Intelligent Question Answering System for Power Regulations based on RAG

Zhiyan Qiao
School of Electrical Engineering
Guang zhou City University of Technology
Guangzhou, Guangdong 510800, China
3232233165@qq.com

Ming Gao
School of Electrical Engineering
Guang zhou City University of Technology
Guangzhou, Guangdong 510800, China

Haixia Ma*
School of Electrical Engineering
Guang zhou City University of Technology
Guangzhou, Guangdong 510800, China
* Corresponding author: mahx@gcu.edu.cn

Xiangyu Wen
School of Electrical Engineering
Guang zhou City University of Technology
Guangzhou, Guangdong 510800, China

Erbao Yan
Guangzhou Power Supply Bureau of Guangdong Power Grid Co., Ltd
Guangzhou, Guangdong 510800, China

Wenxuan Zhao
School of Electrical Engineering
Guang zhou City University of Technology
Guangzhou, Guangdong 510800, China

Hongyue Cao
School of Electrical Engineering
Guang zhou City University of Technology
Guangzhou, Guangdong 510800, China

Hui Wu
School of Electrical Engineering
Guang zhou City University of Technology
Guangzhou, Guangdong 510800, China

2024 4th International Conference on New Energy and Power Engineering (ICNEPE) | 979-8-3315-1608-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICNEPE64067.2024.10860461

***Abstract*—The complexity and large volume of regulations in the power industry make it difficult for professionals to quickly access the information they need. To address the issues posed by the complexity and vast amount of regulations in the power industry, this paper proposes an intelligent question answering system based on the Retrieval-Augmented Generation (RAG) model. The system utilizes DeepDoc document parsing technology to convert power regulation documents into a vector database. By leveraging a multi-channel recall and fusion ranking strategy, it can swiftly locate relevant content from a massive collection of documents. The system integrates search results to generate precise natural language responses, improving the efficiency of information retrieval for professionals. Experimental results show that the RAG-based question answering system achieves an Exact Match (EM) accuracy of 84.0%, a Precision of 85.0%, a Recall of 86.0%, and an F1-score of 85.5%. Compared to the traditional Bertserini algorithm, it demonstrates significant improvements, exhibiting superior retrieval and answering capabilities.**

Keywords-RAG, Power Regulations, Intelligent Question Answering System, Natural Language Processing

I. INTRODUCTION

In the power industry, regulations and standards serve as core pillars of safe operation and compliance management. They play a critical role in the skill training of both new and experienced employees, efficient knowledge acquisition for operation and maintenance personnel, on-site guidance and decision-making assistance for field staff, and effective management for supervisors [1]. However, with the increasing complexity of these regulations and the growing number of documents, it has become an urgent problem for power industry workers to quickly and accurately find specific clauses from the vast array of power regulations, as they cannot possibly memorize all the rules and standards. Traditional manual retrieval methods are not only inefficient but also prone to omissions or errors, which can create safety hazards. In recent years, Natural Language Processing (NLP) technology has rapidly developed, and the application of intelligent question answering (QA) systems has provided new technical means for fast information retrieval and precise information extraction. Among these, the intelligent QA system based on the Retrieval-Augmented Generation (RAG) model is capable of generating coherent and accurate natural language answers, offering crucial support for precise query of regulations in the power industry.

Currently, some researchers have already conducted related studies. Literature 2 constructed a question answering system model that integrates text and knowledge graphs, which shows some performance advantages compared to other models[2]. However, the generated answers are not fluent enough, and the human-computer interaction is poor; Literature 3 focused on the intelligent auxiliary warning decision-making needs for coal mine safety, using gas over-limit coal mine safety hazard knowledge as the data source[3]. Based on RAG, they developed an intelligent QA model for coal mine safety, but the document retrieval accuracy of this model still needs improvement. Literature 4 developed an intelligent QA system for power standards by combining knowledge graphs and large models, which improved the accuracy of professional knowledge QA[4]. However, it still could not completely avoid the "hallucination" problem of large models. In summary, there are currently few

enhanced retrieval-based intelligent QA systems for power regulations, and the performance of QA systems that use knowledge graphs or other algorithms is unsatisfactory. To address this issue, this paper proposes an intelligent question answering system for power regulations based on the RAG architecture. By combining deep document parsing technology with the RAG model, this system improves RAG's retrieval accuracy and mitigates the "hallucination" problem of large models.

II. RAG ENHANCED RETRIEVAL-GENERATION MODEL

Retrieval-Augmented Generation (RAG) is a technique that harnesses information from private or proprietary data sources to facilitate text generation[5-6]. This approach integrates retrieval and generation models to boost the quality and precision of the generated text. In detail, the model initially retrieves information snippets relevant to the input query from a vast array of documents, and then uses these snippets in conjunction with the input as context to direct the generation model in crafting responses.

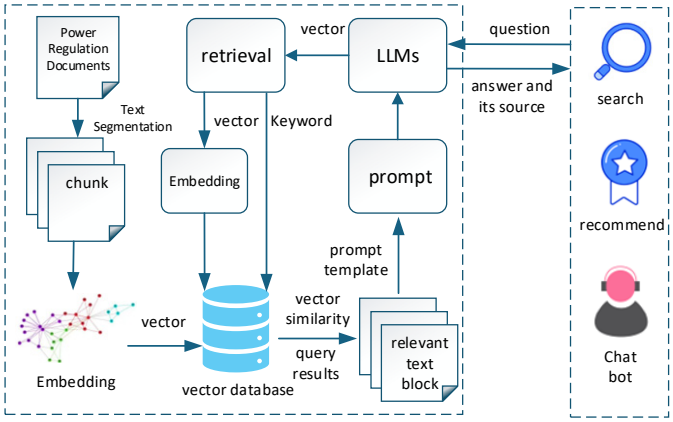

Figure 1. The logical diagram of an Intelligent Question Answering System for Power Regulations based on RAG

The RAG system represents an implementation method in which modules and algorithms are flexibly combined and arranged by developers according to specific application scenarios, rather than following a fixed structure. This workflow is referred to as the flexible implementation of the RAG system. The architecture adopted in this paper is based on the RAG framework, with modifications made to the original design. The modified system framework is illustrated in Figure 1. First, a large set of electricity regulations and guidelines is segmented into text blocks, and an embedding model is used to convert these text blocks into vectors, creating a vector database for electricity regulations. When power sector personnel submit a query via a front-end interface, such as an intelligent dialogue system, the system inputs the query into a large model, which converts the query into a vector. The system then searches the vector database for the most relevant document blocks (Top-K chunks). These retrieved document blocks, along with the original query, are then provided as prompts to a large language model (LLM) to generate the final response.

A. Vector Database Construction

The DeepDoc document parsing technology forms the core component of the Intelligent Question Answering System for Power Regulations based on RAG. DeepDoc encompasses both visual information processing and parsing technologies, enabling comprehensive analysis of complex documents and extracting various types of information, such as text, tables, and images. The visual processing mainly employs OCR (Optical Character Recognition) technology [7] to extract textual information from images or PDF files. The parser analyzes the layout structure of the document, identifying and parsing elements like paragraphs, headings, tables, and images. Additionally, the parser is capable of recognizing table structures and converting them into a data format suitable for vectorized processing.

1) Document Parsing: In the first step of building the vector database, the system uses DeepDoc technology to fully analyze power regulation documents. Through OCR and parsing capabilities, DeepDoc converts complex document content into structured data.

2) Vectorization Processing: The parsed document content is first transformed into dense vectors with up to 1024 dimensions using the Embedding-2 model. Let the query vector be denoted as q, the vector representation of a document d_i be v_i, and the embedding model as Embedding-2, which converts the text into high-dimensional vectors. The vectorization process of the document is shown in equation (1).

$$v_i = Embedding(d_i) = (title_i, content_i, keywords_i) \quad (1)$$

After vectorization, the generated vectors of text, tables, and images are stored in the Elasticsearch database using Bulk indexing, forming the vector database of the system.

B. Retrieval and Ranking

In the Intelligent Question Answering System for Power Regulations based on RAG, the retrieval and ranking module achieves efficient and accurate question-answering functionality through a multi-path retrieval and fusion ranking strategy. The multi-path retrieval strategy leverages the Elasticsearch database, combining vector and keyword recall channels along with structured data search methods to retrieve relevant documents from different perspectives, ensuring comprehensive coverage of the query content.

Based on the candidate documents generated by the multi-path retrieval, the system performs a comprehensive evaluation and ranking of the results through fusion ranking. By assigning weights to each retrieval path and calculating the weighted score for each candidate, the top-K candidate documents related to power regulations are selected and fed into the generative model (ChatGLM4) using prompt optimization. Finally, the system outputs the most relevant answer.

1) Multi-Path Retrieval

The intelligent Q&A system uses a multi-path retrieval mechanism to improve the accuracy of Q&A within large-scale power regulation document collections. Multi-path retrieval employs keyword matching and semantic vector similarity of document content to enhance retrieval precision and comprehensiveness. Semantic similarity $Sim_{vec}(q, v_i)$ is

measured by calculating the cosine similarity between the query vector q and the document vector v_i, as shown in equation (2).

$$Sim_{vec}(q, v_i) = \frac{q \cdot v_i}{\| q \| \| v_i \|} \quad (2)$$

In this equation, $q \cdot v_i$ represents the dot product of the vectors, while $\| q \|$ and $\| v_i \|$ represent the Euclidean norms of the vectors. This calculation method can quantify the semantic similarity between the query and the document, with a value range between [-1, 1]. Values closer to 1 indicate a higher degree of similarity.

The keyword matching equation is shown in equation (3), where q represents the set of keywords contained in query K_q, and d_i represents the set of terms contained in document K_i.

$$Sim_{key}(q, tokens_i) = \sum_{k \in Kq} w_k * \delta(k, K_i) \quad (3)$$

In equation (3), w_k represents the weight of the keyword k, which can be adjusted based on the importance of the keyword or its frequency in the query. The function $\delta(k, K_i)$ is an indicator function, where k if the keyword appears in the vocabulary set of document $\delta(k, K_i) = 1$, and $\delta(k, K_i) = 0$ otherwise.

2) Fusion Ranking

In the Intelligent Question Answering System for Power Regulations based on RAG, the retrieval module acquires a rich set of candidate documents related to power regulations through multi-channel recall. However, in order to distill the most relevant answers from these candidate documents, fusion ranking is required. The purpose of fusion ranking is to comprehensively evaluate the candidates returned from different retrieval paths and generate a final ranked result to ensure that the system provides the most accurate and relevant answers. Fusion ranking is achieved by performing weighted calculations of the semantic similarity between the query vector and document vector, as well as the keyword match score, ultimately generating a comprehensive document score F_i. The calculation equation is shown in (4).

$$F_i = \alpha * Sim_{vec}(q, v_i) + \beta * Sim_{key}(q, tokens_i) \quad (4)$$

$Sim_{vec}(q, v_i)$ represents the semantic similarity between the query vector and the document vector, calculated using cosine similarity or other similarity metrics.

$Sim_{key}(q, tokens_i)$ denotes the text matching score based on keywords, calculated from the frequency of keyword matches or weighted matching scores.

α and β are the weights for semantic similarity and keyword matching similarity, respectively, satisfying $\alpha + \beta = 1$.

C. Generation of Answers

In this Intelligent Question Answering System for Power Regulations based on RAG, answer generation is accomplished through the API call to the large model provided by Zhipu Qingyan, and is optimized using prompt-tuning techniques. The neural network architecture of Zhipu Qingyan consists of multiple layers of Transformer encoders and decoders [8-9]. The encoder processes the input text, extracting relevant information, while the decoder generates output text based on this information. During the generation process, the model adopts an autoregressive approach, meaning that when generating each word, it takes into account all previously generated words. This ensures the coherence and naturalness of the generated content.

III. EXPERIMENT AND RESULTS

A. Electricity Regulation Data Set

The dataset used in this study was sourced from relevant regulations in the power industry, consisting of 268 documents. These documents cover a wide range of topics related to the operation, maintenance, safety management, and emergency response of power systems, ensuring the comprehensiveness and diversity of the data. To construct the test set for this study, 10 representative documents were selected from the 268 documents and then 10 questions were extracted from each document, ultimately forming a dataset consisting of 100 questions

B. Evaluation index

The Intelligent Question Answering System for Power Regulations based on RAG adopts four indicators of accurate matching rate (EM), Precision, Recall and F1 value to measure the response effect of the algorithm[10].

1) Answer accurate match rate (EM)

EM is used to measure the accuracy of question answering. It is the percentage of the total number of questions answered by the question answering system in which the answers given by the question answering system exactly match the correct answers determined by the human. The calculation equation is shown in (5).

$$EM = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} I(\hat{y}_i = y_i) \quad (5)$$

In equation (5), $n_{samples}$ represents the total number of samples; $I(x)$ is the indicator function, which takes a value of 1 when $\hat{y}_i$ is exactly equal to y_i, and 0 otherwise. It can be observed that the larger the EM value, the higher the accuracy of the match.

2) Precision, Recall and F1 values

Recall is an important index used to calculate the recall rate of a model, which represents the proportion of correctly classified samples in all positive samples. The calculation equation is shown in equation (6). Precision represents the proportion of positive samples correctly classified by the model in all samples classified as positive. The calculation equation is shown in equation (7). F1 is the harmonic average of Precision and Recall, which is used to comprehensively evaluate the

performance of the model. The calculation equation is shown in equation (8).

$$\Pr ecision = \frac{TP}{TP + FP} \tag{6}$$

$$\mathrm{Re}\, call = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = 2 \cdot \frac{\Pr ecision \cdot \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call} \tag{8}$$

In equation (6), TP represents true cases, FP represents false positive cases, and Precision represents the proportion of positive samples correctly classified by the model in all classified positive samples. In equation (7), TP represents true cases, FP represents false negative cases, and Recall represents the proportion of samples correctly classified by the model in all positive samples. The higher the recall rate, the stronger the detection ability of the model for positive samples. In equation (8), $F1$ value is the harmonic average of accuracy rate and recall rate, which is used to comprehensively evaluate model performance.

C. Comparison of test results of intelligent question answering system

In order to test the Q&A effect of the system built in this paper, 100 questions are designed for the power regulation documents, and the initial Bertserini algorithm and Intelligent Question Answering System for Power Regulations based on RAG are respectively tested and compared with manual answers. Some test results are shown in Table 1 and Table 2. RAG system is significantly better than Bertserini in accuracy and answer traceability, especially for complex questions. Bertserini falls short on accuracy and traceability, whereas RAG not only answers accurately, but also provides document traceability.

Table 1 Comparison of test results of intelligent question answering system

Serial number	Query problem	Question-and-answer tool	Answer	Exact Match	Traceable	The correct answer is in the document
1	What devices should be installed in order to maintain a safe distance when working between a high voltage drop type fuse and a cable head?	Manual answer	Transition connection	/	/	"Electric Power Safety Operating Regulations of xxxx Power Grid Co., Ltd." (Q/CSSG 510001-2015) Section 15.2.17
		Initial Bertserini algorithm	Transition connection	Yes	No	
		Intelligent Question Answering System for Power Regulations based on RAG	Transition connection	Yes	Yes	
2	Height operation refers to the height of the fall height of how many meters and above the datum level, is likely to fall to the height of the operation?	Manual answer	2m	/	/	"Electric Power Safety Operating Regulations of xxxx Power Grid Co., Ltd." (Q/CSSG 510001-2015) Section 3.1
		Initial Bertserini algorithm	Short-range link	No	No	
		Intelligent Question Answering System for Power Regulations based on RAG	2m	Yes	Yes	
3	What measures should be taken when responsible persons, managers, site staff and related personnel at all levels find violations of the NCP Safety Regulations?	Manual answer	Stop immediately	/	/	"Electric Power Safety Operating Regulations of xxxx Power Grid Co., Ltd." (Q/CSSG 510001-2015) Section 4
		Initial Bertserini algorithm	Emergency interrupt	No	No	
		Intelligent Question Answering System for Power Regulations based on RAG	Stop immediately	Yes	Yes	
...	...	...	...	...	...	...
100	What is the "first line and three rows" principle of hidden danger investigation and management?	Manual answer	The "first line" is to adhere to the red line of safety production, and the "three rows" include the investigation, sorting, and exclusion of hidden dangers	/	/	"Safety Production Risk Grading Control and Hidden Danger Investigation and Treatment Dual Prevention Mechanism Management Guidelines of xx Power Supply Bureau" 5.1.2
		Initial Bertserini algorithm	The "first line" is the bottom line to ensure safe production, and the "three rows" refer to the identification, prioritization and complete removal of risks.	No	No	
		Intelligent Question Answering System for Power Regulations based on RAG	The "first line" is to adhere to the red line of safety production, and the "three rows" include the investigation, sorting, and exclusion of hidden dangers	Yes	Yes	

In order to evaluate the system, this paper tests the initial Bertserini algorithm and the Intelligent Question Answering System for Power Regulations based on RAG respectively, and calculates the evaluation indexes as shown in Table 2. Table 2 shows that the RAG based question answering system is obviously superior to the initial Bertserini algorithm in all indexes. The EM value of RAG system is 84.00%, which is significantly higher than the 26.1% of Bertserini algorithm.

Precision value increased from 38.25% of Bertserini algorithm to 85.00%; Recall was increased from 45.3% to 86.00%. The F1 value also increased from 33.1% to 85.50%. These data show that the RAG based Q&A system has improved significantly in accuracy and overall performance, and can better meet the actual Q&A needs.

Table 2 Evaluation Metrics Comparison for Intelligent Question Answering Systems

Test object	EM	Precision	Recall	F1
Initial Bertserini algorithm	26.1%	38.25%	45.3%	33.1 %
Intelligent Question Answering System for Power Regulations based on RAG	84.00 %	85.00%	86.00 %	85.50 %

IV. CONSTRUCTION OF THE INTELLIGENT QUESTION ANSWERING SYSTEM FOR POWER REGULATIONS BASED ON RAG

A. Construction of an Intelligent Question Answering System for Power Regulations based on RAG.

As shown in Figure 2, the Intelligent Question Answering System for Power Regulations based on RAG built by this paper using the RAG enhanced retrieval generation model can be divided into four parts: the first is the power text database, which is used to store a large number of power regulation documents after being processed by DeepDoc technology, and then stored in the vector database using embedding model; the second is user interaction, providing a intuitive visual Web front-end; the third is query and retrieval, when the user asks a question, the system will parse the question, use the combination of vector search and keyword search to find the related document fragments; the fourth is the generation of system answers, the top-K fragments retrieved by the system will be input into the generation module, and the generation module will generate the answer by prompting. This intelligent answering system uses the API service of the application of big models such as ZhiPuQingYan to call the embedded model and generation model, and processes the data uploaded by the user through the DeepDoc technology, including document parsing, OCR and vectorization, and stores the data in the vector database through the embedded model. The user inputs the question through the Web interface based on React framework, and the front-end forwards the request to the backend system for processing. The backend system uses the query parsing module to use the combination of vector search and keyword search to find the related document fragments. The top-K fragments retrieved by the system are fused and sorted, and the fused and sorted top-K fragments are input into the generation model to generate the final answer. The final answer is displayed on the user's browser.

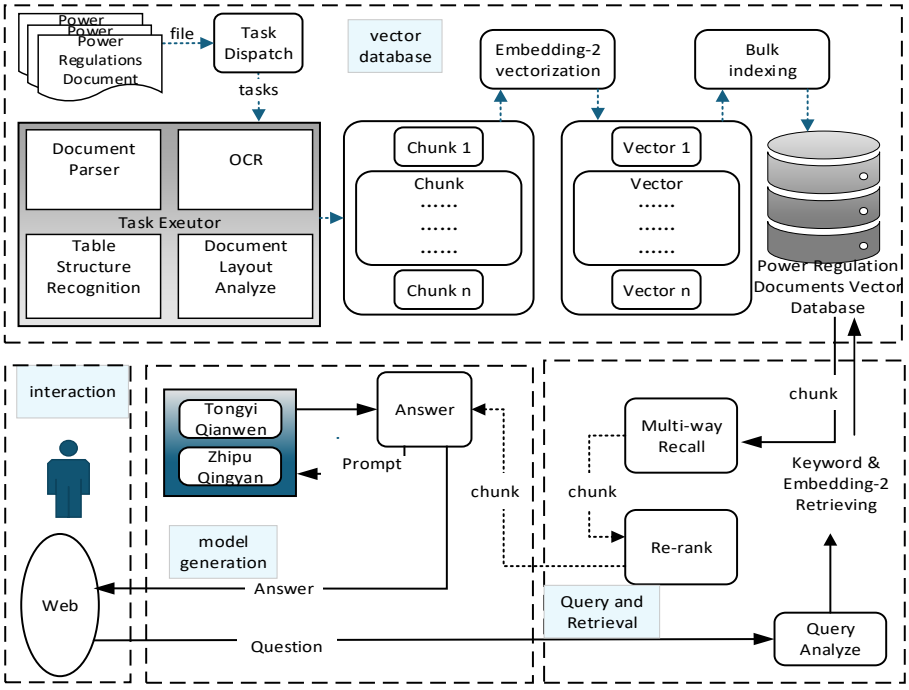

Figure 2. Intelligent Question Answering System for Power Regulations based on RAG

B. Power rules and regulations question and answer system function

1) Interactive interface of question and answer system

The interactive interface of the Q&A system designed in this paper is simple and intuitive, and the operation is smooth. Users can quickly upload documents in various formats to realize efficient analysis and question retrieval. The interface is responsive, supports real-time feedback and multilingual parsing, and significantly improves the user experience. Whether it is document management, question input or answer presentation, the whole process is smooth and natural, ensuring that users have easy access to information. The interactive interface of the intelligent question answering system is shown in Figure 3.

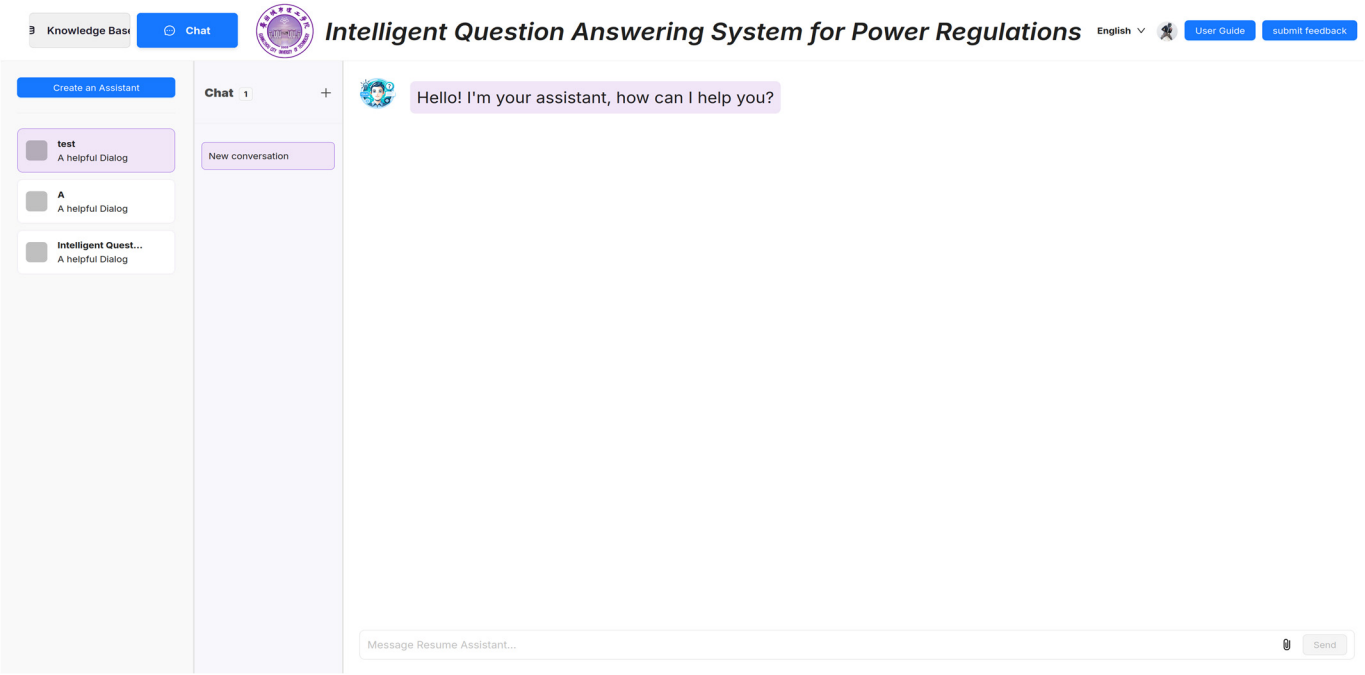

Figure 3. Front-end Interface of the Intelligent Question Answering System for Power Regulations based on RAG

2) Multiple large models are optional

The system provides flexible model selection, supports personal customization and seamless integration of open source large models, such as "Tongyi Qianwen" and other large models. Users can invoke different models to process data through API interfaces or local deployment. This variety of options can adapt to different scenarios, improve the system's processing power and scalability. The large model selection interface is shown in Figure 4.

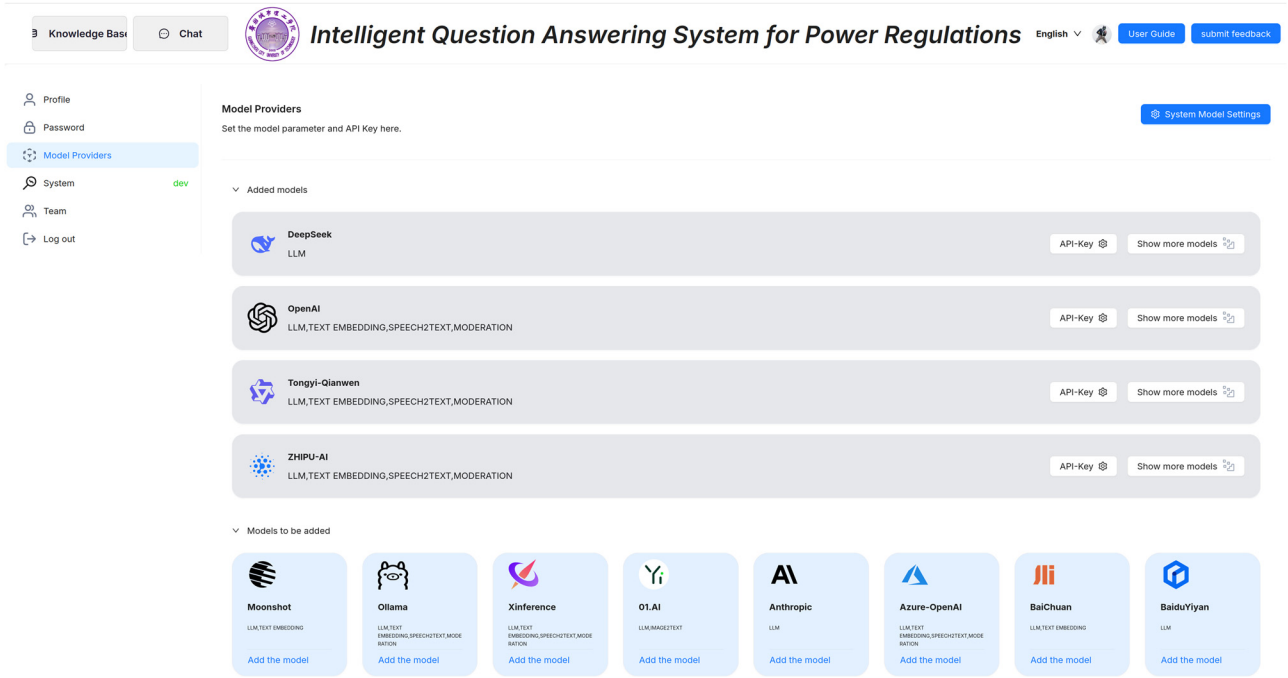

Figure 4. Big model selection

V. CONCLUSIONS

The complexity and large number of rules and regulations in the power industry have brought challenges to the information acquisition of employees. The Intelligent Question Answering System for Power Regulations based on RAG proposed in this paper realizes efficient and accurate question answering service by combining retrieval and generation technology. The system

uses DeepDoc technology to parse documents, build a vector database, and quickly locate relevant content from massive documents through multi-channel recall and fusion sorting strategies. The large model is responsible for generating natural language answers and optimizing them through prompt-tuning technology to ensure the accuracy and relevance of the answers. The experimental results show that the system is superior to the traditional method in terms of precision matching rate, precision rate, recall rate and F1 value, showing its advantages in the field of power rules and regulations. The application of the system will provide a convenient information query tool for the power industry practitioners, improve work efficiency and reduce security risks. In the follow-up work, we will continue to optimize the system performance, such as exploring more advanced retrieval algorithms, improving model training strategies, etc. At the same time, we will expand the application scope of the system and extend it to other industrial fields to provide efficient and accurate intelligent question answering services for more industries.

ACKNOWLEDGMENT

This paper is supported by the 2021 Key Research Platform Project for Ordinary Universities of Guangdong Provincial Department of Education, "Coordinated New Energy Digital Grid Technology Engineering Research Center" (2021GCZX003)

REFERENCES

[1] Chen Peng, Cai Bing, He Xiaoyong, et al. Named Entity Recognition for Power Regulations and Rules [J]. Computer Systems Applications, 2022, 31(06): 210-216.

[2] Zhang Jiahao, Huang Bo, Wang Chenming, et al. A Question Answering System Model Integrating Text and Knowledge Graph [J]. Journal of Chongqing University, 2024, 47(08): 55-64.

[3] Hong Liang, Guo Yao, Liu Xingli, et al. Intelligent Question Answering Model for Coal Mine Safety Based on RAG [J]. Journal of Heilongjiang University of Science and Technology, 2024, 34(03): 487-492.

[4] Liu Yanjuan, Zhang Dongdong, Yu Hailiang, et al. Research on Power Standard Intelligent Question Answering System Based on Knowledge Graph [J]. Electrical Engineering Technology, 2024, (16): 143-146.

[5] Hao Shibo, Shi Donghao, Tang Yuchen. Research on Patent Technology Cooperation Question Answering Application Based on Open Source RAG Architecture for University-Enterprise Collaboration [J]. Technology and Market, 2024, 31(05): 1-11.

[6] Zhong Yi, Leng Yan, Chen Sihui, et al. Research on Battery Acceleration Based on RAG Architecture of Large Language Model: Current Status and Prospects [J/OL]. Energy Storage Science and Technology, 1-11 [2024-09-03].

[7] Rubesh Kumar, T., Purnima, C. Text Extraction using OCR: A Systematic Review [C] Proceedings of the 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE, 2020: 357-362. DOI: 10.1109/ICIRCA48905.2020.9183326.

[8] Wei Yijin, Fan Jingchao. Agricultural Policy Question Answering System Based on ChatGLM2-6B [J]. Data and Computing Development Frontiers (Bilingual), 2024, 6(04): 116-127.

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I. Attention is All You Need [J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[10] Shirdel, M., Di Mauro, M., Liotta, A. Exploring Evaluation Metrics for Binary Classification in Data Analysis: The Worthiness Benchmark Concept [C]. In: Wrembel, R., Chiusano, S., Kotsis, G., Tjoa, A.M., Khalil, I. (eds) Big Data Analytics and Knowledge Discovery. DaWaK 2024, Lecture Notes in Computer Science, vol 14912. Springer, Cham, 2024: 97-112. DOI: 10.1007/978-3-031-65647-7_7.