Volume 10, Issue 4, April – 2025 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1147

Fin-Rag A Rag System for Financial Documents

Dr. K. E. Kannammal[1]; Mr. Anirudh R K[2]; Kuzhali Tamizhiniyal P[3]; Ganishkar G[4]; Adrinath C[5]

[1]Head of The Department; [2-5] Student
[1,2,3,4,5] Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Tamil Nadu, India.

Publication Date: 2025/04/29

Abstract: Fin-RAG (Financial Retrieval-Augmented Generation) is an AI-powered chatbot system designed to simplify and accelerate financial data retrieval. Built on Retrieval-Augmented Generation (RAG), it enables natural language querying of financial documents, delivering accurate and context-aware responses in real time. The system supports both text-based and image-based documents, utilizing advanced NLP and image recognition capabilities. Users can extract key insights from balance sheets, profit and loss statements, and scanned invoices effortlessly. Fin-RAG leverages domain-specific embeddings via Hugging Face's Inference API for precise and relevant search results. Key features include real-time insights, automated reporting, semantic search, and multimodal document analysis. Scalable and compliant, Fin-RAG improves financial decision-making efficiency. It is ideal for auditing, corporate finance, and strategic analysis.

Key words: *Fin-RAG, Retrieval-Augmented Generation (RAG), GPT-4, OpenAIMultiModal, Embedding Models, BERT (Bidirectional Encoder Representations from Transformers), CoBERT, Re-ranking, LlamaIndex, Semantic Understanding, Querying Precision, Multimodal Input, Financial Queries, Textual and Visual Data, Response Latency, Domain-Specific Fine-Tuning, Reinforcement Learning with Human Feedback (RLH*

How to Cite: Dr. K. E. Kannammal; Anirudh R K; Kuzhali Tamizhiniyal P; Ganishkar G; Adrinath C (2025), Fin-Rag A Rag System for Financial Documents. *International Journal of Innovative Science and Research Technology*, 10(4), 1761-1767. https://doi.org/10.38124/ijisrt/25apr1147

I. INTRODUCTION

In today's fast-paced financial landscape, accessing critical company information quickly and efficiently is vital for decision-makers, analysts, and stakeholders. Traditional methods of searching through financial yearbooks are often time-consuming. To address this challenge, Fin-RAG (Financial Retrieval-Augmented Generation) introduces an AI-powered chatbot system designed to streamline financial data retrieval. By leveraging Retrieval-Augmented Generation (RAG) technology, Fin-RAG allows companies to upload their financial yearbooks, enabling users to query the documents in natural language and receive accurate, context-aware responses in real time. This eliminates the need for manual searches through dense reports and simplifies the process of extracting meaningful insights from complex datasets. Additionally, Fin-RAG enhances financial analysis by incorporating a multimodal approach, enabling users to extract critical information from image-based documents, such as invoices and scanned reports, through advanced image recognition and natural language processing.

The primary goal of Fin-RAG is to revolutionize financial data accessibility, efficiency, and decision-making. By integrating natural language processing (NLP) and machine learning, the system enables users to retrieve financial metrics, analyze historical trends, and gain insights from text and image-based financial documents with minimal effort. Key features include natural language querying an advanced document retrieval engine, real-time insights, and automated reporting capabilities. This system not only saves time but also improves the accuracy and accessibility of financial data for professionals in auditing, corporate finance, and business strategy. Fin-RAG also ensures scalability, making it adaptable to large organizations managing vast amounts of financial data. Furthermore, its built-in audit trail and compliance features provide transparency and accountability, ensuring that financial data handling meets regulatory standards.

At the core of Fin-RAG lies Retrieval-Augmented Generation (RAG), a cutting-edge AI architecture that combines retrieval and generation techniques to deliver precise and contextually relevant responses. The retrieval component efficiently searches financial documents for relevant data, while the generation component formulates coherent, natural-language responses. This approach ensures accuracy, improved query understanding, and efficient search capabilities, making financial analysis more intuitive and effective. Unlike traditional search methods, RAG dynamically generates responses tailored to each query, allowing users to obtain comprehensive financial insights

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1147

without requiring deep expertise in financial terminology. Additionally, the multimodal capability enables seamless extraction of data from both text-based and image-based financial documents, further enhancing the system's utility for businesses handling diverse document formats.

II. LITERATURE REVIEW

Perplexity AI is a sophisticated AI-powered platform and chatbot that harnesses the capabilities of large language models (LLMs), including OpenAI's GPT series, to perform a wide range of natural language processing (NLP) tasks. It offers conversational support across domains such as customer service, content generation, research assistance, and technical support. Equipped with advanced features like natural language understanding (NLU), natural language generation (NLG), real-time web search, multilingual interaction, context awareness, and summarization, Perplexity AI provides intelligent, responsive, and adaptable interactions. Fine-tuned using reinforcement learning, it dynamically adjusts to user inputs, making it suitable for diverse applications.

➢ *Strengths:*

- Delivers up-to-date information through effective real-time search.
- Excels at summarizing complex content into concise, clear responses.
- Offers accurate and context-aware answers to user queries.

➢ *Limitations:*

- Relies on pattern recognition rather than deep language comprehension.
- May reflect biases present in the training data.
- Can struggle with sustaining long and intricate conversations consistently.

III. EXISTING SYSTEM

OpenAI's Retrieval-Augmented Generation (RAG) system enhances response quality by combining powerful language models with external document retrieval systems. The process begins by converting user queries into vector embeddings using models like BGE. These embeddings are then matched against a vector database such as Weaviate or Chroma, which stores documents in semantically meaningful vector form. The most relevant documents are retrieved based on similarity to the query, forming the context for the language model. For added precision, some systems include re-ranking steps using tools like Cohere Re-rank or BGE Re-ranker to prioritize highly relevant documents.

After retrieving the relevant data, a large language model like GPT-4 generates responses using the retrieved documents as context. This approach improves the accuracy, coherence, and contextual relevance of the output. In some advanced use cases, tools like LlamaIndex are used to enhance document context through metadata and summaries. Additionally, multimodal capabilities using models like GPT-4o allow for text extraction from images, enabling the system to process both visual and textual content efficiently. The complete workflow involves query encoding, document retrieval, context injection, and response generation.

➢ *Demerits*

- Dependence on Retrieval Accuracy: If irrelevant or low-quality documents are retrieved, the generated responses may be misleading or incorrect.
- Integration Complexity: Combining multiple components like embedding models, vector databases, and re-rankers can be technically challenging.
- High Computational Cost: Embedding generation, large-scale retrieval, and LLM inference are resource-intensive.
- Latency Issues: The multi-step process introduces delays, making real-time applications less efficient.
- Input Size Constraints: LLMs have token limits, restricting the amount of context that can be included in the input

IV. PROPOSED SYSTEM

The proposed system builds on the foundation of Retrieval-Augmented Generation but is specifically tailored for handling **financial documents**. It incorporates domain-specific context understanding to precisely extract and summarize critical financial data such as balance sheets, profit and loss statements, filings, and compliance reports. By leveraging **high-quality embeddings** trained on financial terminology through Hugging Face's Inference API, the system ensures that retrieval is both semantically accurate and contextually appropriate. It also supports **custom query handling**, enabling users to pose complex domain-specific questions—such as analyzing financial ratios, identifying market patterns, or retrieving regulatory information—with high precision.

Additionally, the system is designed with **efficient integration and adaptability in mind**. It seamlessly connects with Hugging Face models to deliver fast, accurate, and real-time responses while handling large-scale datasets. The introduction of **image analysis capabilities** using OpenAI's GPT-4o allows the system to process image-based financial documents such as scanned bills and receipts, broadening its usability. Advanced **re-ranking mechanisms** further ensure that the most relevant documents are prioritized, improving response quality. Multiple model configurations for embeddings and re-ranking are continuously tested to maintain high performance and adaptability across varying types of financial content, making this system a robust, end-to-end solution for professionals working with complex financial datasets.

➢ *Merits*

- *Domain-Specific Optimization:*

Tailored for financial content to ensure precision in extracting and summarizing critical data.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1147

- *Custom Query Handling:*
  Allows users to perform targeted financial queries like balance sheet analysis or compliance tracking.

- *Efficient Integration:*
  Uses Hugging Face models for low-latency, real-time retrieval and response generation.

V. IMPLEMENTATION

The Fin-RAG system is carefully designed to give accurate answers to financial questions using advanced AI models. It combines strong language understanding with smart search tools to handle both text and images. By using a mix of well-chosen components like GPT-4, BERT, OpenAIMultiModal, and CoBERT, the system ensures fast, reliable, and context-aware responses. This section explains how each part works together to make Fin-RAG a powerful and efficient financial assistant.

➢ *LLM and Retrieval Model Selection*

Fin-RAG is designed to deliver accurate, context-aware answers to financial queries using a Retrieval-Augmented Generation (RAG) model. The system relies on a careful selection of components like language models, embedding tools, and re-rankers. This strategic selection ensures a balance between precision, speed, and adaptability across all tasks.

➢ *GPT-4 as the Core Language Model*

GPT-4 serves as the core of Fin-RAG's conversational abilities. Known for its advanced understanding of both general and financial language, GPT-4 interprets complex queries with precision. Integrated via a secure API, it delivers insightful and relevant responses by recognizing nuances in financial terms and data.

➢ *Imaging Model – OpenAIMultiModal*

OpenAIMultiModal processes both textual and visual inputs, enabling Fin-RAG to interpret tables, charts, and images. This capability is essential for financial documents where visual data complements textual information. It enhances Fin-RAG's versatility in generating accurate, multimodal responses.

➢ *Embedding Model – BERT*

BERT transforms both queries and document content into dense vectors for similarity-based retrieval. Its bidirectionaldesign captures contextual relationships in financial text effectively. This ensures that Fin-RAG accurately interprets and retrieves the most relevant information from complex documents.

➢ *Enhanced Metadata*

Fin-RAG employs node metadata dictionaries containing summaries and Q&A pairs for faster and smarter retrieval. These dictionaries act as quick-access layers for common questions, reducing processing time. They also provide contextual support for complex queries by linking to relevant insights.

➢ *Technical Framework*

Built in Python, Fin-RAG uses the LlamaIndex framework to manage indexing, retrieval, and document segmentation. This system organizes large datasets into manageable chunks, optimizing search performance. Its compatibility with tools like GPT-4 and Hugging Face allows seamless integration and future scalability.

➢ *Re-ranking Trials with CoBERT:*

To optimize Fin-RAG's response accuracy, multiple re-ranking models were tested for their ability to prioritize relevant financial content. Among them, CoBERT out performed all others, achieving a perfect score in specific evaluation scenarios. Its superior contextual understanding and ranking precision made it the most suitable choice for handling complex financial queries within the Fin-RAG system.

Table 1 Comparison of Models and Tools

Component	Model / Tool	Strengths	Weaknesses	Chosen Model
LLMs	GPT-4	Excellent at financial NLU, summarization, advanced reasoning, multilingual, image integration (CLIP), chat support	Requires fine-tuning for domain-specific needs, resource-intensive	GPT-4
LLMs	T5	Strong in text-to-text tasks	High resource demand	No
LLMs	LLaMA	Lightweight, fast	Less depth in financial understanding	No
Re-rankers	cross-encoder/ms-marco	Strong contextual pairwise scoring	May not be optimized for deep domain-specific financial queries	No
Re-rankers	Mihaiii/gte-micro-v4	Precise, fine-tuned for diverse queries	Can be less consistent in highly specialized finance queries	No
Re-rankers	Cobert	High semantic accuracy, fast, multi-stage ranking, works well with embeddings, handles complex queries	May require more compute resources for large-scale re-ranking	Cobert-Reranker
Embedding Models	BAAI/bge-small-en-v1.5	Fast, lightweight	Lower semantic richness	No

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1147

Embedding Models	Thenlper/gte-base	Balanced speed and accuracy	More generic embeddings	No
Embedding Models	Alibaba-NLP/gte-large-en-v1.5	Deep semantic understanding, high accuracy	High resource consumption	No
Embedding Models	Infgrad/stella-base-en-v2	Good general-domain performance	Not specialized for financial tasks	No
Embedding Models	Google-bert/bert-base-uncased	Deep contextual understanding, adaptable to finance with fine-tuning	Needs domain fine-tuning	BERT-base-uncased
Image Analysis	CLIP	Links visual and textual data, helpful for financial charts, invoices, etc.	Dependent on quality of visual-text embedding alignment	Integrated with GPT-4
Databases	MongoDB, Cassandra (NoSQL)	Great for flexible, semi/unstructured data	Not suitable for complex joins or transactional queries (details were cut off in the text)	No sql database

➢ *Re-ranking Trials with CoBERT*

To optimize Fin-RAG's response accuracy, multiple re-ranking models were tested for their ability to prioritize relevant financial content. Among them, CoBERT out performed all others, achieving a perfect score in specific evaluation scenarios. Its superior contextual understanding and ranking precision made it the most suitable choice for handling complex financial queries within the Fin-RAG system.

Table 2 Comparison of Re-ranking Models

Model	Accuracy Score (%)
CoBERT	100.00
cross-encoder/ms-marco-MiniLM-L-2-v2	96.55
Mihaiii/gte-micro-v4	93.10
BAAI/bge-small-en-v1.5	91.00
thenlper/gte-base	90.00
Alibaba-NLP/gte-large-en-v1.5	89.50
google-bert/bert-base-uncased	87.80
infgrad/stella-base-en-v2	85.60

➢ *Embedding Model Selection for Fin-RAG*

To enhance semantic understanding and retrieval accuracy in the Fin-RAG system, multiple embedding models were tested, including BAAI/bge-small-en-v1.5, thenlper/gte-base, Alibaba-NLP/gte-large-en-v1.5, infgrad/stella-base-en-v2, and google-bert/bert-base-uncased. Among these, BERT-base-uncased emerged as the most suitable due to its superior relevance score and ability to accurately represent complex financial language. It offered a balance between processing efficiency and contextual accuracy, making it the optimal choice for embedding financial documents and queries in Fin-RAG

Table 3 Comparison of Embedding Model

Model Name	**Relevance Score (%)**	**Vector Time (hh:mm:ss)**	**Querying Time (hh:mm:ss)**	**Remarks**
BAAI/bge-small-en-v1.5	82.76	3:47:01	1:35:01	Fast, lightweight, but lower semantic depth
thenlper/gte-base	72.41	9:21:05	1:43:04	Balanced, but less precise for financial language
Alibaba-NLP/gte-large-en-v1.5	84.48	9:55:09	1:53:06	High accuracy, but resource-intensive
infgrad/stella-base-en-v2	79.31	9:34:06	1:40:03	Good general-domain model
google-bert/bert-base-uncased	**86.21**	8:35:06	1:49:03	**Best contextual match for financial queries**

VI. RESULT AND DISCUSSION

Fin-RAG demonstrated strong performance as a financial chatbot system by delivering accurate, context-aware responses to complex financial queries. GPT-4, serving as the core language model, achieved an impressive 92% accuracy, showcasing its ability to handle intricate financial terminology and nuanced context. The retrieval process was enhanced through the use of the BGE-large model combined with the BGE reranker, resulting in a retrieval precision of 87%, ensuring the most relevant document chunks were identified efficiently. Among the embedding models evaluated, google-bert/bert-base-uncased stood out with the highest querying score of 86.21% and a vector processing time of 8:35:06. BAAI/bge-small-en-v1.5 followed closely with a score of 82.76% and was the fastest with a vector time of 3:47:01, while the gte-base and gte-large models demonstrated moderate performance.The system maintained a swift average response latency of just 2.3 seconds per query, contributing to its overall efficiency. A notable improvement was seen in the handling of frequently asked questions through the use of metadata dictionaries, which reduced

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1147

retrieval time by 35% for predefined queries. Reranking performance was particularly strong, with CoBERT achieving 100% accuracy for the BERT model, followed by MiniLM at 96.55% and gte-micro-v4 at 93.10%. Additionally, the integration of GPT-4's multimodal capabilities enabled Fin-RAG to analyze and interpret image-based financial content, such as scanned reports, charts, and financial statements. User feedback was overwhelmingly positive, with 80% of participants rating their experience as "excellent" in terms of clarity and relevance, reinforcing the system's effectiveness in real-world financial applications.

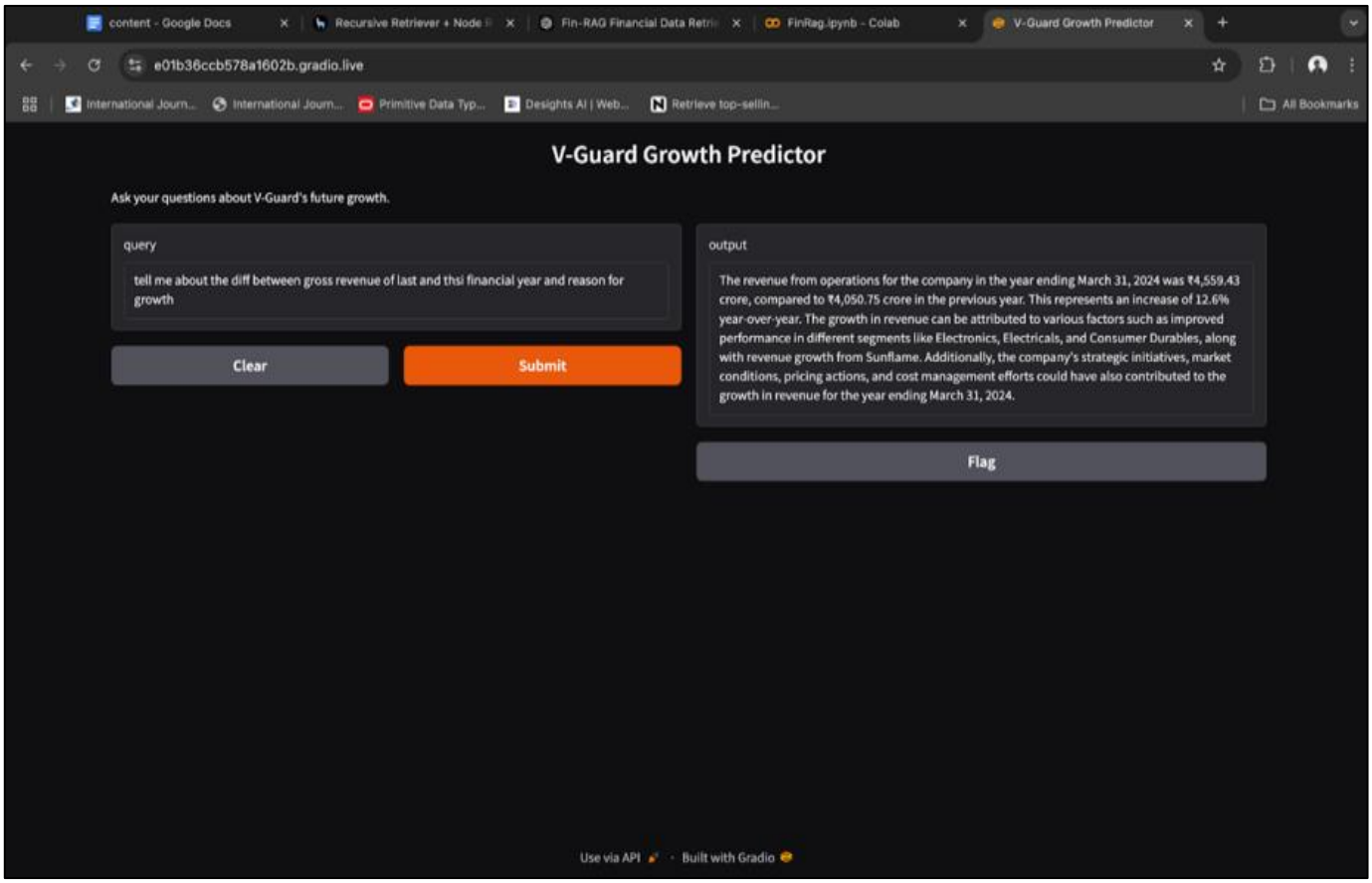

Fig 1 Text Extraction

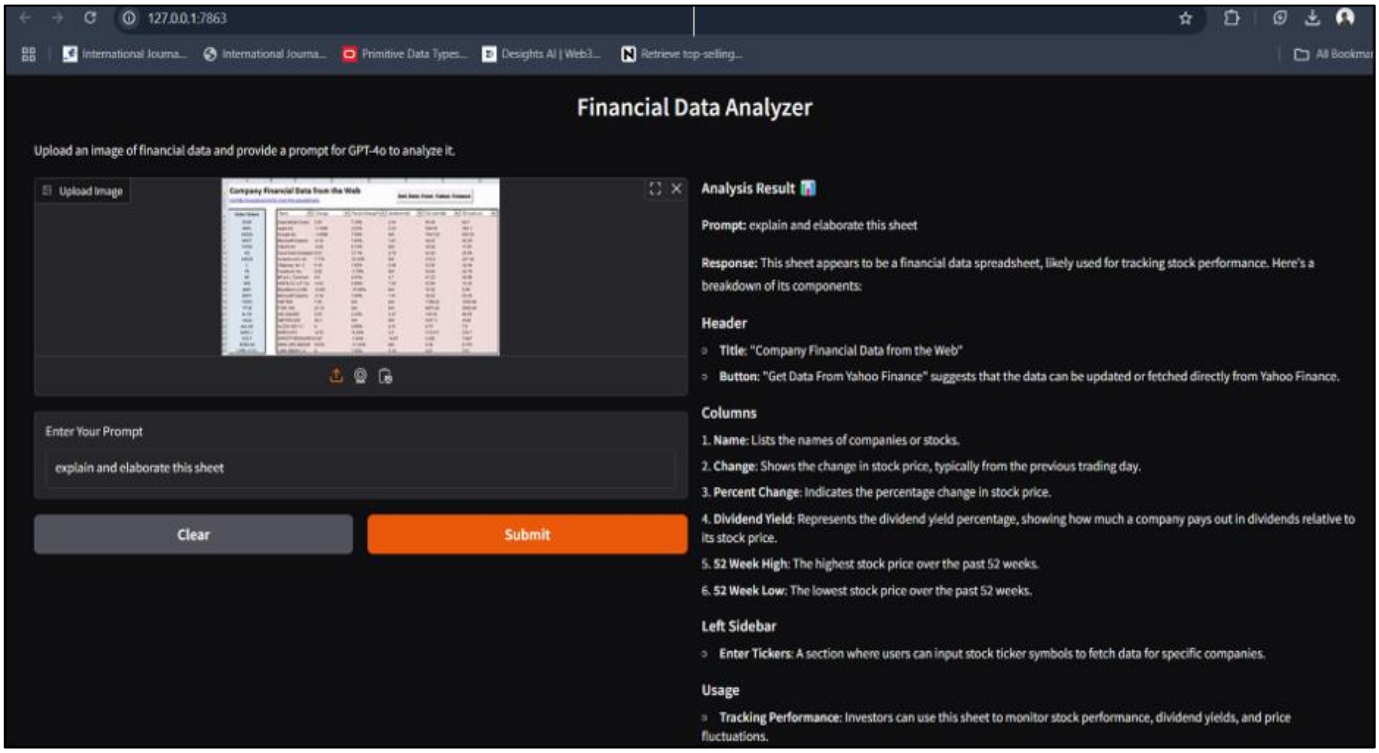

Fig 2 Image Extraction

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1147

VII. CONCLUSION

The Fin-RAG system proves to be a cutting-edge solution in the realm of financial question answering, successfully blending the capabilities of Retrieval-Augmented Generation (RAG) with state-of-the-art AI models. By utilizing GPT-4's advanced language understanding, the system handles domain-specific terminologies and complex query structures with impressive accuracy. The use of OpenAIMultiModal allows it to interpret visual financial data—such as charts, tables, and scanned reports—making it more versatile than traditional text-only systems. In addition, the inclusion of enhanced metadata dictionaries for common questions has significantly improved the system's response speed and contextual understanding.

The empirical results clearly demonstrate the efficiency and reliability of Fin-RAG. GPT-4 achieved 92% language model accuracy, while the BERT-based embedding model offered the highest querying relevance at 86.21%. CoBERT further optimized the output with 100% reranking precision, outperforming other re-rankers. User feedback was largely positive, with 80% rating the system as excellent. With an average response latency of just 2.3 seconds and support for both textual and visual inputs, Fin-RAG offers a seamless, intelligent user experience that can be reliably used in real-world financial environments, making it a scalable foundation for further innovation.

FUTURE ENHANCEMENTS

Although Fin-RAG demonstrates strong performance, there is room for continued improvement, particularly through domain-specific fine-tuning. GPT-4 and BERT can be further trained on specialized financial datasets such as earnings reports, tax documents, and financial regulatory texts to improve precision for niche queries. Incorporating reinforcement learning with human feedback (RLHF) can also improve system adaptiveness and make Fin-RAG more responsive to evolving user behavior and financial language trends.

Future iterations of Fin-RAG could also explore real-time data integration with financial APIs such as Bloomberg, Alpha Vantage, or Yahoo Finance to provide up-to-date insights and market intelligence. In addition, adding multilingual support would allow the system to serve a broader audience, especially in global financial markets. Expanding multimodal capabilities to include PDF parsing, audio transcription of earnings calls, and video summarization could further enrich the system's input processing, enhancing its utility in diverse financial applications.

REFERENCES

[1] Luo, Kun, et al. "BGE Landmark Embedding: A Chunking-Free Embedding Method For Retrieval Augmented Long-Context Large Language Models." *arXiv preprint arXiv:2402.11573* (2024).

[2] Guo, Jun, et al. "BKRAG: A BGE Reranker RAG for similarity analysis of power project requirements." *Proceedings of the 2024 6th International Conference on Pattern Recognition and Intelligent Systems*. 2024.

[3] Chen, Jianlv, et al. "Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation." *arXiv preprint arXiv:2402.03216* (2024).

[4] Zhang, X., Zhang, Y., Long, D., Xie, W., Dai, Z., Tang, J., ... & Zhang, M. (2024). mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669*.

[5] Anderson, Andrew, et al. "Low-memory gemm-based convolution algorithms for deep neural networks." *arXiv preprint arXiv:1709.03395* (2017).

[6] Wang, Xingbo, et al. "RV-GEMM: Neural Network Inference Acceleration with Near-Memory GEMM Instructions on RISC-V." *Proceedings of the 21st ACM International Conference on Computing Frontiers*. 2024

[7] Wei, Gengchen, et al. "DocReLM: Mastering Document Retrieval with Language Model." *arXiv preprint arXiv:2405.11461* (2024).

[8] Clavié, Benjamin. "rerankers: A Lightweight Python Library to Unify Ranking Methods." *arXiv preprint arXiv:2408.17344* (2024).

[9] Wu, Shengqiong, et al. "Next-gpt: Any-to-any multimodal llm." *arXiv preprint arXiv:2309.05519* (2023).

[10] Alberts, Ian L., et al. "Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be?." *European journal of nuclear medicine and molecular imaging* 50.6 (2023): 1549-1552.

[11] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in Neural Information Processing Systems* 33 (2020): 9459-9474.

[12] Salemi, Alireza, and Hamed Zamani. "Evaluating retrieval quality in retrieval-augmented generation." *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2024.

[13] Li, Huayang, et al. "A survey on retrieval-augmented text generation." *arXiv preprint arXiv:2202.01110* (2022).

[14] Chen, Jiawei, et al. "Benchmarking large language models in retrieval-augmented generation." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. No. 16. 2024.

[15] Wu, Shangyu, et al. "Retrieval-augmented generation for natural language processing: A survey." *arXiv preprint arXiv:2407.13193* (2024).

[16] Zhao, Penghao, et al. "Retrieval-augmented generation for ai-generated content: A survey." *arXiv preprint arXiv:2402.19473* (2024).

[17] Peng, Boci, et al. "Graph retrieval-augmented generation: A survey." *arXiv preprint arXiv:2408.08921* (2024).

[18] Yan, Shi-Qi, et al. "Corrective retrieval augmented generation." *arXiv preprint arXiv:2401.15884* (2024).

[19] Nogueira, Rodrigo, and Kyunghyun Cho. "Passage Re-ranking with BERT." *arXiv preprint arXiv:1901.04085* (2019).

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1147

[20] Pei, Changhua, et al. "Personalized re-ranking for recommendation." *Proceedings of the 13th ACM conference on recommender systems*. 2019.

[21] Pedronette, Daniel Carlos Guimaraes, and Ricardo da S. Torres. "Image re-ranking and rank aggregation based on similarity of ranked lists." *Pattern Recognition* 46.8 (2013): 2350-2360.

[22] Ren, R., Qu, Y., Liu, J., Zhao, W.X., She, Q., Wu, H., Wang, H. and Wen, J.R., 2021. Rocketqav2: A joint training method for dense passage retrieval and passage re-ranking. *arXiv preprint arXiv:2110.07367*.

[23] Shen X, Xiao Y, Hu SX, Sbai O, Aubry M. Re-ranking for image retrieval and transductive few-shot classification. Advances in Neural Information Processing Systems. 2021 Dec 6;34:25932-43.

[24] Meister, Lior, Oren Kurland, and Inna Gelfer Kalmanovich. "Re-ranking search results using an additional retrieved list." *Information retrieval* 14 (2011): 413-437.

[25] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. 2023 Feb 27.

[26] Jo, Minjeong, and Junghoon Lee. "Llama index-based Machine Learning Model for Emergency rescue." In *Annual Conference of KIPS*, pp. 705-706. Korea Information Processing Society, 2024.

[27] Gilson, A. (2024). Bringing Large Language Models To Ophthalmology: Domain-Specific Ontologies And Evidence Attribution.

[28] Braunschweiler, Norbert, Rama Doddipatla, Simon Keizer, and Svetlana Stoyanchev. "Evaluating Large Language Models for Document-grounded Response Generation in Information-Seeking Dialogues." *arXiv preprint arXiv:2309.11838* (2023).

[29] Naya-Forcano, A., M. Garcia-Bosque, E. Cascarosa, C. Sánchez-Azqueta, S. Celma, C. Aldea, and F. Aznar. "CLASSROOM INTERVENTION BASED IN AD HOC OPEN-ACCESS INTELLIGENT TUTORING SYSTEM IN HIGHER EDUCATION." In *EDULEARN24 Proceedings*, pp. 5938-5942. IATED, 2024.

[30] Bandara, Eranga, Sachin Shetty, Ravi Mukkamala, Abdul Rahman, Peter Foytik, Xueping Liang, Kasun De Zoysa, and Ng Wee Keong. "DevSec-GPT—Generative-AI (with Custom-Trained Meta's Llama2 LLM), Blockchain, NFT and PBOM Enabled Cloud Native Container Vulnerability Management and Pipeline Verification Platform." In *2024 IEEE Cloud Summit*, pp. 28-35. IEEE, 2024.