Integrating Retrieval-Augmented Generation and Large Language Models for Financial Question Answering

Yu-Jen Chen[1]
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
E-mail: b11902113@ntu.edu.tw

Ping-Han Chen[2, 3]
Internet of Things Laboratory
Chunghwa Telecom Laboratories
Taoyuan, Taiwan
Department of Engineering Science
National Cheng Kung University
Tainan, Taiwan
E-mail: rayche522@cht.com.tw

Tzu-Chia Tung[4]
Graduate Institute of Vehicle Engineering
National Changhua University of Education
Changhua, Taiwan
E-mail: tct@cc.ncue.edu.tw

Yung-Chien Chou[5, *]
Department of Mechatronics Engineering
National Changhua University of Education
Changhua, Taiwan
E-mail: ycc@cc.ncue.edu.tw

Yu-Chin Chu[1]
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
E-mail: d13944008@ntu.edu.tw

Sian-Wun Du[1]
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
E-mail: d14944001@ntu.edu.tw

Wei-Chien Wang[1]
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
E-mail: d08922038@ntu.edu.tw

Chiou-Shann Fuh[1]
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
E-mail: fuh@csie.ntu.edu.tw

Chung-Ming Yang[2]
Internet of Things Laboratory
Chunghwa Telecom Laboratories
Taoyuan, Taiwan
E-mail: cmyang@cht.com.tw

***Abstract*—Customer service systems in the financial industry require accurate and effi-cient question-answering solutions. Traditional methods, such as rule-based chatbots, struggle with complex queries, while Large Language Models (LLMs) face challenges like hallucination and outdated knowledge. This study explores the effectiveness of Retrieval-Augmented Generation (RAG) in answering financial questions using various retrieval methods, including BM25, embedding models, and reranker models. Experimental results show that BM25 is the fastest but less accurate, while the Reranker Model achieves the highest accuracy (0.8933) at a high computational cost. The best balance is found by combining BM25, a Reranker Model, and Recursive Token Chunker, improving accuracy (0.92) while reducing execution time (2427 seconds). This approach enhances AI-driven financial services by providing reliable and up-to-date responses.**

Keywords*—Financial Question Answering; Large Language Models (LLMs); Retrieval-Augmented Generation (RAG)

I. Introduction

Customer service systems in the financial industry rely heavily on accurate and timely information to provide reliable support. However, traditional question-answering systems, such as rule-based chatbots and FAQs, often fail to effectively handle complex queries. Large Language Models (LLMs) improve financial question answering by generating context-aware responses [1]. However, they also have some limitations, including hallucinations and outdated knowledge, which pose significant risks to financial decisions [2].

Retrieval-augmented generation (RAG) has emerged as a promising solution that ensures more accurate and up-to-date responses by combining real-time data retrieval with LLM [3]. This study aims to explore the effectiveness of RAG-enhanced LLM in financial question answering to address challenges such as response accuracy and efficiency. The findings could enhance AI-driven financial services, improving in-vestment analysis, customer support, and regulatory compliance.

II. Methodology

The dataset used in this study is provided by AI CUP 2024 [4]. The dataset is di-vided into three categories: Frequently Asked Questions (FAQ), Financial Statements, and Insurance Contract. For the FAQ category, 616 sample questions and answers are stored in JSON format, while the Financial Statements and Insurance Contract catego-ries contain 1,034 and 643 documents in PDF format, respectively. Additionally, the dataset includes 150 queries with ground truth (the ID of JSON entry or PDF file containing relative information) for evaluating the accuracy of RAG, evenly distribut-ed across the three categories.

Several methods are attempted to retrieve the most relevant document to the given query, including the BM25 algorithm [5], Embedding models (multilingual-e5-large) [6], and Reranker models (bge-reranker-v2-m3) [7]. In addition, a hybrid strategy is considered: BM25 first selects the top three most relevant documents, and a Reranker model then identifies the best match. Another approach involves truncating long documents into smaller chunks before applying a Reranker model. The trunking method (Recursive Token Chunker) is also tested to improve retrieval accuracy [8]. The above-mentioned RAG methods are evaluated using 150 queries with ground truth, measuring both accuracy and execution time.

To implement an automatic customer service system, we designed a comprehensive workflow that integrates RAG and LLM. The workflow is illustrated as follows:

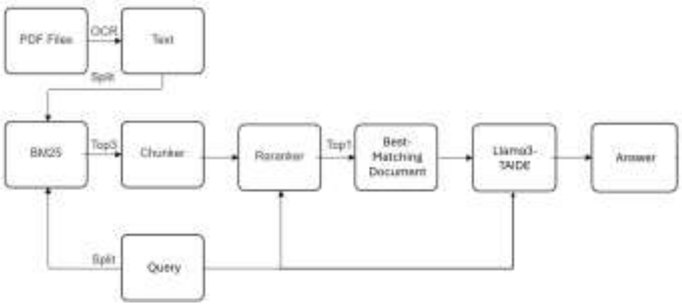

Fig. 1. That integrates RAG and LLM.

After finding the best-matching document with the RAG method (BM25 and Re-ranker model, with chunking), the query and the document are passed to the Llama3-TAIDE model to generate responses. Llama3-TAIDE is an LLM based on Meta's LLaMA3-8b model and pre-trained on traditional Chinese datasets. A structured prompt is designed to ensure the precise response from the model:

"Based on the given {Article Passage} and {Query}, generate the correct {Answer}. For example: <an example, including article passage, query, and answer>

Output Constraints:

1. Directly answer {Query} and generate {Answer} text.

2. The answer should be concise and not exceed 50 charac-ters.

3. Only output the answer without additional explanations or symbols.

{Article Passage}: <Put the document retrieve by RAG here.>

{Query}: <Put the query here.> "

III. Results and Observations

The following table shows the accuracy and execution time (complete 150 queries) of different RAG methods:

TABLE I. Accuracy and execution time (complete 150 queries) of different RAG methods.

Method	Accuracy	Execution Time (seconds)
BM25	0.74	103
Embedding Model	0.66	1,847
BM25 and Embedding Model	0.72	116
BM25 and Embedding Model, with Recursive Token Chunker	0.8133	152
Reranker Model	0.8933	21,750
BM25 and Reranker Model	0.86	1,305
BM25 and Reranker Model, with Recursive Token Chunker	0.92	2,427

978-8-3315-0166-2/25/$31.00 ©2025 IEEE

The experiment results show that BM25 is the fastest (103 seconds) but has rela-tively low accuracy (0.74). The Embedding Model alone is much slower (1,847 sec-onds) and, without chunking, performs worse than BM25 (0.66). Combining BM25 significantly improves retrieval speed, while Recursive Token Chunker enhances ac-curacy (up to 0.92). The Reranker Model achieves the highest accuracy (0.8933–0.92) but comes with extremely high computational costs (21,750 seconds). Overall, using both the BM25 and Reranker model with chunking provides the best balance between accuracy (0.92) and efficiency (2,427 seconds).

Each retrieval method has its own advantages and limitations. BM25 has high computational efficiency, but it relies on precise keyword matching and lacks semantic understanding, making its performance limited when processing synonymous or context-related queries. Embedding Models can capture semantic similarity, but they usually require inefficient candidate retrieval mechanisms and simple vector aggregation strategies; therefore, their accuracy is limited and the computational cost is high. In addition, the limitation on the input tokens further reduces the model performance in the absence of an effective chunking strategy. Reranker Models perform detailed analysis of queries and documents with large transformer models. Although they can achieve the highest accuracy, their huge computational resource requirements make it difficult to apply to real-time or large-scale retrieval scenarios. Considering both accuracy and efficiency, the ideal retrieval approach usually adopts a multi-stage strategy, using BM25 as a preliminary candidate retrieval, and then combining it with a Reranker Model for fine screening to achieve the best balance between performance and accuracy.

IV. Conclusion

This study evaluates the effectiveness of RAG-enhanced LLM in financial question answering, addressing challenges in accuracy and efficiency. The results show that BM25 is the fastest retrieval method but has moderate accuracy (0.74), while the Em-bedding Model alone is slower and less accurate (0.66) due to the lack of chunking. The Reranker Model achieves the highest accuracy (0.8933) but at a high computational cost (21,750 seconds). The optimal balance is achieved by combining BM25, a Reranker Model, and Recursive Token Chunker, resulting in an accuracy of 0.92 with a significantly reduced execution time (2,427 seconds). This approach ensures precise and efficient responses, making it a viable solution for AI-driven financial services. Future work could explore further optimizations in retrieval speed and response gen-eration quality, enhancing customer support, investment analysis, and regulatory com-pliance in the financial industry.

Acknowledgment

This work was supported by the National Science and Technology Council (NSTC) of Taiwan, under grants NSTC 114-2221-E-002-055, NSTC 114-2221-E-018 -011.

References

[1] N. Chinaksorn and D. Wanvarie, "LLM-RAG for Financial Question Answering: A Case Study from SET50," 2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 2025, pp. 0952-0957

[2] G. Perković, A. Drobnjak and I. Botički, "Hallucinations in LLMs: Understanding and Addressing Challenges," 2024 47th MIPRO ICT and Electronics Convention (MIPRO), Opatija, Croatia, 2024, pp. 2084-2088

[3] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401.

[4] https://tbrain.trendmicro.com.tw/Competitions/Details/37

[5] Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond (Vol. 3, No. 4). Foundations and Trends® in Information Retrieval, 333–389.

[6] Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., Wei, F. (2024). Multilingual E5 Text Embeddings: A Technical Report. arXiv preprint arXiv:2402.05672.

[7] Ma, X., Zhang, X., Pradeep, R., Lin, J. (2023). Zero-Shot Listwise Document Reranking with a Large Language Model. arXiv preprint arXiv:2305.02156.

[8] Gong, H., Shen, Y., Yu, D., Chen, J., Yu, D. (2020). Recurrent Chunking Mechanisms for Long-Text Machine Reading Comprehension. arXiv preprint arXiv:2005.08056.