Leveraging Large Language Models for Enhanced Insights in Multi-Document Question Answering

Juluri Shiva Sai[1], Vantala Swamynath[2], Bethi Mayookh[3],
Undrakonda Ananth Lakshmi Srinivas[4], Hilly Gohain Baruah[5] and Rakcinpha Hatibaruah[6]
{shivasaijuluri2004@gmail.com[1], swamynathvantala14@gmail.com[2], mayookhbethi@gmail.com[3],
srinivasundrakonda95@gmail.com[4], hgbaruah1990@gmail.com[5], rak300aei@gmail.com[6] }

Department of Advanced Computer Science and Engineering Vignan's Foundation for Science,
Technology and Research Guntur, Andhra Pradesh, India[1, 2, 3, 4, 5, 6]

Abstract. For data retrieval and comprehensive natural language understanding, answering multiple documents at the same time (MDQA) is essential with the basics of Large Language Models (LLMs). This paper is an attempted research work in formulating a MDQA system using the recent LLMs to solve questions from multiple passing documents. Moreover, we use F1-score, BLEU and ROUGE as evaluation metrics to assess the quality of the generated responses. Our results show that our approach improves significantly in answer correctness, coherence and contextual relevance over the baseline Answer extraction models. In doing so we present important insights into generation of LLMs for MDQA and pave ground towards further advancements in multi-document reasoning and knowledge incorporation.

Keywords: Multi-document question answering, Large Language Models, Information Retrieval, Natural Language Understanding, Evaluation Metrics

1 Introduction

The Multi-Document Question Answering (MDQA) System is meant to provide more comprehensive and detailed information from multiple sources, than traditional Question Answering (QA) systems that rely on information extraction from a single document. It addresses the rising need for contextually relevant and accurate responses in domains such as the health-care, law, and academic sectors, simplyfying information retrieval to enhance precision and speed.

MDQA is built on advances in linguistic computing, information extraction, and machine learning methods. For example, traditional extractive models such as BERT, which are pre-trained on datasets such as SQuAD and then fine-tuned, struggle with queries that require synthesis over several documents. Novel methods including Retrieval-Augmented Generation and Fusion-in-Decoder are proposed to better integrate information but have difficulty in preserving consistency and handling contradicting information.

Large language models as Mistral-7B, Mixtral-8x7B and LLaMA-3. 2-3B) enhance MDQA with successful processing of massive texts. Retrieval-based methods like Dense Passage Retrieval (DPR) and ColBERT improve accuracy, while generative models like GPT and LLaMA aid comprehensive summarization. While progress has been made, challenges persist concerning reducing duplication, addressing contractions and ensuring verifiable correctness. MDQA is an active subject of research to provide valuable knowledge discoveries from multiple

ICITSM-Part I 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357854

sources of information. This research examines contemporary studies on techniques and advancements in MDQA systems. Traditional approaches employing NLP were investigated in [1]. In [2], the use of quasi-knowledge graphs to integrate evidence from multiple documents has been shown which demonstrated improvement in accuracy. It enhances information retrieval by structuring and linking extracted facts for better contextual understanding. ”FanOutQA” is a benchmark designed to assess large language models on multi-hop reasoning across multiple doc- uments, enhancing MDQA evaluation. It tests large language model (LLM)s’ ability to integrate and infer information from diverse sources [3]. The work in [4] explores joint optimization of retrieval and reading components in Open Domain Question Answering, enhancing accuracy and system performance. It emphasizes end-to-end training to improve information retrieval and comprehension. ”Visconde” explores how GPT-3 and neural reranking enhance answer selection by refining relevance and accuracy. It demonstrates how these techniques improve MDQA by prioritizing the most relevant information [5]. Datasets are essential for the advancement of MDQAsystems. The HOTPOTQA dataset [6] enables multi-hop question answering by demanding reasoning that synthesizes information from several documents. It emphasizes explainability and diverse reasoning paths, aiding MDQA system development. Furthermore, in [7] the novel strategies using knowledge graphs to enhance LLM prompting, improving context understanding and response accuracy in MDQA tasks. This approach ensures more structured, reliable, and context- aware answers. In [8] [11], the scope of MDQA is expanded using a pre-training strategy that boosts its effectiveness. enhances their capacity to manage several documents successfully. ”HiQA” improves large-scale document QA by using cascading metadata and a multi-route retrieval technique for more accuracy and efficiency. It improves context interpretation and retrieval in large document contexts [9] [12].

[10] [13] proposes a multi-document setting QA evaluation framework that additionally considers various information forms (table, chart and presentation). It measures how good these systems are at assimilating and processing visual and textual input to give accurate responses.

Recent progresses further enrich the MDQA scene. The Mistral-7B model [17] [18] [21] [14] enhances the understanding of multi- document context, whereas the Mixtral8x7b model [19] [20] [22] [15], leverages a dynamic mixture-of-experts approach and the notion of resource optimisation to scale ef- ficiently and performantly.

These trends highlight the dynamically changing nature of MDQA research, which has adopted a diverse set of approaches to improve the performance of systems that integrate information from multiple documents along several dimensions of performance. MDQA System enhances the quality of information retrieved by aggregating multiple sources of information, reducing the bias, and maintaining its precision. Its context-based approach enables elaborative relevant responses, improves decision capacities and summary of information. Scalability is still a challenge requiring efficient data handling. Generalization across domains is crucial for broader applicability. Addressing these issues is critical for enhancing its performance in various applications

There are many challenges in building a MDQA System. One major challenge is the preparation of the data, with documents coming in different formats (PDFs, webpages, scanned images) which need to be consistently cleaned and standardized. Another problem is to understand con- text across multiple documents, especially when different documents provide divergent, redundant, or overlapped information items. Setting up a retrieval system is difficult, due to the fact, that indexing and ranking documents both have to overcome the trade-off between speed and accuracy for large datasets. Additionally, around comprehension problems there are more NLP issues that come into play when interpreting complex or ambiguous or specialized text, often requiring entity identification and coreference resolution. Generating answers to queries faced further difficulties, as the synthesis and generation of answers had to ensure that the latter would be concise, logical and correct, without falling on the trap of providing wrong information or even fabrications especially when treating great masses of data and maintaining reactivity on the fly. System accuracy assessment is also challenging as it requires setting up evaluation metrics covering relevance and completeness of the perception, and accounting for subjective aspects in multi-document responses [16] .

The important contributions of our research are listed below:

- Created a multi-document QA system that integrates data from several sources to deliver answers that are both truthful and thorough.
- Implemented an efficient NLP-based retrieval mechanism to improve response relevance and consistency.
- Evaluate performance using several measures, such as F1- score, BLEU score, and ROUGE.
- Showcase the system's application in many domains, including healthcare, law and finance.

This paper is structured as follows: Section 1 introduces the problem and analyzes existing QA systems. Section 2 outlines the suggested technique, which includes a flowchart and procedure specifics. Section 3 assesses the system's performance using the F1-score, BLEU score, and ROUGE metrics. Section 4 summarizes major results and outlines next research options.

2 Proposed Methodology

This chapter describes the MDQA system's pipeline in conjunction with LLM. The work flow diagram is illustrated in Fig. 1:

The details are discussed below.

- Step 1: Extracting Information from a Source Document (PDF, Link to Web Site). This extract step is when raw material is getting harvested & ready to go further for processing. After the data is fetched, it divides into small piece. In other words, these segments allow us to process more text faster and give better retrieval results since the document is divided into smaller pieces.
- Step 2: Once content has been broken down into segments, for all identified segment one must go ahead and construct an embedding. This article is about one of these methods: word embeddings — mapping words to a high-dimensional feature space. Words that occur in similar contexts have similar embeddings and again this property crucially enables the capture of meaning. An embedding is the vector representation of a word; enabling the model to learn connections between the text. All of these text expressions are then simply mapped to the embedding layer and bound together in some way, so model understands where they each stand with respect to one another (keeping long term dependencies). Next, we use these embeddings to create an semantic index that assists us in ordering and keeping structure for the data thereby making the search intelligent. The

embeddings are then indexed on a vector database to enable the efficient and accurate search operations.

- Step 3: Turning the query into an embedding: The technology can transform a user's question into an embedding. System Considerations Semantic Search: Matching of the query embedding to vector database embeddings indexed before. This is how to obtain contextually relevant information, as opposed to single out word match.
- Step 4: Upon determining the relevant text segments, they are ranked based on the similarity to the user's query. The most relevant results are then directly shown to the user, or used as additional context for an LLM to generate a response. The LLM takes a look at the ranked results, and generates a logical and contextually correct response that is served back to the user.

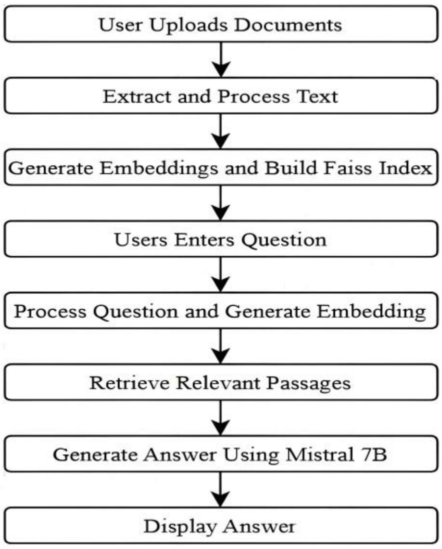
User Uploads Documents
Extract and Process Text
Generate Embeddings and Build Faiss Index
Users Enters Question
Process Question and Generate Embedding
Retrieve Relevant Passages
Generate Answer Using Mistral 7B
Display Answer

Fig. 1. The flowchart of the MDQA System with LLM

This full pipeline ensures that users may efficiently extract relevant information from large documents by combining se- mantic search with LLM-driven answers to increase accuracy and comprehension.

Fig.2 shows a schematic of the MDQA system with LLM. It depicts the MDQA system which effectively handles docu- ments (such as PDFs or URLs) by extracting text, segmenting it into smaller parts, and transforming each segment into em- beddings (concise vector representations). These embeddings are used to create a semantic index in a vector database, enabling quick similarity-based searches. When a user asks a question, it's transformed into a query embedding and compared to the indexed embeddings. The most relevant text segments are retrieved, ranked by similarity, and fed into a generative AI (powered by an LLM) to produce a coherent and contextually appropriate answer. This system enhances search accuracy and generates human-like responses, providing thorough insights from multiple documents efficiently.

Mistral-7B is a dense transformer model consisting of 7 billion parameters, designed for effective reasoning and organized text creation [21]. It is especially adept at deriving insights from various documents while preserving contextual integrity. Mixtral-8x7B, a mixture of experts architecture, is made up of eight networks with 7 billion parameters each, with just two experts engaged at any moment, operating like a 14 billion parameter model while maintaining lower computational expenses than entirely dense models [22]. This architecture enhances its efficiency and allows it to manage intricate multi-document reasoning tasks proficiently. LLaMA- 3.2-3B, the smallest among the trio, is a streamlined model tailored for quick inference and resource-constrained settings. Although it is not as strong as the other two, it is essential for managing simpler queries and performing pre-filtering tasks before sending them to the larger models for more in- depth analysis. By integrating these models, system achieves a balance of high-quality answers, scalability, and computational efficiency, providing strong and thorough insights from various documents.

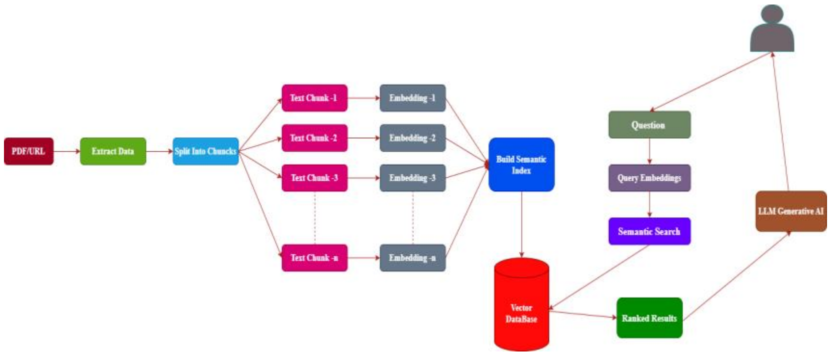

Fig. 2. The block schematic of the MDQA System with LLM

3 Evaluation and Analysis

Here, it explains experimentally the results of system performance evaluation with such metrics in different aspects to examine how well and good approach work. Here are the specific assessment criteria.

F1-Score

F1-score is one of the critical metrics used in classification models especially when data is itself class-imbalanced. It represents the harmonic mean of accuracy and recall, so as to be a balance measurement for both false positive and false negative. Precision measures the accuracy of the identified positive instances, whereas recall indicates how many true and actual positive cases were found correctly.

$$F1 = \frac{2PR}{P+R} \qquad (1)$$

where P = Precision, R = Recall

Bleu Score

The BLEU score is a metric used to evaluate a generated text by a machine especially those produced by the MT(Machine Translation) and NLG (Natural Language Generation). Measures the similarity of generated text to one or multiple reference texts by comparing n-gram overlap. Its value will always be between 0 and 1 and a score of 1.0 means that the candidate text is exactly the same as a reference in every possible way while on the other hand, BLEU score of 0.0 indicates that the candidate generates has no resemblance with any reference.

$$BLEU \;=\; BP.\,e^{\frac{1}{N}(\log p1+\log p2+\ldots+\log pN)} \tag{2}$$

where:

$$BP \;= \begin{cases} 1, & if\ c > r \\ e^{\left(1-\frac{r}{c}\right)}, & otherwise \end{cases} \tag{3}$$

where e = Euler's constant,

r = reference length,

c = candidate length,

pn = n-gram precision,

N = maximum n-gram order,

BP = brevity penalty

Rouge:

The f1 score for ROUGE-1 is the most often metric in evaluating text summarisation and NLG models. It calculates the unigram overlap of generated and reference summaries It consists of three parts- Precision, Recall, and a harmonic mean between them(F1-score) And precision is a measure of how accurate the generated text is in relation to the original one. and Recall and breaks them in the sum of words in reference summary and the words available in system summmary.

ROUGE-2 is also a common metric for NLP. It measures the number of two-word sequences that are in common between the generated text and a human-authored reference. ROUGE-2 is grounded in the idea that a good summary should conserve the important words or word n-grams of the original document.

ROUGE-L is a commonly used measure for evaluating the effectiveness of automatically generated text, particularly in tasks like text generation and summarization. Unlike typical ROUGE metrics, which focus on n-gram overlap, t identifies the longest matching sequence of words between the predicted text and the original content.

ROUGE-L (LongestCommonSubsequence-LCS) is assessed by looking at precision, recall, and the F1 score based on the longest shared word sequence.

$$ROUGE - L_R = \frac{LCS(Ref,Cand)}{Length(Ref)} \tag{4}$$

$$ROUGE - L_P = \frac{LCS(Ref,Cand)}{Length(Cand)} \tag{5}$$

$$ROUGE - L_{F1} = \frac{(1+\beta^2).ROUGE-L_P.ROUGE-L_R}{\beta^2.ROUGE-L_P+ROUGE-L_R} \quad (6)$$

A comparison of the performances of the three models—Mistral-7B, Mixtral-8x7B, and LLaMA3.2-3B—across multiple metrics reveals unique strengths and shortcomings, as demonstrated in Table 1 . Presented here is an evaluation of which model excelled for each metric and the likely explanations for these outcomes. Mixtral-8x7B excels in ROUGE and BERTScore, signifying its improved ability to align n-grams and semantic similarity. This is consistent with its MoE structure, which specializesin managing varied and intricate inquiries by engaging expertstailored to specific tasks. LLaMA-3.2-3B excels in BLEURT, indicating it generates more human-like replies, likely a result of fine-tuning on user feedback or its reduced size preventing over-optimization for basic metrics. Mistral-7B is competitive, yet not the top choice, probably due to its absence of MoE flexibility and being surpassed by Mixtral in many metrics. However, it continues to outperform LLaMA-3.2-3B in both ROUGE and BERTScore, highlighting its capabilities as a dense model.

The findings emphasize that the model architecture (MoE versus dense) and training goals have a considerable effect on performance across various metrics. The MoE design of Mixtral-8x7B provides an advantage in factual and semantic tasks, while the LLaMA-3.2-3B tuning could benefit the outputs favored by humans. Mistral-7B acts as a solid foundation, but is outperformed by its MoE equivalent. For applications that focus on precision and recall (e.g., QA systems), Mixtral8x7B is the optimal option. For fluency similar to that of humans, LLaMA-3.2-3B may be a better choice.

Table 1. Performance Comparison of Models on Different Metrics.

Metrics	Mistral-7B	Mixtral-8x7B	LLama-3.2-3B
ROUGE-1	0.5669	0.6524	0.3668
ROUGE-2	0.4182	0.4956	0.2316
ROUGE-L	0.5272	0.6037	0.3273
BERT Score	0.7122	0.7341	0.5840
BLEURT	-0.4581	-0.3616	-0.8017

4 Conclusion

This is achieved in this work through semantic search and LLMs that drive information retrieval from large text sources such as PDFs and URLs. It is based on making a smart caching of related information by using text chunking, embeddings, vector databases and rating algorithms. The incorporation of LLM enhances the replies by rendering them more coherent, accurate and informative. Users can simply ask natural language questions and get correct answers, rather than spending time sifting through lengthy documentation. Through the semantic indexing and vector search, the fast and accurate retrieval provides very valuable support for academic research, corporate documentation, legal text mining, medical treatment and so on. In conclusion, the effort showcases a knowledge extraction approach that is scalable, intelligent, and automatic in efficiently extracting and consuming information.

References

[1] Shah, Shalin, Srikanth Ryali, and Ramasubbu Venkatesh. "Multi-Document Financial Question Answering using LLMs." arXiv preprint arXiv:2411.07264 (2024).

[2] Lu, Xiaolu, et al. "Answering complex questions by joining multi-document evidence with quasi knowledge graphs." Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. 2019.

[3] Zhu, Andrew, et al. "FanOutQA: A multi-hop, multi-document question answering benchmark for large language models." arXiv preprint arXiv:2402.14116 (2024).

[4] Singh Sachan, Devendra, et al. "End-to-End Training of Multi-Document Reader and Retriever for Open Domain Question Answering." arXiv e-prints (2021): arXiv-2106.

[5] Pereira, Jayr, et al. "Visconde: Multi-document qa with gpt-3 and neural reranking." European Conference on Information Retrieval. Cham: Springer Nature Switzerland, 2023.

[6] Yang, Zhilin, et al. "HotpotQA: A dataset for diverse, explainable multi-hop question answering." arXiv preprint arXiv:1809.09600 (2018).

[7] Wang, Yu, et al. "Knowledge graph prompting for multi-document question answering." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 17. 2024.

[8] Caciularu, Avi, et al. "Peek across: Improving multi-document modeling via cross-document question-answering." arXiv preprint arXiv:2305.15387 (2023).

[9] Chen, Xinyue, et al. "HiQA: A Hierarchical Contextual Augmentation RAG for MultiDocuments QA." arXiv preprint arXiv:2402.01767 (2024).

[10] Suri, Manan, et al. "VisDoM: Multi-Document QA with Visually Rich Elements Using Multimodal Retrieval-Augmented Generation." arXiv preprint arXiv:2412.10704 (2024).

[11] Bai, Yushi, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du et al. "Longbench: A bilingual, multitask benchmark for long context understanding." arXiv preprint arXiv:2308.14508 (2023).

[12] Kwan, Wai-Chung, Xingshan Zeng, Yufei Wang, Yusen Sun, Liangyou Li, Lifeng Shang, Qun Liu, and Kam-Fai Wong. "M4le: a multi-ability multi-range multi-task multi-domain long-context evaluation benchmark for large language models." arXiv preprint arXiv:2310.19240 (2023).

[13] Abujabal, Abdalghani, et al. "Automated template generation for question answering over knowledge graphs." Proceedings of the 26th international conference on world wide web. 2017.

[14] Aly, R., Guo, Z., Schlichtkrull, M., Thorne, J., Vlachos, A., Christodoulopoulos, C., ... , Mittal, A. (2021). Feverous: Fact extraction and verification over unstructured and structured information. arXiv preprint arXiv:2106.05707.

[15] Lin, C. Y.,Hovy, E. (2002, July). From single to multi-document summarization. In Proceedings of the 40th annual meeting of the association for computational linguistics (pp. 457- 464).

[16] Wan, X. and Yang, J., 2008, July. Multi-document summarization using cluster-based link analysis. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 299-306).

[17] Christensen, J., Soderland, S. and Etzioni, O., 2013, June. Towards coherent multi-document summarization. In Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies (pp. 1163-1173).

[18] Bolotova-Baranova, Valeriia, Vladislav Blinov, Sofya Filippova, Falk Scholer, and Mark Sanderson. "WikiHowQA: A comprehensive benchmark for multi-document non-factoid question answering." In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 5291-5314. 2023.

[19] Caciularu, A., Peters, M.E., Goldberger, J., Dagan, I. and Cohan, A., 2023. Peek across: Improving multi document modeling via cross-document question-answering. arXiv preprint arXiv:2305.15387.

[20] Abualhaija, S., Arora, C., Sleimi, A. and Briand, L.C., 2022, August. Automated question answering for improved understanding of compliance requirements: A multi-document study. In 2022 IEEE 30th international requirements engineering conference (RE) (pp. 39-50). IEEE.

[21] Jiang, Albert Q., et al. "Mistral 7B." arXiv preprint arXiv:2310.06825 (2023).

[22] Chaplot, Devendra Singh, et al. "Mixtral of Experts." arXiv preprint arXiv:2401.04088 (2024).