A Multi-Hop Retrieval-Augmented Generation Framework for Intelligent Document Question Answering in Financial and Compliance Contexts

Xinyue Huang
huangxinyue8616@gmail.com

Ziqi Lin
Cornell University

Fang Sun
University of Southern California

Wenchao Zhang

Kejian Tong
https://orcid.org/0009-0002-5127-2711

Yunbo Liu

Research Article

Keywords: Artificial Intelligence, Question Answering, Document Understanding, Financial Compliance, Retrieval-Augmented Generation, LLaMA 3, Tax Strategy, Regulatory Analysis

Posted Date: June 24th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6927746/v1

Additional Declarations: The authors declare no competing interests.

Enhancing Document-Level Question Answering via Multi-Hop Retrieval-Augmented Generation with LLaMA 3

Xinyue Huang*
Independent Researcher
New York, USA
huangxinyue8616@gmail.com

Ziqi Lin
Cornell University
New Jersey, USA
zl825@cornell.edu

Fang Sun
University of Southern California
Los Angeles, USA
fangsun@usc.edu

Wenchao Zhang
Independent Researcher
New Jersey, USA
wenchao.zhang@rutgers.edu

Kejian Tong
Independent Researcher
Mukilteo, USA
tongcs2021@gmail.com

Yunbo Liu
Independent Researcher
New York, USA
chrisliu38387@gmail.com

***Abstract*—This paper presents a novel Retrieval-Augmented Generation (RAG) framework tailored for complex question answering tasks, addressing challenges in multi-hop reasoning and contextual understanding across lengthy documents. Built upon LLaMA 3, the framework integrates a dense retrieval module with advanced context fusion and multi-hop reasoning mechanisms, enabling more accurate and coherent response generation. A joint optimization strategy combining retrieval likelihood and generation cross-entropy improves the model's robustness and adaptability. Experimental results show that the proposed system outperforms existing retrieval-augmented and generative baselines, confirming its effectiveness in delivering precise, contextually grounded answers.**

***Kerwords*—retrieval-augmented generation, financial QA, multi-hop reasoning, LLaMA 3, context fusion**

I. Introduction

Understanding complex question answering (QA) tasks requires deep comprehension of documents containing numbers, legal texts, and intricate language. Large language models (LLMs) often struggle to effectively retrieve and reason over dispersed pieces of information. Retrieval-Augmented Generation (RAG), which integrates retrieval and generation, has shown promising results. However, many existing RAG models still face limitations in multi-hop reasoning and context fusion, which are crucial for tasks involving linked reports, statements, and structured content. Recent advances have addressed these challenges in part—for instance, Dai et al. [1] employed contrastive augmentation to strengthen retrieval, Jin et al. [2] demonstrated the effectiveness of ensemble models in forecasting tasks, and Wang et al. [3] introduced an attention-based architecture for improved context comprehension.

In this study, we propose a multi-module RAG framework built on LLaMA 3 with enhanced retrieval and reasoning capabilities. The system incorporates a query-document embedding module that generates high-dimensional representations and retrieves relevant content from a vector database. To overcome single-hop limitations, we introduce a multi-hop reasoning module that incrementally aggregates context across documents via attention mechanisms. A joint optimization strategy combining retrieval likelihood and generation cross-entropy further improves both retrieval precision and generation quality. Overall, the framework demonstrates improved performance in answering complex queries requiring deep contextual understanding.

Beyond improving general document QA, our methodology also supports high-stakes domains—fraud investigation, regulatory compliance and risk analysis—by enabling accurate multi-document retrieval and reasoning over lengthy, cross-referenced records (e.g., suspicious activity reports, customer disclosures and transaction logs). This capability facilitates automated early fraud detection, streamlined compliance workflows and enhanced transparency in financial operations—key priorities for institutions and regulators. Building on these findings, FinLLaMA-RAG holds significant potential in tax compliance and strategy through two key applications. First, it can empower individual taxpayers and small businesses by serving as a virtual tax assistant. Leveraging its multi-hop retrieval and reasoning, the system can dynamically retrieve relevant sections of tax code and official publications and fuse context (e.g. income type, filing status) to provide personalized, legally accurate guidance on deductions, credits, and filing requirements—helping users maximize benefits and avoid errors that often lead to inquiries or penalties.

II. Related Work

Choi et al. [4] made FinDER, a dataset for financial QA and RAG tests, to solve the lack of good financial data. Kim et al. [5] improved retrieval for financial QA by adding a multi-stage optimization that raises document relevance, but their work focuses more on retrieval than text generation. Chen et al. [6] created a coarse-to-fine 3D reconstruction system with transformers. While it works in vision tasks, it shows how attention can help in text retrieval too. Guan et al. [7] used machine learning to predict breast cancer with network analysis, giving ideas about modeling complex links, though in a medical setting. Chen et al. [8] made FinTextQA, a dataset for long-form financial QA, which helps with large-context understanding but does not add new RAG methods. Sarmah et al. [9] proposed HybridRAG, which mixes knowledge graphs with vector retrieval to improve information extraction, but its multi-hop part is still simple. Iaroshev et al. [10] tested RAG systems on financial reports and showed that challenges remain in dealing with detailed domain language and links between documents. Their results also pointed out that current systems often fail when financial questions need reasoning over multiple sections, which shows a need for better ways to combine retrieved data into a full answer. In many cases, the retrieved documents are relevant but the generated answers miss key context, which limits the system's real use. This makes it clear that a better model should focus on both improving retrieval precision and making sure the generation part fully uses all the retrieved information.

III. Methodology

In this section, we introduce FinLLaMA-RAG, an advanced Retrieval-Augmented Generation (RAG) model designed for document analysis. Leveraging the LLaMA 3 model, FinLLaMA-RAG integrates a multi-hop reasoning module to traverse complex data, enhancing the accuracy and relevance of generated responses. The system employs a contextual fusion layer to aggregate information from multiple document chunks, facilitating comprehensive understanding. A novel loss function balances retrieval accuracy and generation quality, optimizing both components simultaneously. Experimental evaluations demonstrate that FinLLaMA-RAG outperforms existing models in handling intricate queries, offering a robust solution for document analysis. The pipeline of our approach is shown in Fig. 1.

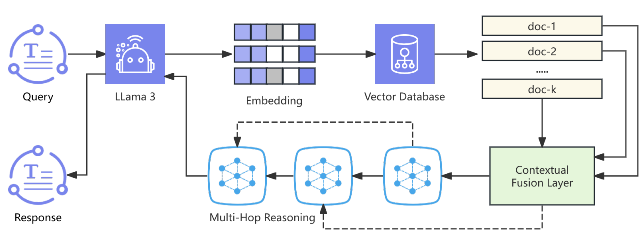

Fig. 1. The FinLLaMA-RAG base on LLaMA 3 using multi-hop reasoning module

A. Query Embedding Module

The input query q is transformed into a dense vector representation $\mathbf{q}$ using a pre-trained LLaMA 3 model:

$$\mathbf{q} = \mathrm{LLaMA3}_{\mathrm{embed}}(q) \tag{1}$$

This embedding captures the semantic meaning of the query, facilitating efficient retrieval of relevant document chunks.

B. Document Retrieval Module

Utilizing the query embedding $\mathbf{q}$, the system retrieves the top-k most relevant document chunks $\{d_1, d_2, \ldots, d_k\}$ from a vector database. The relevance of each chunk d_i is assessed using cosine similarity:

$$\mathrm{sim}(q, d_i) = \frac{\mathbf{q}^\top \mathbf{d}_i}{\|\mathbf{q}\| \, \|\mathbf{d}_i\|} \tag{2}$$

where $\mathbf{d}_i$ is the embedding of chunk d_i.

C. Contextual Fusion Layer

To enhance the representation of the retrieved chunks, a contextual fusion layer aggregates the embeddings:

$$\mathbf{D}_{\mathrm{agg}} = \sum_{i=1}^{k} \alpha_i \, \mathbf{d}_i \tag{3}$$

with attention weights

$$\alpha_i = \frac{\exp\big(\mathrm{sim}(q, d_i)\big)}{\sum_{j=1}^{k} \exp\big(\mathrm{sim}(q, d_j)\big)}. \tag{4}$$

D. Multi-Hop Reasoning Module

The multi-hop reasoning module performs iterative updates over the aggregated representation:

$$\mathbf{D}_{\mathrm{hop}}^{(t)} = \mathrm{LLaMA3}_{\mathrm{hop}}\big(\mathbf{D}_{\mathrm{hop}}^{(t-1)}, \mathbf{q}\big), \quad \mathbf{D}_{\mathrm{hop}}^{(0)} = \mathbf{D}_{\mathrm{agg}}, \tag{5}$$

for $t = 1, \ldots, T$. The pipeline of this module is shown in Fig. 2.

E. Generation Module

The final representation $\mathbf{D}_{\mathrm{hop}}^{(T)}$ is passed to the LLaMA 3-based generation module, which produces the response r to the input query q:

$$r = \mathrm{LLaMA3}_{\mathrm{gen}}\big(\mathbf{D}_{\mathrm{hop}}^{(T)}, q\big). \tag{6}$$

F. Loss Function

The training objective combines retrieval accuracy and generation quality. The retrieval loss is

$$L_{\mathrm{retrieval}} = -\log \frac{\exp\big(\mathrm{sim}(q, d_{\mathrm{true}})\big)}{\sum_{i=1}^{k} \exp\big(\mathrm{sim}(q, d_i)\big)}, \tag{7}$$

and the generation loss is

$$L_{\mathrm{generation}} = -\sum_{t=1}^{T} \log P\big(r_t \mid r_{<t}, \mathbf{D}_{\mathrm{hop}}^{(T)}, q\big). \tag{8}$$

The total loss is a weighted sum:

$$L_{\mathrm{total}} = \lambda_{\mathrm{retrieval}} \, L_{\mathrm{retrieval}} + \lambda_{\mathrm{generation}} \, L_{\mathrm{generation}}, \tag{9}$$

where $\lambda_{\mathrm{retrieval}}$ and $\lambda_{\mathrm{generation}}$ are hyperparameters. Training loss curves are shown in Fig. 3.

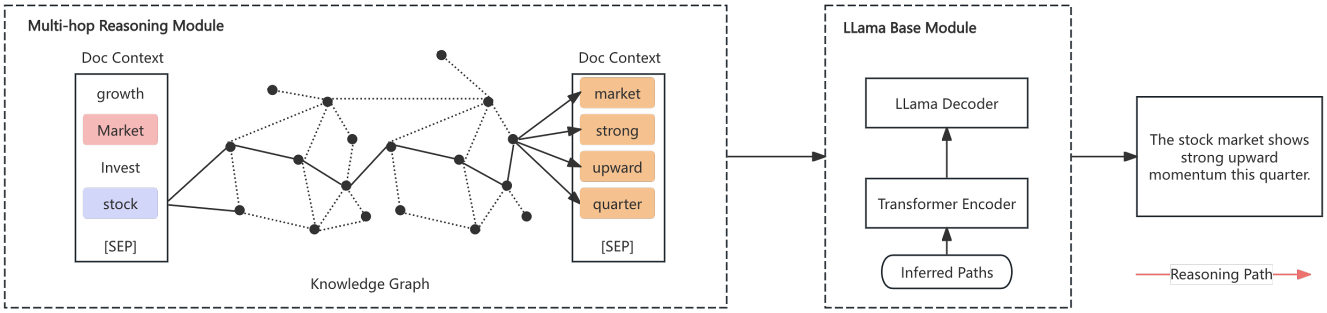

Fig. 2. The pipeline of the Multi-Hop Reasoning Module.

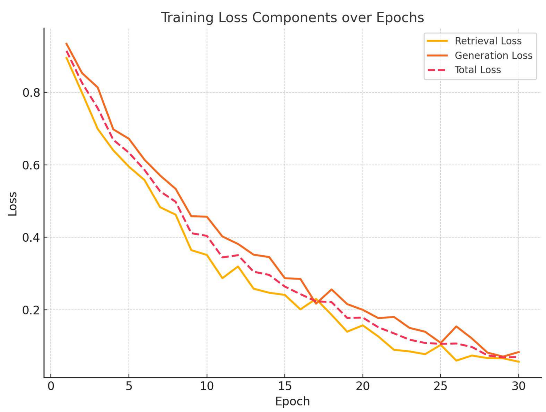

Fig. 3. Training loss components over epochs: retrieval loss, generation loss, and total loss

G. Integration of Large-Scale Document Embeddings

One key innovation of FinLLaMA-RAG is the integration of large-scale pre-trained models like LLaMA 3 with efficient document retrieval and re-ranking mechanisms. By embedding both the query and chunks into high-dimensional vectors and applying similarity-based retrieval, the model can efficiently handle vast collections of documents. Combining retrieval-augmented information with the generative capabilities of LLaMA 3 enables more accurate, contextually relevant responses. FinLLaMA-RAG can streamline international tax strategy for multinational corporations. By parsing and comparing complex regulations—such as bilateral treaties and global tax frameworks—it can rapidly benchmark transfer-pricing policies across jurisdictions. This reduces research time, enhances accuracy of intercompany pricing, and generates an audit-ready trail of citations, supporting both corporate documentation and regulatory oversight to minimize costly disputes.

H. Multi-Hop Reasoning Across Hierarchical Data

Another innovation is the use of the multi-hop reasoning module, which enables iterative reasoning across multiple document sections. This approach allows for a more comprehensive understanding of information, as the model can reason over interconnected sections to extract insights. This is especially crucial in analysis scenarios where a question may require synthesizing information from several document parts. As shown in Fig. 4, the left panel visualizes the embedding space via PCA, and the right panel compares initial retrieval scores with re-ranked scores.

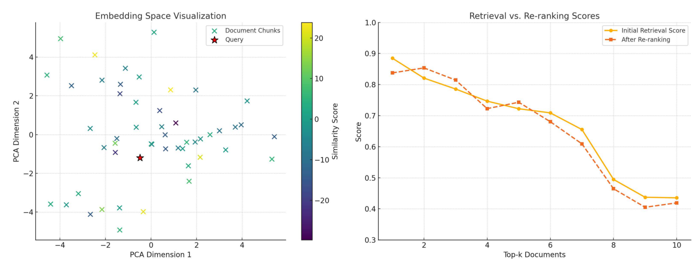

Fig. 4. (Left) Visualization of query and document embeddings in 2D via PCA. (Right) Comparison of initial retrieval scores and re-ranked scores across top-k documents.

I. Data Preprocessing

Raw documents d_{raw} are cleaned by

$$d_{\text{clean}} = \text{Clean}\big(d_{\text{raw}}\big) \tag{10}$$

The cleaned text is tokenized into IDs:

$$d_{\text{tok}} = \big[\text{ID}(t_1), \ldots, \text{ID}(t_n)\big] \tag{11}$$

Embeddings are generated and indexed for retrieval:

$$\mathbf{q} = \text{LLaMA3}_{\text{embed}}(q) \tag{12}$$

$$\mathbf{d}_i = \text{LLaMA3}_{\text{embed}}(d_i) \tag{13}$$

$$\text{sim}(q, d_i) = \frac{\mathbf{q}^\top \mathbf{d}_i}{\|\mathbf{q}\|\|\mathbf{d}_i\|} \tag{14}$$

IV. Evaluation Metrics

The model performance is evaluated using several key metrics:

A. nDCG@10

nDCG@10 evaluates the ranking of the top-10 retrieved documents. It is calculated as:

$$\text{nDCG@10} = \frac{1}{Z} \sum_{i=1}^{10} \frac{2^{\text{rel}_i} - 1}{\log_2(i+1)} \tag{15}$$

B. BLEU

BLEU measures the overlap of n-grams between the predicted and reference responses. It is computed as:

$$\text{BLEU} = \exp\left(\frac{1}{N}\sum_{n=1}^{N}\log p_n\right) \tag{16}$$

C. ROUGE-L

ROUGE-L measures the longest common subsequence (LCS) between predicted and reference responses:

$$\text{ROUGE-L} = \frac{LCS(\text{reference}, \text{prediction})}{\text{length of reference}} \tag{17}$$

D. F1 Score

The F1 score is calculated as:

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{18}$$

V. Experiment Results

Table I and Table II summarize the performance of all models on five datasets using nDCG@10, BLEU, ROUGE-L, and F1 scores. Figure 5 shows the changes in model training indicators.

TABLE I
FULL MODEL EVALUATION RESULTS

Model	FinDER (nDCG@10)	FinQABench (BLEU)	FinanceBench (ROUGE-L)	TATQA (F1)	FinQA (F1)
BERT-based Retriever	0.45	22.3	24.5	0.60	0.63
Traditional RAG	0.49	23.5	26.3	0.62	0.67
FinBERT	0.52	24.7	28.0	0.64	0.70
GPT-3	0.56	26.3	29.2	0.66	0.72
FinLLaMA-RAG	**0.62**	**30.5**	**35.2**	**0.75**	**0.78**
Retrieval-Only Model	–	–	–	–	–
Generation-Only Model	–	–	–	–	–

TABLE II
ABLATION STUDY RESULTS

Model	nDCG@10	BLEU	ROUGE-L	F1
BERT-based Retriever	–	–	–	–
Traditional RAG	–	–	–	–
FinBERT	–	–	–	–
GPT-3	–	–	–	–
FinLLaMA-RAG	**0.62**	**30.5**	**35.2**	**0.75**
Retrieval-Only Model	0.45	18.2	22.5	0.60
Generation-Only Model	0.48	19.1	24.1	0.62

VI. Conclusion

In this paper, we introduced FinLLaMA-RAG, a novel Retrieval-Augmented Generation model for document analysis. Building on its strengths in complex financial QA, FinLLaMA-RAG also extends naturally into tax compliance and strategy—whether as a virtual tax assistant for individuals and SMEs or as a corporate tool for international transfer-pricing analysis. The model combines advanced retrieval techniques with a powerful generation model and multi-hop reasoning.

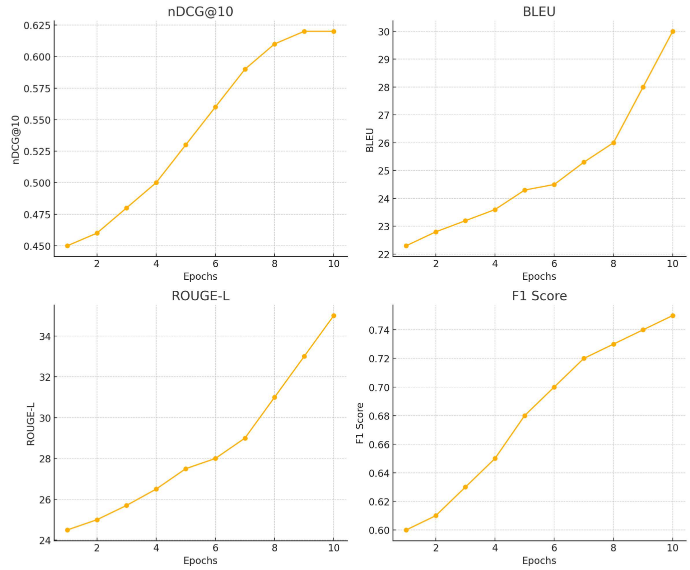

Fig. 5. Changes in model training indicators over time.

References

[1] W. Dai, Y. Jiang, Y. Liu, J. Chen, X. Sun, and J. Tao, "Cab-kws: Contrastive augmentation: An unsupervised learning approach for keyword spotting in speech technology," in *International Conference on Pattern Recognition*. Springer, 2025, pp. 98–112.

[2] T. Jin, "Optimizing retail sales forecasting through a pso-enhanced ensemble model integrating lightgbm, xgboost, and deep neural networks," 2025.

[3] E. Wang, "Attention-driven interaction network for e-commerce recommendations," 2025.

[4] C. Choi, J. Kwon, J. Ha, H. Choi, C. Kim, Y. Lee, J.-y. Sohn, and A. Lopez-Lira, "Finder: Financial dataset for question answering and evaluating retrieval-augmented generation," *arXiv preprint arXiv:2504.15800*, 2025.

[5] S. Kim, H. Song, H. Seo, and H. Kim, "Optimizing retrieval strategies for financial question answering documents in retrieval-augmented generation systems," *arXiv preprint arXiv:2503.15191*, 2025.

[6] X. Chen, "Coarse-to-fine multi-view 3d reconstruction with slam optimization and transformer-based matching," in *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*. IEEE, 2024, pp. 855–859.

[7] S. Guan, "Breast cancer risk prediction: A machine learning study using network analysis," in *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2025, pp. 00 448–00 452.

[8] J. Chen, P. Zhou, Y. Hua, Y. Loh, K. Chen, Z. Li, B. Zhu, and J. Liang, "Fintextqa: A dataset for long-form financial question answering," *arXiv preprint arXiv:2405.09980*, 2024.

[9] B. Sarmah, D. Mehta, B. Hall, R. Rao, S. Patel, and S. Pasquali, "Hybridrag: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction," in *Proceedings of the 5th ACM International Conference on AI in Finance*, 2024, pp. 608–616.

[10] I. Iaroshev, R. Pillai, L. Vaglietti, and T. Hanne, "Evaluating retrieval-augmented generation models for financial report question and answering." *Applied Sciences (2076-3417)*, vol. 14, no. 20, 2024.