2024 International Conference on Artificial Intelligence and Power Systems (AIPS) | 979-8-3503-9221-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/AIPS64124.2024.00057

Enhancing Large Model Document Question Answering through Retrieval Augmentation

Jingwen Zeng [1a], Rongrong Zheng [1b], Chenhui Wang [1c], Wenting Xue [1d], Xiaoyang Yu [1e], Tao Zhang [2*]
[1] State Grid Information & Telecommunication Branch, Beijing, China
[2] State Grid Sgitg Digital Technology(Beijing) Co., Ltd, Beijing, China
[a]sgitb3099@163.com, [b]christinazrr@126.com, [c]264025558@qq.com, [d]xuewt442@163.com, [e]yu1224754419@163.com,
*Corresponding author Email: zhangtao01@sgitg.sgcc.com.cn

Abstract **Document question answering(DocQA) requires models to provide comprehensive answers, given a series of document content and questions. In recent years, large neural models have been widely applied in fields such as natural language processing and computer vision, yielding significant results. However, these applications usually face a challenge that the large language model often produce fake information, owing to model illusions. To tackle this problem, the recent proposed document question answering has proven to be effective. Current methods primarily rely on vector for passage retrieval, but the effectiveness is often limited, which restrict the model's performance. To address this issue effectively, we propose a dual-path retrieval along with precise ranking framework to enhance existing knowledge retrieval. To better evaluate this task, we also constructed a manually annotated test set for validation. Experimental results demonstrate the significant advantages of our model.**

Keywords: document question answering, large language modal

I. Introduction

Document question answering(DocQA) requires models to provide comprehensive answers based on a series of document content and questions. Currently, the emergence of numerous large-scale language models, such as ChatGPT [4], Baichuan [7], qwen [6], Vicuan [9], and Llama [8] attracts wild attention. They possess advantages in terms of data, parameters, and architecture, exhibiting powerful language comprehension and generation capabilities, leading to outstanding performance in natural language processing tasks. Therefore, they are widely applied in downstream natural language processing and information extraction tasks. However, these models may encounter data distributions during training that differ from those in actual application scenarios, making it impossible to access real-time, non-public, or offline data, resulting in a lack of relevant knowledge and the generation of false information in practical applications, known as model illusions. Hence, it is risky to directly apply such large-scale language models for information retrieval and queries.

To address this issue, some methods [12,13,14,15,16] adopt external knowledge to alleviate the illusions of large models. By incorporating an external knowledge base, models can access additional background knowledge and information, which helps enhance the model's generalization ability and robustness. These external knowledge resources can encompass various types of data, including structured data, text data, graph data, and more. By integrating this external knowledge, models can better understand and interpret data, enrich their semantic expression capabilities, and improve their ability to generalize to unknown data. External knowledge can also mitigate data sparsity issues and assist models in learning domain-specific information. The introduction of an external knowledge base can enhance the model's cognitive abilities, improve its performance across various tasks, effectively mitigating the hallucination issues present in large models.

In order to accurately generate answers to questions, the core module of introducing external knowledge relies on the model's ability to find the most relevant parts from a massive corpus of knowledge. Currently, there are two mainstream methods for implementing the document retrieval step: 1) the use of vector retrieval, such as models like M3E [2], Text2Vec [1], OpenAI-Ada [4], etc. These methods, trained on large-scale sentence pair datasets, can cover the majority of retrieval scenarios and are highly efficient, making them wildly used in industrial applications. 2) The use of matching models, such as BERT [10] and BGE [5]. Although these models often achieve promising results, they are less efficient in large-scale text retrieval and reasoning, thus limiting their application in document question answering.

Current DocQA methods often employ vector retrieval to retrieve document fragments. Although this method can handle some specific usage scenarios, it has two drawbacks that limit the capabilities of document question answering: 1) Since the vector retrieval model is based on retrieving the entire sentence, it performs poorly when dealing with sparse documents (i.e., documents containing a large amount of scattered keyword information); 2) Compared to the deep interactive capabilities of matching models, vector models often shows low accuracy, making it easier to retrieve a large number of irrelevant documents during the retrieval process, thus affecting the generated results. In this paper, we propose a dual-path retrieval composite recall architecture, to overcome the aforementioned limitations. To address the shortcomings of vector models in dealing with sparse information retrieval, we introduce lexical matching to enhance the retrieval of sparse information. Through this dual-path retrieval approach, effective information can be confined to a very small range. This step is also known as coarse-ranking process. To further enhance the model's effectiveness, we also design a matching model for precision ranking retrieval, thereby further filtering out irrelevant information. Since the dual-path retrieval has removed most irrelevant

979-8-3503-9221-0/24/$31.00 ©2024 IEEE
DOI 10.1109/AIPS64124.2024.00057

information, the candidate texts are greatly reduced, ensuring the efficiency of the precision ranking retrieval. Furthermore, as there is currently no publicly available Chinese test set for document question answering, we constructed a document question answering test set to evaluate the effectiveness of different models, which has been annotated by professionals. Experimental results demonstrate that our proposed approach surpasses previously proposed models, validating the effectiveness of this method.

II. Related Work

DocQA [12,13,14,15,16] has become an important paradigm in open-domain question answering, where the retriever model first obtain relevant passages, which are then processed by the reader model to generate answers. Reader models are typically classified as either extractive or generative models, with the former pinpointing the answer span in the provided context and the latter generating answers token by token. However, current DocQA [11] mainly relies on vector-based passage retrieval, which limits the model's capabilities.

III. Method

Our framework is illustrated in Figure 1, comprising three main steps: creating knowledge base, relevant passage retrieve, and LLM generation.

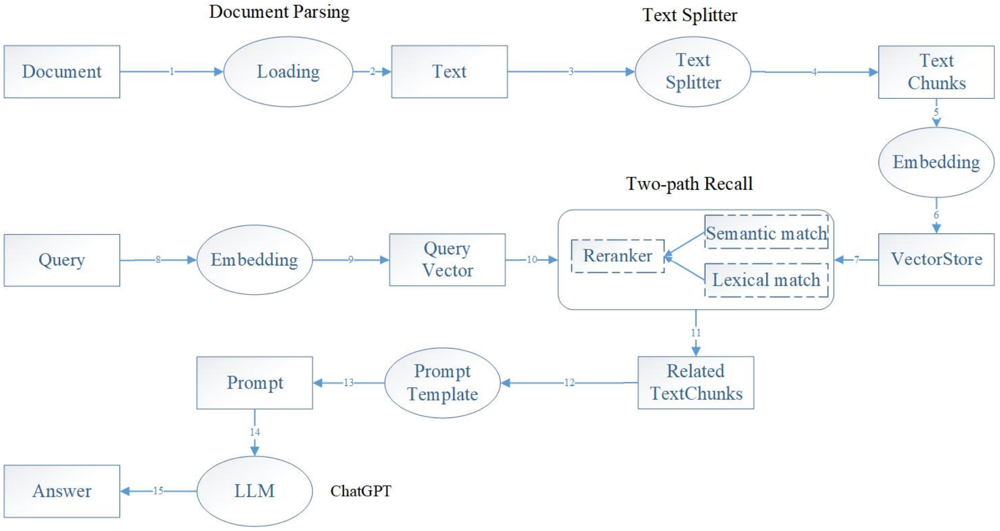

Figure 1. The framework of our method, comprising three main steps: creating knowledge base, relevant passage retrieve, and LLM generation.

A. Creating Knowledge Base

The creation of the knowledge base involves three steps: knowledge loading, knowledge segmentation, and knowledge vectorization. Generally, new data other than the training dataset is referred to as external data. The data loading phase includes operations such as loading data from documents, and processing the data into a unified format based on its characteristics. In the text segmentation phase, two main factors are considered: 1) the token limitations of the embedding model, and 2) the impact of semantic integrity on overall retrieval effectiveness. Segmenting text into fixed lengths (e.g., 256/512 tokens) results in a loss of significant semantic information. Therefore, we segment the text based on sentences to preserve the complete semantic meaning of each sentence. Common segmentation symbols include periods, exclamation points, question marks, line breaks, etc. Subsequently, context concatenation is performed within 512 words to reduce computing burden.

Vectorization is the process of converting text data into vectors, which directly influences the subsequent retrieval effectiveness. It transform data into vector representations and store them in a vector database by creating a knowledge base. The M3E [2] is the most commonly used embedding models, generally meet most requirements. Considering that this is a common scenarios, we do not further finetune embedding models. Generally, the process of indexing the vectorized data and loading it into the database can be summarized as the data entry process. In this case, we utilize the FAISS as vector knowledge base, consider its outstanding performance.

B. Relevant Passage Retrieve

The purpose of this step is to retrieve the most relevant content snippets from the knowledge base for a given query. Specifically, the retrieval module we designed consists of two main modules: dual-path coarse retrieval and precise matching retrieval. For a given query, we first utilize the effective M3E [2] model and the bm25 [17] lexical retrieval model to retrieve several candidate text blocks most relevant to the current query from the vector database. This dual-path fusion search approach provides superior retrieval results by combining two complementary search algorithms—it considers both semantic similarity between the query and stored documents and keyword matching, thus catering to retrieval needs at different levels of information granularity.

Although some candidate retrieval results are obtained through the dual-path algorithm, they may still contain irrelevant information. Due to the differing evaluation criteria of vector retrieval and keyword retrieval, precise sorting cannot be achieved through them. To address this issue, we re-rank the coarse-ranking passages using a matching model [3] to filter out irrelevant content, ensuring that subsequent large-scale models can access higher-quality knowledge to enhance generation quality. It is important to note that since a significant amount of irrelevant content has already been filtered out through dual-path coarse ranking, applying the matching model for precise sorting does not significantly increase computational overhead, thereby maintaining the efficiency of the model's operation.

C. LLM Generation

This is the final step of the whole system—generating answers based on all the retrieved contexts and the given user query. Here, we concatenate all the acquired contexts with the query statement and provide them to the LLM. By incorporating the retrieved relevant data, we enhance the capabilities of the language model. This step effectively combines prompt engineering techniques with LLM, empowering prompts to enable large language models to generate accurate answers to user queries.

IV. Experiment

A. Dataset and Evaluation

We selected 50 real documents from the financial domain, covering the current performance and expectations across various industries. We invited three senior analysts from the financial sector to construct 150 common queries in the financial domain based on their expertise, with each question being associated with a paragraph from the corresponding text document. Subsequently, these questions were annotated with standard answers by the analysts. We provide data analysis in Table 1, where the dataset contains 150 questions and 50 documents. The longest question is 67 words. In our experiments, we split the documents into 6800 snippets, each containing less than 1000 words.

In the evaluation process, we used matching accuracy and generation accuracy to assess the model's quality. Matching accuracy can be automatically calculated, considering a retrieved text snippet as correct as long as it fully contains the knowledge segment; otherwise, it is considered incorrect if it does not contain or only partially contains the segment. Generation accuracy, on the other hand, was manually annotated by the three analysts based on their expertise.

Table 1. data analysis.

dataset	statistics
question	150
document	50
passage snippet	6800
Max length of question	67
Max length of snippet	1000

B. Experiment Setting

In the table 2, we compared different models. ChatGPT represents the large model directly providing answers, DocQA(V) represents using a vector model to retrieve one knowledge snippets to enhance the capabilities of ChatGPT, which is currently the most common approach. Additionally, DocQA(V+S) introduces lexical retrieval and semantic retrieval, each retrieving one snippets to enhance the model's capabilities. For DocQA (V+S+R), we also introduced a precision ranking model using BGE-base model. DocQA (V+S+R+) indicates further domain adaptation of the BGE model using document data. Specifically, we segmented the articles into numerous fragments, utilizing ChatGPT to generate a question for each fragment, and then fine-tuned the BGE model using the questions and knowledge snippets.

Table 2. Information on video and audio files that can accompany a manuscript submission.

Model	Generation ACC	Recall ACC
ChatGPT	10.67%	None
DocQA(V)	53.33%	56%
DocQA(V+S)	62%	64.67%
DocQA(V+S+R)	73.33%	76.67%
DocQA(V+S+R+)	80.67%	84%

C. Main Result

In our experimental results shown in Table 2, it can be observed that without introducing any external knowledge, ChatGPT performs the worst due to generating a significant amount of illusions and false knowledge. With the introduction of vector retrieval, i.e., DocQA(V), the generation accuracy increased from 10.67% to 53.33%, highlighting the importance of incorporating external knowledge. Furthermore, in cases of correct retrieval, ChatGPT is likely to be correct, confirming the critical role of the retrieval stage. When both semantic and lexical retrieval are employed, i.e., DocQA(V+S), the model's performance further improves the performance to 62% and 64.67 in terms of recall and generation accuracy respectively, demonstrating the importance of dual-path retrieval in retrieving information in situations of sparse knowledge. Additionally, the introduction of a precision ranking model, i.e., DocQA (V+S+R), further enhances generation accuracy to 73.33%, showcasing the effectiveness of our approach. Finally, DocQA (V+S+R+) further finetunes the BGE model, which increase the accuracy to 80.67%, as it boost the domain adaptability. Through these experiments and comparative studies, the results indicate a significant improvement of our method on this dataset. This helps better understand the issue of knowledge illusions in large models, enhancing the model's robustness and performance. Finally, we obverse that

D. Case Study

To visually demonstrate the superiority of our model, we provide a case study in Table 3, which compares our model with mainstream vector retrieval methods. In this example, the question is "What impact does the El Niño phenomenon have on agricultural products," involving terms such as "El Niño" and "agriculture." We can observe that traditional retrieval models easily locate incorrect segments due to the presence of these terms in erroneous snippets. However, by employing a composite retrieval approach, the retrieval capability significantly improves, enabling the correct retrieval of knowledge snippets and consequently providing the accurate answer.

Table 3. Information on video and audio files that can accompany a manuscript submission.

Model	Recall passage	Answer
Recall model	According to the latest report released by the World Meteorological Organization (WMO) on May 3, this year's possibility of the occurrence of the El Niño climate phenomenon is increasing, which may bring about high temperatures and extreme weather. The WMO predicts in the report that in May to July 2023, the likelihood of transitioning from a neutral ENSO state to an El Niño event is 60%, rising to 80% from July to September. The El Niño event will lead to global climate patterns becoming unstable, resulting in frequent disastrous weather events. Typically, the El Niño phenomenon causes drought conditions in regions such as Southeast Asia and Australia, while increasing rainfall in the central equatorial Pacific region extending to the west coast of South America. For China, El Niño events are likely to cause warm winters, heavy rain and flooding in the south, hot and dry conditions in the north, and cold summers in the northeast. It is believed that extreme weather events such as droughts and floods may disrupt global food production patterns, bringing more uncertainty to international food supplies. Agriculture, Forestry, Animal Husbandry, and Fishery Industry Tracking Report | May 23, 2023. Please read the disclaimer and statement after the text. Nino 3.4 Index (°C) Data source: National Climate Prediction Center of the United States (including forecasts), CITIC Securities Research Department Note: It is generally considered that when the Nino 3.4 Index exceeds 0.5°C and persists for more than 6 months, it can be classified as an El Niño event; if it exceeds 1.5°C and persists for more than 9 months, it can be classified as a strong El Niño event.	Based on the information available, the El Niño phenomenon may have the following impacts on agricultural products: It may lead to unstable climate patterns globally, triggering frequent disastrous weather events, including extreme weather phenomena such as droughts and floods. It may cause drought conditions in regions such as Southeast Asia and Australia, while increasing rainfall in the central equatorial Pacific region extending to the west coast of South America. In China, it may lead to abnormal climate phenomena such as warm winters, heavy rain and flooding in the south, hot and dry conditions in the north, and cold summers in the northeast.
ours	The uncertainty surrounding El Niño is increasing, and the tight supply situation in the sugar market may continue. According to the World Meteorological Organization (WMO) forecast, this year's likelihood of the El Niño climate phenomenon is increasing, which may bring about high temperatures and extreme weather globally. It is believed that extreme weather events such as droughts and floods may disrupt the global food production order, bringing more uncertainty to international food supplies. According to the forecast data from the global trade consulting firm Czapp in May 2023, it is indicated that for the 2023/24 fiscal year, there still exists a gap between supply and demand for sugar globally, and the tight supply situation is expected to continue. The El Niño event may lead to drier weather in Asia this year compared to normal years, and it may result in reduced sugar production in Brazil and India. Domestic sugar supply is expected to be tight, and in the short term, domestic sugar prices may continue to rise. It is recommended that domestic sugar industry giants such as COFCO Sugar, with a business model encompassing "domestic + overseas," "processing + trading," and "futures + spot trading," and a full industry chain layout, are well positioned in the sugar industry.	The El Niño phenomenon may have the following impacts on agricultural products: It could lead to global high temperatures and extreme weather, potentially triggering extreme climate events like droughts and floods, disrupting the global food production order, and increasing uncertainty in international food supply. The Asian region may become drier, which could result in reduced sugarcane production in countries like Brazil and India. There is still a gap between supply and demand for sugar globally, and the tight supply situation is expected to continue, potentially affecting the sugar industry and prices.
Ground-truth	El Niño uncertainty is increasing, and the tight sugar supply situation may persist. According to the World Meteorological Organization (WMO), the likelihood of El Niño climate phenomenon occurring this year is increasing, potentially bringing high temperatures and extreme weather globally. We believe that extreme weather events such as droughts and floods may disrupt the global food production order, leading to more uncertainty in international food supply. Data from the global trade consulting firm Czapp in May 2023 indicates that there is still a gap between supply and demand for sugar in the 2023/24 period, and the tight supply situation is expected to continue. The El Niño event may cause drier weather in Asia this year compared to normal years and could result in reduced sugarcane production in Brazil and India. Domestic white sugar supply is expected to be tight, and domestic sugar prices may continue to rise in the short term. We recommend that domestic sugar industry leader COFCO Sugar should adopt a business model encompassing "domestic + overseas," "processing + trading," "futures + spot trading," and establish a complete industry chain layout.	Based on the information mentioned in the document, the El Niño phenomenon may lead to a tendency for global climate patterns to become unstable, potentially bringing high temperatures and extreme weather. Typically, the El Niño phenomenon leads to drought conditions in Southeast Asia, Australia, and increased rainfall in the central equatorial Pacific to the west coast of South America. For China, the El Niño event is likely to result in a mild winter, increased risk of heavy rainfall and flooding in the south, elevated risk of high temperatures and drought in the north, and the potential for a cooler summer in the northeast. In the forecast for the Yangtze River Basin, the overall precipitation is expected to be below average this year due to the influence of the El Niño phenomenon, resulting in a greater prevalence of drought over flooding. Additionally, the El Niño event may disrupt the global food production order, leading to increased uncertainty in international food supply. Therefore, it can be anticipated that the El Niño phenomenon will have a certain impact on agricultural production and supply, including the potential for drought, flooding, and extreme weather. However, the specific extent and scope of these impacts still require further observation and assessment.

V. Conclusion

This paper introduces a novel dual-path method to address the issue of knowledge illusions in large models, achieving superior results. To better evaluate this task, we also constructed a manually annotated test set for validation. In the future, we will further optimize retrieval and ranking models, and explore applications in other domains to validate the generalizability and scalability of this method.

Acknowledgement

This work is sponsored by State Grid Information & Telecommunication Branch self-managed technology project (ID: 529939230008).

References

[1] Xu Ming, Text2vec: A Tool for Text to Vector, 2022.

[2] Wang Yuxin, Sun Qingxuan, He sicheng, M3E: Moka Massive Mixed Embedding Model, 2023.

[3] J ohnson J, Douze M, Jégou H. Billion-scale similarity search with GPUs [J]. IEEE Transactions on Big Data, 2019, 7(3): 535-547.

[4] Ouyang, Long and Wu, Jeffrey and Jiang, Xu and Almeida, Diogo and Wainwright, Carroll and Mishkin, Pamela and Zhang, Chong and Agarwal, Sandhini and Slama, Katarina and Ray, Alex and others, Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems, 2022

[5] Xiao S, Liu Z, Zhang P, et al. C-pack: Packaged resources to advance general chinese embedding [J]. arXiv preprint arXiv:2309.07597, 2023.

[6] Qwen Group, Qwen Technical Report, arXiv preprint arXiv, 2023.

[7] baichuan2023baichuan2, Bichuan 2: Open Large-scale Language Models, arXiv preprint, 2023.

[8] LLaMA: Open and Efficient Foundation Language Models.

[9] Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

[10] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[11] Soong D, Sridhar S, Si H, et al. Improving accuracy of gpt-3/4 results on biomedical data using a retrieval-augmented language model[J]. arXiv preprint arXiv:2305.17116, 2023.

[12] Es S, James J, Espinosa-Anke L, et al. Ragas: Automated evaluation of retrieval augmented generation[J]. arXiv preprint arXiv:2309.15217, 2023.Dense passage retrieval for opendomain question answering.

[13] Guu K, Lee K, Tung Z, et al. Retrieval augmented language model pre-training[C]//International conference on machine learning. PMLR, 2020: 3929-3938.

[14] Jiang Z, Araki J, Ding H, et al. How can we know when language models know? on the calibration of language models for question answering [J]. Transactions of the Association for Computational Linguistics, 2021, 9: 962-977.

[15] Izacard G, Grave E. Leveraging passage retrieval with generative models for open domain question answering [J]. arXiv preprint arXiv:2007.01282, 2020.

[16] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[J]. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[17] Robertson S, Zaragoza H, Taylor M. Simple BM25 extension to multiple weighted fields[C]//Proceedings of the thirteenth ACM international conference on Information and knowledge management. 2004: 42-49.