates the development of a sophisticated system capable of efficiently processing and analyzing vast amounts of textual data to accurately forecast market trends. Additionally, the task of navigating through and extracting actionable insights from intricate financial reports and articles can be daunting for users, potentially hindering their understanding and ability to make informed decisions.

To address these challenges, we propose the construction of an advanced financial system that employs Information Retrieval (IR) techniques not only to derive insights from daily news and articles for trend prediction in the Indian market but also to facilitate easier access to critical information. Moreover, we aim to develop a comprehensive Retrieve-and-Generate (RAG)-based system that can respond to user queries by summarizing and simplifying the content from complex financial reports and articles. This will enhance user comprehension and support more effective decision-making.

Furthermore, for those interested in obtaining detailed informational data from a company's financial statements, this system will include capabilities to parse and analyze these documents. By integrating specific NLP techniques tailored for financial documents, the system will enable users to extract key financial metrics and narratives, thereby providing a deeper understanding of a company's financial health and future prospects. This functionality is essential for investors who need to perform thorough due diligence and for analysts who require precise information to build investment theses or company profiles.

II. Motivation

- RAG models can improve the accuracy of retrieved financial data by integrating powerful language understanding capabilities. This ensures that the financial statements are relevant and precisely match the query requirements.
- Automating the retrieval of financial statements reduces the time and effort required by analysts to manually search through vast amounts of data. This increases efficiency and allows for quicker decision-making.
- By using a RAG-based system, the process can scale to accommodate the retrieval of data from a large number of companies without a significant increase in operational complexity or resource requirement.
- Financial markets are dynamic, and having a system that can provide real-time or near-real-time data retrieval can be crucial for making timely financial decisions and analyses.
- With more accurate and timely data, financial analysts can perform more complex and precise analyses, potentially leading to better investment decisions and market predictions.
- Ensuring compliance with financial regulations requires access to accurate and current financial statements. A RAG-based system can facilitate this by ensuring that all retrieved documents are up-to-date and compliant with the latest regulations.
- Reducing the reliance on human labor for data retrieval and increasing the speed of data processing can significantly reduce costs associated with financial analysis and research.

III. Literature Review

A. *Advancements in Financial Report Chunking*

In "Financial Report Chunking for Effective Retrieval Augmented Generation," a transformative approach to chunking in Retrieval Augmented Generation (RAG) systems is presented, which focuses on the structural elements of documents rather than traditional paragraph-level segmentation[1]. This method significantly enhances the precision and relevance of information retrieved, proving especially effective in complex, structured financial documents where traditional methods fall short. By utilizing structural elements for chunking, RAG systems can offer more contextually relevant information, improving decision-making processes by providing stakeholders with accurate and pertinent data. This innovation in chunking marks a substantial advancement in handling sophisticated document structures across various data-intensive sectors, including finance.

B. *Enhancements in RAG for Financial Question Answering*

"Improving Retrieval for RAG based Question Answering Models on Financial Documents" by Setty, Jijo, Chung, and Vidra, introduces innovative methodologies to refine text retrieval processes within Retrieval Augmented Generation

Identify applicable funding agency here. If none, delete this.

(RAG) systems for financial documents[2]. Recognizing that the effectiveness of Large Language Models (LLMs) is contingent on the quality of input, this paper emphasizes enhancing RAG techniques to ensure more accurate and relevant responses. By incorporating advanced chunking techniques, query expansion, metadata annotations, re-ranking algorithms, and fine-tuning embedding algorithms, the authors propose a series of enhancements that address existing limitations in RAG systems. These improvements aim to increase the precision of text retrieval, significantly boosting the performance and reliability of LLMs in responding to complex financial queries.

C. Enhancing Stock Investment Strategies through Annual Report Analysis with Large Language Models

Gupta's 2023 study, "GPT-InvestAR: Enhancing Stock Investment Strategies through Annual Report Analysis with Large Language Models," introduces a transformative approach using LLMs to analyze complex annual reports, traditionally a labor-intensive process[3]. This integration not only automates the extraction of key financial data, enhancing efficiency and reducing the need for expert analysis, but also incorporates machine learning to improve the accuracy and depth of stock performance predictions. This advancement signifies a major leap in financial analytics, providing more precise and actionable insights for investment strategies, showcasing a novel and promising direction in the field (Gupta, 2023).

IV. Novelty

Real-Time Financial Statement Prediction from User-Provided Documents

The proposed novelty in our RAG-based information retrieval system is the capability to predict future financial statements and trends directly from the latest financial documents provided by the user in PDF format. This innovative approach enables real-time, contextual analysis of financial data, offering insights that are immediately relevant and highly personalized. By extracting and utilizing both structured and unstructured data from these documents—ranging from numerical data in tables to narrative descriptions in text—the system can provide comprehensive predictions about financial health, operational efficiency, and potential future outcomes. This method is particularly valuable for making timely decisions based on the most current financial information available, and it addresses the challenge of rapidly changing financial environments where the latest data can significantly influence forecasts.

Moreover, this system incorporates advanced natural language processing and machine learning algorithms to adaptively learn from new data as it becomes available, enhancing its predictive accuracy over time. Such features also allow the system to handle the diverse presentation styles and formats used in different companies' financial documents, making it robust and versatile across various sectors. Additionally, the system is designed to recognize and adjust to regulatory changes that affect financial reporting standards, ensuring compliance and relevance in its predictions. Through these capabilities, the system not only supports financial analysts and corporate decision-makers but also empowers individual investors and stakeholders by providing them with tools to analyze and predict financial trends from merely a single document input. This approach innovates in the realm of financial analytics by integrating cutting-edge technology with user-driven data input, creating a dynamic, responsive tool that advances the frontier of financial information systems.

V. Methodology

The methodology for implementing the real-time financial statement prediction system from user-provided PDF documents involves several key steps, outlined below. Each step is crucial for ensuring the accurate extraction, processing, and prediction of financial data.

1. **Document Reading and Pre-processing**: Tools Used: PyPDF2 for PDF text extraction, Camelot for table reading. Process: Initially, the PDF document is read using PyPDF2 to extract text data. Camelot is employed to extract data from tables within the PDF. This helps in parsing structured financial data accurately. The extracted text data is then segmented into chunks of 100 words each, with an overlap of 30 words to ensure continuity and context preservation across chunks. Text data from tables is converted into .txt files for easier processing and integration with textual data chunks.

2. **Data Mapping and Word Embedding**: Mapping: A mapping is created between the text chunks and the corresponding tables extracted. This links narrative content with structured data, facilitating a holistic analysis. Word Embedding: Word embeddings are generated for each chunk to transform textual information into a numerical format that can be processed by machine learning models. This step is crucial for preparing the data for further analysis and ensuring that textual data can be effectively compared and analyzed.

3. **Contextual Analysis Using NLP Models**: Condition Based Analysis: If the stock mentioned in the document exists in our pre-existing database, a TF-IDF (Term Frequency-Inverse Document Frequency) model is used to analyze the text. This helps in identifying the importance of words in relation to the document set, focusing on financial terms specific to the known stocks. If the stock is not present in the database, BERT (Bidirectional Encoder Representations from Transformers) is employed to understand the context and extract meaning from the new or less common financial data. LLM Integration: The embeddings and contextual data are then passed to large language models (LLMs) like Llama and Falcon. These LLMs are used to perform advanced predictions and analyses based on the combined textual and structured data.

4. **Performance Comparison and Analysis**: The outputs from Llama and Falcon models are compared to evaluate which model performs better in terms of accuracy, reliability, and relevance of the financial predictions. This comparative

analysis helps in refining the model selection for specific types of financial documents and prediction needs.

VI. Database and Data Usage

A. Database Setup

Data Collection:

- Collected financial statements from the National Stock Exchange (NSE) website, covering a period of the last 30 years.
- Also gathered corporate news data spanning the same 30-year period.
- Conducted thorough data cleaning to eliminate repeated and irrelevant news items.

Data Chunking and Mapping:

- Segmented the textual data into manageable chunks to facilitate detailed analysis.
- Established a mapping between these data chunks and corresponding tables to enhance contextual relevance and data retrieval accuracy.

B. Data Usage

Stock-Based Queries:

- Implemented TF-IDF (Term Frequency-Inverse Document Frequency) analysis for stock-based queries.
- Stored TF-IDF vectors for text chunks to quickly assess the relevance of text based on stock-related queries.
- Preserved TF-IDF vectorizers specifically tailored for the stocks dataset, enabling dynamic query processing and efficient data retrieval.

General Queries:

- For queries not directly related to specific stocks, utilized BERT (Bidirectional Encoder Representations from Transformers) embeddings.
- Stored BERT embeddings for text chunks to ensure robust handling of general queries by capturing nuanced contextual meanings.

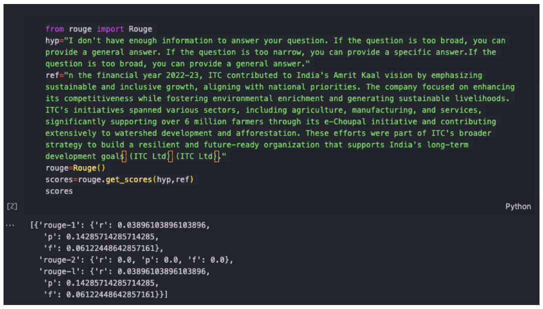
Structured Data Storage:

- Created a dedicated folder to store tables extracted from financial documents in a structured format.
- This organization facilitates straightforward access to structured data for processing by Large Language Models (LLM) such as Llama and Falcon.

VII. Evaluation

In this study, we evaluate the performance of two advanced language models, Llama 2.0 and Falcon, in summarization tasks by analyzing their Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores. ROUGE is a widely recognized metric used to assess the quality of text summaries by measuring the overlap of n-grams, longest common subsequences, and skip-bigrams between the machine-generated text and a reference. For our analysis, we employ text generated by ChatGPT 4.0 as the reference standard, owing to its established accuracy and reliability in text generation.

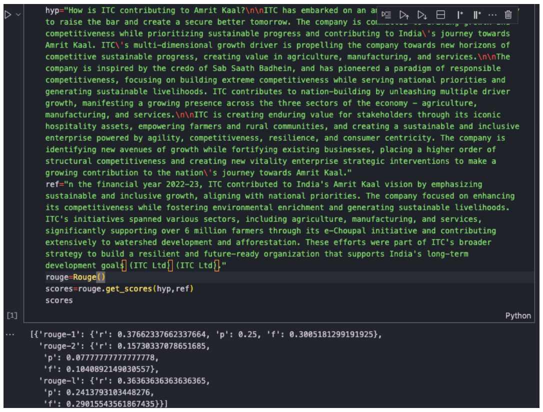
Our comparative analysis involves calculating ROUGE scores for summaries produced by both models across a diverse set of texts, thereby providing insights into each model's ability to capture essential information and present it concisely. This methodology not only highlights the strengths and weaknesses of each model in terms of text generation and summarization capabilities but also aids in understanding the practical implications of deploying these models in real-world applications. Through this detailed evaluation, we aim to discern which model demonstrates superior performance in creating accurate and coherent summaries, thereby guiding future developments in natural language processing technologies.



```
from rouge import Rouge
hyp="I don't have enough information to answer your question. If the question is too broad, you can
provide a general answer. If the question is too narrow, you can provide a specific answer.If the
question is too broad, you can provide a general answer."
ref="n the financial year 2022-23, ITC contributed to India's Amrit Kaal vision by emphasizing
sustainable and inclusive growth, aligning with national priorities. The company focused on enhancing
its competitiveness while fostering environmental enrichment and generating sustainable livelihoods.
ITC's initiatives spanned various sectors, including agriculture, manufacturing, and services,
significantly supporting over 6 million farmers through its e-Choupal initiative and contributing
extensively to watershed development and afforestation. These efforts were part of ITC's broader
strategy to build a resilient and future-ready organization that supports India's long-term
development goals (ITC Ltd) (ITC Ltd)."
rouge=Rouge()
scores=rouge.get_scores(hyp,ref)
scores
```

```
[{'rouge-1': {'r': 0.03896103896103896,
   'p': 0.14285714285714285,
   'f': 0.06122448642857161},
  'rouge-2': {'r': 0.0, 'p': 0.0, 'f': 0.0},
  'rouge-l': {'r': 0.03896103896103896,
   'p': 0.14285714285714285,
   'f': 0.06122448642857161}}]
```

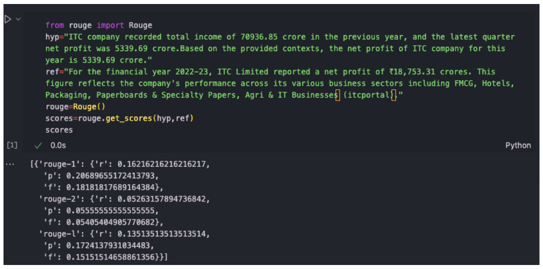
Fig. 1. Query 1.Rouge score of the generated output from Flacon(GPT4.0 content as reference).



```
hyp="How is ITC contributing to Amrit Kaal?\n\nITC has embarked on an a
to raise the bar and create a secure better tomorrow. The company is co
competitiveness while prioritizing sustainable progress and contributing to India\'s journey towards
Amrit Kaal. ITC\'s multi-dimensional growth driver is propelling the company towards new horizons of
competitive sustainable progress, creating value in agriculture, manufacturing, and services.\n\nThe
company is inspired by the credo of Sab Saath Badhein, and has pioneered a paradigm of responsible
competitiveness, focusing on building extreme competitiveness while serving national priorities and
generating sustainable livelihoods. ITC contributes to nation-building by unleashing multiple driver
growth, manifesting a growing presence across the three sectors of the economy - agriculture,
manufacturing, and services.\n\nITC is creating enduring value for stakeholders through its iconic
hospitality assets, empowering farmers and rural communities, and creating a sustainable and inclusive
enterprise powered by agility, competitiveness, resilience, and consumer centricity. The company is
identifying new avenues of growth while fortifying existing businesses, placing a higher order of
structural competitiveness and creating new vitality enterprise strategic interventions to make a
growing contribution to the nation\'s journey towards Amrit Kaal."
ref="n the financial year 2022-23, ITC contributed to India's Amrit Kaal vision by emphasizing
sustainable and inclusive growth, aligning with national priorities. The company focused on enhancing
its competitiveness while fostering environmental enrichment and generating sustainable livelihoods.
ITC's initiatives spanned various sectors, including agriculture, manufacturing, and services,
significantly supporting over 6 million farmers through its e-Choupal initiative and contributing
extensively to watershed development and afforestation. These efforts were part of ITC's broader
strategy to build a resilient and future-ready organization that supports India's long-term
development goals (ITC Ltd) (ITC Ltd)."
rouge=Rouge()
scores=rouge.get_scores(hyp,ref)
scores
```

```
[{'rouge-1': {'r': 0.37662337662337664, 'p': 0.25, 'f': 0.3005181299191925},
  'rouge-2': {'r': 0.15730337078651685,
   'p': 0.07777777777777778,
   'f': 0.1040892149030557},
  'rouge-l': {'r': 0.36363636363636365,
   'p': 0.2413793103448276,
   'f': 0.29015543561867435}}]
```

Fig. 2. Query 1.Rouge score of the generated output from Llama 2.0(GPT4.0 content as reference).



```
from rouge import Rouge
hyp="ITC company recorded total income of 70936.85 crore in the previous year, and the latest quarter
net profit was 5339.69 crore.Based on the provided contexts, the net profit of ITC company for this
year is 5339.69 crore."
ref="For the financial year 2022-23, ITC Limited reported a net profit of ₹18,753.31 crores. This
figure reflects the company's performance across its various business sectors including FMCG, Hotels,
Packaging, Paperboards & Specialty Papers, Agri & IT Businesses (itcportal)."
rouge=Rouge()
scores=rouge.get_scores(hyp,ref)
scores
```

```
[{'rouge-1': {'r': 0.16216216216216217,
   'p': 0.20689655172413793,
   'f': 0.18181817689164384},
  'rouge-2': {'r': 0.05263157894736842,
   'p': 0.05555555555555555,
   'f': 0.05405404905770682},
  'rouge-l': {'r': 0.13513513513513514,
   'p': 0.1724137931034483,
   'f': 0.15151514658861356}}]
```

Fig. 3. Query 2.Rouge score of the generated output from Flacon(GPT4.0 content as reference).

Query 1 on ITC's contribution to Amrit Kaal will use a TF-IDF vectorizer due to its specific context. Query 2, asking about ITC's 2022-23 net profit, will utilize BERT for its broader applicability and generalized content, which actually is the new data to the model hence handles differently from all other queries related to the text.

VIII. CODE

https://github.com/om21481/RAG_Financial_GPT

REFERENCES

[1] Yepes, A. J., You, Y., Milczek, J., Laverde, S., & Li, R. (2024). "Financial Report Chunking for Effective Retrieval Augmented Generation," arXiv preprint arXiv:2402.05131.

[2] Setty, S., Jijo, K., Chung, E., & Vidra, N. (2024). "Improving Retrieval for RAG based Question Answering Models on Financial Documents," arXiv preprint arXiv:2404.07221.

[3] Gupta, U. (2023). "GPT-InvestAR: Enhancing Stock Investment Strategies through Annual Report Analysis with Large Language Models," ArXiv.