Proceedings of the Third International Conference on Automation, Computing and Renewable Systems (ICACRS-2024)
IEEE Xplore Part Number: CFP24CB5-ART; ISBN: 979-8-3315-3242-0

2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS) | 979-8-3315-3242-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICACRS62842.2024.10841711

Enhancing Financial Risk Analysis using RAG-based Large Language Models

Abhishek Darji
Department of Computer Science & Engineering
Devang Patel Institute of Advance Technology and Research (DEPSTAR)
Charotar University of Science and Technology (CHARUSAT)
Anand, Gujarat, India
abhishekdarji653@gmail.com

Fenil Kheni
Department of Computer Science & Engineering
Devang Patel Institute of Advance Technology and Research (DEPSTAR)
Charotar University of Science and Technology (CHARUSAT)
Anand, Gujarat, India
fenilkheni1307@gmail.com

Dhruvil Chodvadia
Department of Computer Science & Engineering
Devang Patel Institute of Advance Technology and Research (DEPSTAR)
Charotar University of Science and Technology (CHARUSAT)
Anand, Gujarat, India
dhruvilchodvadia@gmail.com

Parth Goel
Department of Computer Science & Engineering
Devang Patel Institute of Advance Technology and Research (DEPSTAR)
Charotar University of Science and Technology (CHARUSAT)
Anand, Gujarat, India
parthgoel.ce@charusat.ac.in

Dweepna Garg
Department of Computer Engineering
Devang Patel Institute of Advance Technology and Research (DEPSTAR)
Charotar University of Science and Technology (CHARUSAT)
Anand, Gujarat, India
dweepnagarg.ce@charusat.ac.in

Bankim Patel
Department of Computer Science & Engineering
Devang Patel Institute of Advance Technology and Research (DEPSTAR)
Charotar University of Science and Technology (CHARUSAT)
Anand, Gujarat, India
bankimpatel.dcs@charusat.ac.in

***Abstract*—The rapid development of Generative AI has brought major changes in way of functioning of different sectors throughout the world. Many research work has been done in the field of financial sector to increase the efficiency and reduce the errors due to human intervention. However, the current financial risk analysis relies on manual reviews and conventional machine learning models which repeatedly failing to process financial risk data. This study investigates how Retrieval-Augmented Generation (RAG) approach can help Large Language Models (LLM) to generate risk analysis reports for audit reports which extract detailed information from the audit reports and avoid overlooking of small details, which was a major drawback in the earlier system. This research study covers how Retrieval Augmented Generation (RAG) enhances the performance financial risk analysis of audit reports using different LLMs like GPT-4o, Gemini-1.5-flash, and LlaMa3.1. This research work includes the performance of LLMs beyond multiple metrics, including faithfulness, context precision-recall-relevancy, and answer relevance. The research findings imply that LlaMa3.1 is a great model in terms of faithfulness of the generated report with a score of 78.26%. In terms of retrieval of the documents and its context, Llama had a very strong performance by getting the score of 79.62% in context-precision, 78.26% in context-recall and 86.99% in context-relevancy. In terms of generated report, the Llama3.1 model have the score of 37.83% for answer-relevancy and Gemini-1.5-flash have a score of 58.64% for answer-correctness.**

***Keywords*—Large language Model (LLM), Retrieval- Augmented Generation (RAG), GPT-4o, Gemini-1.5-flash, Ollama, Generative AI, Financial Risk Analysis**

I. INTRODUCTION

Financial risk assessment is very important for evaluating a company's financial performance, especially in current-day scenario where the world produces ample amount of data on daily basis. In the past financial data analysis mostly depended on manual reviews and older machine learning models [1]. But now, with big improvements in the field of Generative AI, Large Language Models have changed the times how we process and analyze financial data. By using retrieval augmented generation (RAG) methods, they help improve risk assessment in the finance sector [2]. Even though LLMs have shown great success in many natural language tasks, using RAG and other techniques is necessary to effectively apply these models in finance. This approach makes LLMs better by pulling extra context from external datasets during the inference phase [3]. This research proposes a more efficient and optimized method in comparison to the existing methods. By integrating Retrieval-Augmented Generation approach with Large Language Models (LLM), risk analysis reports would be generated directly by giving audit reports as the input. Our approach tries to eliminate the avoidance of small details from the audit reports that can be crucial for analyzing the risk. This approach not only helps to improve the faithfulness and context-awareness of the generated risk analysis report of the given audit report by using comprehensive data processing but

helps to get a customized report along with high scalability.

While conducting financial risk analysis, multiple data sources like balance sheets, income statements, historical pricing and macroeconomic indicators or directors are considered. This effort is helped by RAG using a pool of financial data from multiple sources, giving quicker and more thorough access to information. In some cases, traditional risk analysis calls for manual or batch processing of financial data such as profit/loss reports and balance sheets [4]. That method is time-consuming and might miss the important point. Retrieval Augmented LLM makes it possible by looking in internal knowledge for contextual information, that can be used. Visual representation of this framework is depicted in Figure 1.

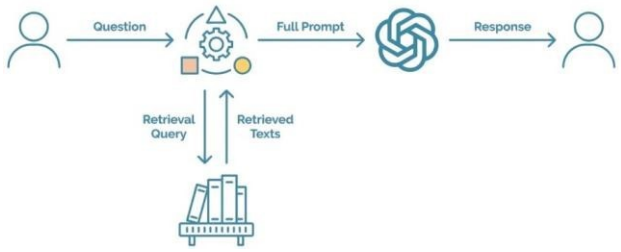

Fig. 1. A Depiction of the RAG Framework in Large Language Models

RAG combines Large Language Models (LLMs) Ope- nAI's GPT-4o or Meta's LlaMA and adds external retrieval mechanisms for real-time information streams. Since LLMs have read a lot of text, they are excellent at understanding and producing human-like responses; however, the nature of their training data is frozen in time [6]. RAG solves this issue by letting LLMs access basic information externally, as they are generating the answer making sure that the conclusions of a generated answers based on external info is valid to relevant data.

II. Literature Review

Large Language Models (LLMs) lies in natural language processing (NLP), which has seen significant recent advancements. Early work in chatbots as question-answer systems utilized NLP techniques to mimic human dialogue [6]. More recently, LLMs with Retrieval - Augmented Generation systems aim to enhance interaction accuracy and mitigate issues such as hallucinations and factual inconsistencies in multiple question-answer systems including finance domain. This section consists of remarkable contributions and developments in FinTech domain using Large Language Models. Initiating the foundation of the work, Islam et al. built the dataset comprising 10,231 questions, Answers and evidence triplets derived from 40 publicly traded companies, and 361 public filings on FinanceBench. GPT-4-Turbo, which was given additional information achieved the worst results and only gave 9% accuracy. The Oracle model was only 85% successful [7]. Udit Gupta proposed LLMs, as used in his work, to determine the trend of stock prices based on the companies' news headings. By applying such an approach in conjunction with the current machine learning models, one can promote the efficiency of an investment plan by using other information from annual reports, which is more qualitative. However, some limitations are there, like only one predictor variable is used, and the model considers that transaction costs are present. However, as mentioned earlier, there are some fallacies or disadvantages to using LLMs, and yet LLMs have some benefits, like delivering a systematic and formalized approach to stock selection [8]. Hongyang (Bruce) Yang et al. introduced FinGPT, a new instrument for financial data analysis, which has already begun to change the financial spheres by utilizing the LLM for preserving the simulation model of prediction and Money State graphical statistics. Although BloombergGPT is one of the transformation models implemented within Bloomberg and has increased the need for open-source methods among the data, it also primarily resembles and focuses on data-driven strategies and auto-data curation workflow [9]. The innovative framework defined by Boyu Zhang et al. posited that financial analysis helps in valuation and investment decisions. Parameterization in traditional models is constrained by size and training sets. However, there are issues with using LLMs for financial sentiment analysis, which involves issues such as shifts in objectives when pre-training and when predicting sentiment labels. This is where a retrieval-augmented LLM framework helps, enabling an approximate 15% to 48% improvement in accuracy and F1 score on large language models [10]. Qianqian Xie et al. and Xiao Zhang et al. proposed that ChatGPT and LlaMA are two financial learning models that have better results on different data sets as compared to other models. Their LLMs surpass other LLMs in various financial NLP tasks; FinMA-30B achieves a 10% improvement in F1 score on the FPB dataset compared with GPT-4, and it has been 37% better than BloombergGPT in the NER dataset. Overall, FinMA-7B-full performs poorly on the two evaluation measures, particularly in the financial prediction tasks. The existing conditions provide chances to advance LLMs in both financial academic research and practical usage [11]. Huaqin Zhao et al. evaluated the financial aspects. LLMs were given financial-related tests that showcased their features in zero- shot learning, mathematical problem-solving, and language sentiment analysis. In an attempt to assess these LLMs, recommendations given by the models were checked with real-life statistics and the overall performance of the market. In cases where concrete datasets were not available, GPT-4 was employed, which gave practical outputs regarding financial engineering, risk evaluation, and market indicators [12]. Doo-Il Kwak et al. proposed automation of report generation by 40% and 25% with the help of Rivet and RAG technologies in the case of data management. Empirical validations involve the measurements of precision, recall, and BLEU, which set basic

barometers for the industry to show the possibilities of such an application of advanced technologies [6]. Brian Gardiner et al. proposed the RAG, a tactical instrument to prevent risk management that increases security since it avails resources to guard investment in security measures [13]. Yixuan Tang et al. and Yi Yang et al. compare the response accuracy of the LLMs such as GPT-4, PaLM, and Llama2-70B to determine the LLM's performance in the multi-hop queries. It is evidenced that prior RAG methods are decently ineffective, with a measure of 0.56 accuracy. These open-source LLMs, Llama02-70B and Mixtral-8 x 7B, yield only 0.32 and 0.36 accuracy, respectively [14]. Kunlun Zhu1 et al. introduce the RAGEval approach produces special datasets for the evaluation of RAG systems. The comparative analysis of the results that were received at the use of the best open-source models showed that the highest rate of work was received by Qwen1. 5-14B-chat in Chinese with a completeness score of 0.4926. Llama3-8B-Instruct in English, which earned it a score of 0.6524. GPT-4o occupies the current position of the best results differs only by a small margin from the top-rated models and achieved a completeness score of 0. 5187 in Chinese and 0. 6845 in English, which means that through subsequent developments of open-source models, the performance disadvantage can be eliminated [15]. Retrieval-Augmented generation to enhance the factual correctness of large language models, in turn mini-mizing hallucination, with domain-specific contexts. The work demonstrates that on factual questions, LLMs exhibit significant improvement when they are used with private knowledge bases for retrieval [16]. Kunal Sawarkar et al. introduce a blended RAG approach stitching together semantic search along with hybrid query-based retrievers directed to improve the accuracy of the RAG models. This study fetches stronger proofs regarding retrieval precision and response quality over domains, particularly over knowledge- intensive domains, as opposed to the traditional RAG model [17]. Paulo Finardi et al. discuss the contribution that retrievers and chunked data may contribute to the overall effectiveness of an RAG system by the well-optimized retrieval mechanism and by an efficient chunking strategy toward improving text generation tasks [18]. Alireza Salemi et al. introduce methods for evaluating the quality of retrieval components within RAG frameworks. Metrics used here play a crucial role in determining the final performance of the overall system. Some improvements are proposed in the evaluation of retrieval relevance and accuracy [19]. Shahul Es et al. bring in RAGAs-the framework for the automation of RAG system evaluation. It addresses a typical set of challenges involved in the evaluation of RAGs, including subjective assessment of the output relevance and is scalable for improving the robustness of RAG-based mod- els [20]. Xiao Yang et al. proposed CRAG benchmark that targets challenging, real-world question-answering tasks from domain, such as Finance to challenge Retrieval-Augmented Generation (RAG) systems. It analyzes the performance of existing LLM and RAG systems, demonstrating their difficulty with dynamism and long-tail facts while showing areas for future investigation in trustworthy question answering [21].

Karthik Meduri et al. introduce optimizing the performance of RAG models using techniques such as Knowledge Distillation, Quantization, and Pruning. The approach aims at achieving better efficiency in handling large knowledge bases without either compromising retrieval accuracy or weakening response quality. It makes RAG promisingly possible as a solution to a host of NLP tasks such as conversational AI or content summarization [22]. Vasileios Katranidis et al. bring forth a novel approach to verifying facts, which might be applied in RAG systems under the ability of LLMs to call functions. This will help minimize the errors and costs associated with unsupported facts, especially in incompleteness or inaccuracies of texts generated. The framework also offers the potential for a more realistic investigation of factual accuracy in retrieval-augmented generation [23].

III. Methodology

This study holds the retrieval-augmented generation (RAG) architecture system, a highly developed methodology that combines the capabilities of information retrieval and natural language generation to generate complete documents. RAG architecture surpasses in two core areas: gathering information from huge data stores and using audit reports of company to create contextually relevant risk analysis reports. Figure 2 shows an overall diagram of the Retrieval-Augmented Generation (RAG) process. It provides detailed information of each step of RAG system which includes indexing and augmentation.

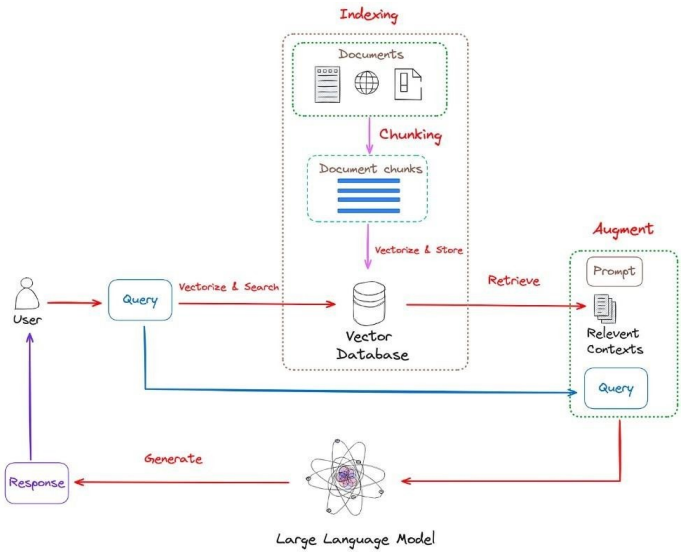

Fig. 2. A depiction of the RAG Framework along with Indexing

A. Retriever Component

The retriever's task is to obtain relevant information from the knowledge base to help in answering a user's query. The retriever uses similarity search to search through a huge vector - embedded knowledge base when a user poses a question. The most relevant vectors are identified and retrieved, and these are ultimately used to deliver a sufficient answer.

1) Document Structuring: The first stage in a RAG system is to prepare documents for efficient retrieval and analysis. In this preprocessing step, the knowledge base is arranged to achieve accurate and efficient information extraction. To maintain consistency across multiple document types, content is gathered from a variety of information sources, including documents and user-provided PDFs transformed into a single plain text format. Audit reports are combined with their associated risk analysis statements into a single document. The merged document is divided into smaller sections called chunks, about 1500 characters long, using a text splitter. To ensure semantic alignment, natural language processing techniques are employed to split the text into sentences or paragraphs. This chunking method for long documents enhances retrieval efficiency by allowing the system to concentrate on contextually relevant segments.

2) Embedding Encoding: Each text chunk needs to be converted into a dense vector embedding after going through the document structuring. By capturing the semantic meaning of the data, these embeddings enable the retrieval of data more accurately and effectively. The visual representation of embedding has been shown in Figure 3. Three different embedding models are utilized in this study's embedding delivery process: for GPT-4, text-embedding-ada-002 is employed [24]; for LLaMA3.1, nomic-text-embedding is used [25]; and for Gemini-1.5-flash, embedding-001 is used [26].

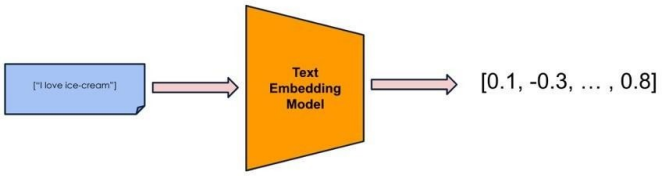

Fig. 3. Embedding of chunks

3) Vector-Based Indexing: The indexing procedure arranges vector embeddings within the vector database and improves data searchability. These embeddings are effectively stored and retrieved using Chroma. Each audit report's unique ID is used to index documents, making it possible to collect both audit reports and related risk analysis statements with the same unique ID. This approach provides simple and quick access to important information.

4) Vectorizing the Query: The user-provided audit report undergoes the same embedding procedure as other documents, ensuring consistency in data representation. The audit report embedding and the system's document embedding work effectively whenever the same embedding approach is applied. This consistency ensures accurate comparison and the retrieval of relevant information from both sources throughout the search and retrieval process. Furthermore, maintaining inter-embedding flexibility improves the retrieval mechanism's overall effectiveness by enhancing relevance and speed.

5) Retrieval Phase: A similarity search technique known as vector similarity search is applied to find a connection between the audit reports and document vectors. The most appropriate chunk of documents is pulled based on how closely they match the query. A retrieval function in the Chroma database combines many search-relevant chunks based on the similarity function by taking different search parameters.

B. Generator Component

The retrieved documents are fed into the generative models along with the user's audit report into a generative model. The language model takes this input and comprehends the potential risk in the user's audit report based on the retrieved documents' risk analysis statement. It then generates the risk analysis statement for the user's audit report, which incorporates the factual information from the audit report and has contextual relevance with the retrieved documents, thus giving an accurate response to the user's audit report. During the implementation of the RAG system, research utilized several parameters to optimize its retrieval and generation processes. Variations in values of parameters lead to difference in the results in evaluation of metrics. The parameters specified in Table 1 play a crucial role in determining the accuracy, efficiency, and overall performance of the RAG system. Every parameter stated in Table 1 have been common throughout all the models.

IV. Results & Discussion

1) Dataset Description: The dataset utilized in this research paper serves as the foundation for evaluating the efficiency and performance of Retrieval Augmented Generation (RAG) system proposed. Dataset is created by manually collecting the audit reports of different firms and creating their risk analysis statements. The dataset is created carefully throughout the process and the test-cases have risk analysis statement, which are created in a similar way for evaluation purpose. The dataset comprises of audit reports of different firms of different quarters of fiscal year 2023-24 and its risk analysis statement. The quarterly reports of Aditya Birla Capital Limited [27], Bajaj Finance Limited [28], Cipla Limited [29], HDFC Bank Limited [30], Hindustan Unilever Limited [31], ICICI Bank Limited [32], LTIMindtree Limited [33], State Bank of India [34], Tech Mahindra Limited [35]. The test- case are similar audit reports and their risk analysis statement are used as the ground report so it can be compared with the generated report. There are four test-cases of which are audit reports of Tata Consultancy Services [36] for all the four quarters of fiscal year 2023-24. The Metadata of the dataset is given in Table II.

2) Evaluation of RAG system: Evaluation of any study or research is the last and most important step that helps one to analyze how much useful the study is. Evaluation not only helps to understand the existing models but also to open new possibilities of further research. RAG systems are mainly evaluated on the basis of retrieval of the context and generation of the output.

TABLE I
PARAMETERS OF MODEL

Parameter	Description	Value
Document Chunk Size	The text chunks generated during the pre-processing stage, typically measured in to-kens, must strike a balance between coherence and manageability. Proper chunk size ensures that the segments are large enough to retain contextual meaning.	1500 characters
Chunk Overlap	To maintain the context between the chunks, few characters are overlapped in each chunk.	100 characters
Retriever Search K	Top relevant audit report and its risk analysis statement is retrieved which is used for generating the risk analysis statement.	1
QA Prompt template	Prompt engineering is an important aspect in any LLM application. It is the structure which is used as a template by the language model for generating the output.	” The user has given quarterly financial results of a company. Learn the data provided in the file” + user_audit_report +” First tell me the name of the company specified in the user given quarterly financial report.” +” The relevant risk analysis reports are:” + combined_risk_analyses +” the quarterly financial results given by the user. Create a risk analysis report of the user’s quarterly financial results using the relevant risk analysis documents by using its terminologies, format, semantics, writing style etc...”
Embedding Model	The model used to generate dense vector embedding from text chunks.	Text-embedding-ada-002-v2, embedding-001, nomic-embed-text
Primary QA LLM	The language model used for generating answers based on the retrieved document chunks and the user query.	GPT 4o, Llama3.1, Gemini-1.5-flash
Vector Database	The type of database used to index and store vector embeddings for efficient retrieval.	Chroma Database

TABLE II
METADATA OF DATASET

Metadata	Description
Audit Reports	The audit reports are quarterly results that each firm releases after the end of every term. The audit reports are used to train the LLM and risk analysis statements are generated on the basis of the audit reports. Audit reports also serve as the test cases which will be used during RAG implementation.
Risk Analysis Statements	Risk analysis statements are documents created manually or by any tool for assessing the risks in any given audit report. It is based on a few parameters that decide the level of risks in the report of the firm. The risk analysis statement also acts as the ground truth for the test cases which can then be used during the evaluation of the generated risk analysis statement.

In this research, different evaluation metrics have been used which are commonly used in LLM’s and NLP. The different evaluation metrics are Faithfulness, Context Precision, Context Recall, Answer Relevancy, Context Relevancy, Answer Correctness. On the basis of risk analysis statement generated by each model, the Table III consists of the results of evaluated model. The comparison of different models on the basis of each metric have been shown in Figure 4. The study clearly states that Llama 3.1 model have overall better performance as compared to the OpenAI’s GPT-4o and Gemini-1.5-flash.

TABLE III
EVALUATION METRICS FOR DIFFERENT MODELS

Evaluation Metric	GPT-4o	Gemini-1.5-flash	LLama3.1
Faithfulness	70.17%	62.45%	78.26%
Context-precision	56.86%	46.16%	79.62%
Context-recall	70.17%	62.45%	78.26%
Answer-relevancy	34.92%	32.43%	37.83%
Context-relevancy	72.63%	66.45%	86.99%
Answer-correctness	57.93%	58.64%	48.47%

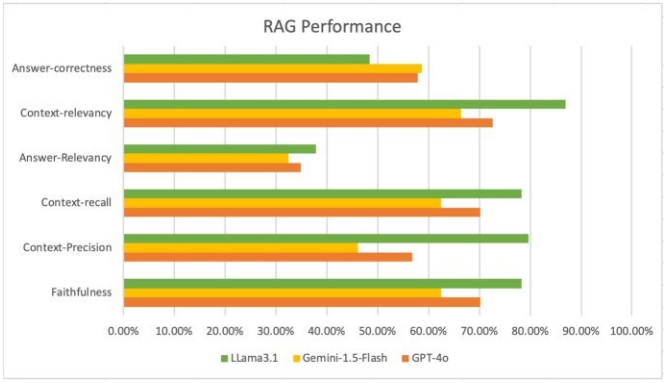

Fig. 4. Comparison of Evaluation Metrics for Different Models

V. CONCLUSION

This research study gives some further improvement in the field of financial analysis using artificial intelligence and machine learning, The RAG model generates the desired output on the basis retrieved documents. The results concluded from the study is that model generated reports were compared to the actual risk analysis reports which gave different results for different models. The Llama3.1 model turned out to be a very efficient model due to its higher accuracy and precision in terms of financial analysis and Natural Language Processing. Large Language Models have proved to be a great tool for financial analysis. Despite the fact that current study has given good results, deeper research is required for future enhancements in the field of financial analysis and LLMs. Future work might involve the integration of advanced retrieval mechanisms and further fine-tuning these models toward specific financial tasks, permitting them to have greater precision. Perhaps the area of study financial tasks can be improved by working with different models and trying to integrate the financial concepts with them. The understanding of financial study will make the implementation of the models and algorithms much easier and give more accurate and precise results.

REFERENCES

[1] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.

[2] Y. Cao, Z. Chen, Q. Pei, F. Dimino, L. Ausiello, P. Kumar, K. Subbalakshmi, and P. M. Ndiaye, "Risklabs: Predicting financial risk using large language model based on multi-sources data," *arXiv preprint arXiv:2404.07452*, 2024.

[3] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "Bloomberggpt: A large language model for finance," *arXiv preprint arXiv:2303.17564*, 2023.

[4] B. H. A. Khattak, I. Shafi, A. S. Khan, E. S. Flores, R. G. Lara, M. A. Samad, and I. Ashraf, "A systematic survey of AI models in financial market forecasting for profitability analysis," *IEEE Access*, 2023.

[5] D.-I. Kwak and K.-Y. Park, "Enhancing automated report generation: Integrating RIVET and RAG with advanced retrieval techniques," in Proceedings of the Korea Information Processing Society Conference, Korea Information Processing Society, 2024, pp. 753–756.

[6] P. Goel and A. Ganatra, "A survey on chatbot: Futuristic conversational agent for user interaction," in 2021 3rd international conference on signal processing and communication (ICPSC). IEEE, 2021, pp. 736– 740.

[7] P. Islam, A. Kannappan, D. Kiela, R. Qian, N. Scherrer, and B. Vidgen, "Financebench: A new benchmark for financial question answering," *arXiv preprint arXiv:2311.11944*, 2023.

[8] U. Gupta, "GPT-Investar: Enhancing stock investment strategies through annual report analysis with large language models," *arXiv preprint arXiv:2309.03079*, 2023.

[9] H. Yang, X.-Y. Liu, and C. D. Wang, "FinGPT: Open-source financial large language models," *arXiv preprint arXiv:2306.06031*, 2023.

[10] B. Zhang, H. Yang, T. Zhou, M. Ali Babar, and X.-Y. Liu, "Enhancing financial sentiment analysis via retrieval-augmented large language models," in *Proceedings of the Fourth ACM International Conference on AI in Finance*, 2023, pp. 349–356.

[11] Q. Xie, W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira, and J. Huang, "Pixiu: A large language model, instruction data and evaluation benchmark for finance," *arXiv preprint arXiv:2306.05443*, 2023.

[12] H. Zhao, Z. Liu, Z. Wu, Y. Li, T. Yang, P. Shu, S. Xu, H. Dai, L. Zhao, G. Mai, et al., "Revolutionizing finance with LLMs: An overview of applications and insights," *arXiv preprint arXiv:2401.11641*, 2024.

[13] B. Gardiner, "E-business security in RAG order," *School of Computing, Dublin Institute of Technology*, Ireland, 2003.

[14] Y. Tang and Y. Yang, "Multihop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries," *arXiv preprint arXiv:2401.15391*, 2024.

[15] K. Zhu, Y. Luo, D. Xu, R. Wang, S. Yu, S. Wang, Y. Yan, Z. Liu, X. Han, Z. Liu, et al., "RAGEVAL: Scenario specific RAG evaluation dataset generation framework," *arXiv preprint arXiv:2408.01262*, 2024.

[16] J. Li, Y. Yuan, and Z. Zhang, "Enhancing LLM factual accuracy with RAG to counter hallucinations: A case study on domain-specific queries in private knowledge-bases," *arXiv preprint arXiv:2403.10446*, 2024.

[17] K. Sawarkar, A. Mangal, and S. R. Solanki, "Blended RAG: Improving RAG (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers," *arXiv preprint arXiv:2404.07220*, 2024.

[18] P. Finardi, L. Avila, R. Castaldoni, P. Gengo, C. Larcher, M. Piau, P. Costa, and V. Carida´, "The chronicles of RAG: The retriever, the chunk and the generator," *arXiv preprint arXiv:2401.07883*, 2024.

[19] A. Salemi and H. Zamani, "Evaluating retrieval quality in retrieval-augmented generation," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 2395–2400.

[20] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval-augmented generation," *arXiv preprint arXiv:2309.15217*, 2023.

[21] X. Yang, K. Sun, H. Xin, Y. Sun, N. Bhalla, X. Chen, S. Choudhary, R. D. Gui, Z. W. Jiang, Z. Jiang, et al., "CRAG–Comprehensive RAG benchmark," *arXiv preprint arXiv:2406.04744*, 2024.

[22] K. Meduri, G. S. Nadella, H. Gonaygunta, M. H. Maturi, and F. Fatima, "Efficient RAG framework for large-scale knowledge bases," 2024.

[23] V. Katranidis and G. Barany, "FAAF: Facts as a function for the evaluation of RAG systems," *arXiv preprint arXiv:2403.03888*, 2024.

[24] OpenAI, "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774v6*, 2023.

[25] Ollama, "LLaMA 3.1," Ollama, 2024. [Online]. Available: https://ollama.com

[26] Gemini Team, Google, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530v4*, 2024.

[27] Aditya Birla Capital, "Quarterly Results," 2024. [Online]. Available: https://www.adityabirlacapital.com/investor-relations/quarterly-results

[28] Bajaj Finserv, "Financial Results," 2024. [Online]. Available: https://www.aboutbajajfinserv.com/finance-investor-relations-financial-results

[29] Cipla, "Quarterly Results," 2024. [Online]. Available: https://www.cipla.com/investors/quarterly-results

[30] HDFC Bank, "Financial Results," 2024. [Online]. Available: https://www.hdfcbank.com/personal/about-us/investor-relations/financial-results

[31] Hindustan Unilever Limited, "Quarterly Results," 2024. [Online]. Available: https://www.hul.co.in/investor-relations/results-presentations/quarterly-results/

[32] ICICI Bank, "Quarterly Financial Results," 2024. [Online]. Available: https://www.icicibank.com/about-us/qfr
[33] LTIMindtree, "Financial Results," 2024. [Online]. Available: https://www.ltimindtree.com/investors/financial-results/
[34] State Bank of India, "Investor Relations Reports," 2024. [Online]. Available: https://sbi.co.in/web/investor-relations/reports
[35] Tech Mahindra, "Quarterly Earnings," 2024. [Online]. Available: https://www.techmahindra.com/investors/quarterly-earnings/
[36] Tata Consultancy Services, "Financial Statements," 2024. [Online]. Available: https://www.tcs.com/investor-relations/financial-statements