Computer Life
ISSN: 1819-4818 | Vol. 13, No. 3, 2025

Intelligent Q&A application of rules and regulations based on large models and RAG technology

Shugang Liu [a], Yu Zhao [b]

School of Computer Science, NORTH CHINA ELECTION POWER University, Baoding 071003, China
[a] lsg69@qq.com, [b] 3108468080@qq.com

Abstract: Employees face high learning costs and inefficient retrieval due to the increasing volume and complexity of company rules and regulations. This paper proposes an intelligent Q&A system using OCR, vectorization, Large Language Models (LLM), and Retrieval-Augmented Generation (RAG). We detail the knowledge base construction, prompt engineering, and Q&A module design. Example validation demonstrates the system's accuracy, completeness, and logical output. Results show it significantly improves employee learning efficiency, offering an innovative solution for rules & regulations informatization.

Keywords: Large Language Model (LLM); Retrieval-augmented Generation (RAG); Optical Character Recognition (OCR); Intelligent question answering system.

1. Introduction

In order to address the problems of high learning costs and low retrieval efficiency caused by the increasing number and complexity of enterprise rules and regulations, this article focuses on using artificial intelligence technology to provide solutions, and explores and implements an intelligent Q&A system for rules and regulations based on Large Language Model (LLM) and Retrieval Augmentation Generation (RAG) technology[1]. The system provides accurate and efficient rule and regulation query services for enterprise employees through core functional modules such as OCR recognition, knowledge base construction (text segmentation and vectorization), intelligent retrieval and prompt engineering enhancement, and large model question and answer generation. This article provides a detailed introduction to the system's requirements background, core technology methods (LLM and RAG), overall architecture design (knowledge layer, model layer, service layer, application layer), and application instance verification.

2. Research Status and Significance

In the context of increasingly strict compliance supervision within the company, the number of corporate rules and regulations is increasing and becoming more and more complex, resulting in higher learning costs and inefficient retrieval for employees and managers. Therefore, the use of modern science and technology to build an intelligent question and answer system has become an important direction of enterprise information construction.

In recent years, with the rapid development of artificial intelligence technology, LLM and RAG have provided new technical support for intelligent Q&A in knowledge-intensive fields. LLM has strong semantic understanding and generation capabilities, which can generate accurate answers based on user input, combined with RAG technology, to improve the accuracy and interpretability of answers[2].

LLM-based intelligent Q&A technology has made significant progress in many fields. In the field of geological investigation, researchers have developed intelligent question-and-answer robots using RAG technology to improve the efficiency of geotechnical engineering professional knowledge retrieval. In the management of enterprise scientific research projects, the intelligent question and answer system has been applied to the whole life cycle of scientific research management, which has improved the scientific research productivity of enterprises. The application of LLM technology provides a reference for the intelligent Q&A of rules and regulations, and provides valuable experience for its technical route and implementation path.

This article will focus on the construction of an intelligent Q&A system for enterprise rules and regulations[3]. Based on OCR technology, the system extracts and preprocesses the texts of relevant rules and regulations, builds a knowledge base through text vectorization technology, and provides users with efficient and accurate intelligent Q&A services with the help of LLM and RAG technologies.

3. Rules and Regulations Intelligent Q&A Methods and Techniques

3.1. Large Language Models (LLM)

LLM is an important breakthrough in the field of natural language processing in recent years, which is based on deep neural networks and has powerful text understanding and generation capabilities through the training of massive text data.

LLMs not only excel in language comprehension and generation, but are also capable of multi-task learning to suit different application scenarios[4]. In the intelligent question answering system, the LLM analyzes the user's query intention, provides accurate responses, and is able to handle multiple rounds of dialogue to achieve natural language interaction with the user. In addition, LLMs have the ability to continuously learn and continuously optimize their performance and improve service quality based on feedback from real-world applications.

3.2. Retrieval enhancement generation.

RAG is an information retrieval framework designed to provide LLMs with information from external knowledge sources[5]. The technical route of intelligent Q&A of rules and regulations includes the construction of knowledge base of rules and regulations, knowledge retrieval and enhancement, and intelligent Q&A.

1). Construction of knowledge base of rules and regulations

Text Preprocessing: Users can upload various knowledge files containing rules and regulations (such as PDF, Word, images, etc.). For image files, the system integrates optical character recognition (OCR) technology to automatically extract text information from the image and convert it into a processable text format. Subsequently, the system performs unified preprocessing on all files, including but not limited to format standardization, irrelevant character filtering, text encoding unification, and basic cleaning, ultimately generating standard file objects with clear structure and standardized format, providing high-quality input for subsequent processing stages.

Text Segmentation: After completing text preprocessing, the system intelligently segments standardized file objects based on semantic coherence and structural features. By combining paragraph delimiter recognition, semantic boundary judgment, and contextual analysis, the text is divided into several paragraphs with independent semantics. This process ensures that the segmentation results maintain the logical integrity of the original text while adapting to the needs of knowledge retrieval and management, laying the foundation for paragraph level vectorization processing.

Vectorization And Index Construction: The system uses natural language processing models to vectorize the segmented text paragraphs, mapping the text semantics into high-dimensional numerical vectors. On this basis, an efficient knowledge indexing structure is constructed through a vector database to achieve fast matching and retrieval of paragraph semantics. Ultimately, a regulatory knowledge base that supports intelligent queries and similarity retrieval will be formed, providing core capability support for subsequent knowledge Q&A, compliance review, and other scenarios.

2). Knowledge retrieval and enhancement

Problem vectorization: Transform the original questions raised by users in natural language into high-dimensional dense vectors through a semantic encoder compatible with large language models. This process enables a deep understanding of the semantic core of user questions, rather than simple keyword matching, laying the foundation for accurate semantic retrieval in the future. The system ensures that the vectorized model is consistent with the model used in building the knowledge base to ensure consistency in the vector space, thereby improving the relevance and accuracy of retrieval.

Retrieval matching: Using similarity matching algorithm, quickly compare and calculate the similarity between the "problem vector" and all paragraph vectors in the "rule and regulation knowledge base". The system will accurately retrieve the K text contents with the highest semantic relevance to the current problem from the knowledge base (such as K=3 or 5), and sort them according to their similarity scores. The core purpose of this step is to quickly identify several key paragraphs that are most likely to contain answers from a vast amount of knowledge, ensuring high relevance of the recalled content.

Hint Template Embedding: Structurally assemble the K key text contents retrieved (as reference context) with the user's original question using a pre-set and carefully optimized prompt engineering template. This template typically clearly defines the role, task background, and response criteria of AI, and combines context and questions with clear instructions (such as "Please answer questions according to the following rules and regulations:") to form an enhanced prompt rich in contextual information and clear task instructions, which is ultimately submitted to the Large Language Model (LLM) for deep understanding and content generation.

3). Intelligent Q&A

Submit the embedded prompt word instructions and relevant contextual information as input to the Large Language Model (LLM). LLM, based on its profound parameterization knowledge and powerful semantic understanding ability, performs deep inference and contextual fusion on instructions and background information, generating accurate, reliable, and standardized final answers. This process effectively combines the accuracy of retrieved information with the generalization ability of LLM, ensuring that the response has both precision and good readability.

4. Design of the system

4.1. Overall architecture

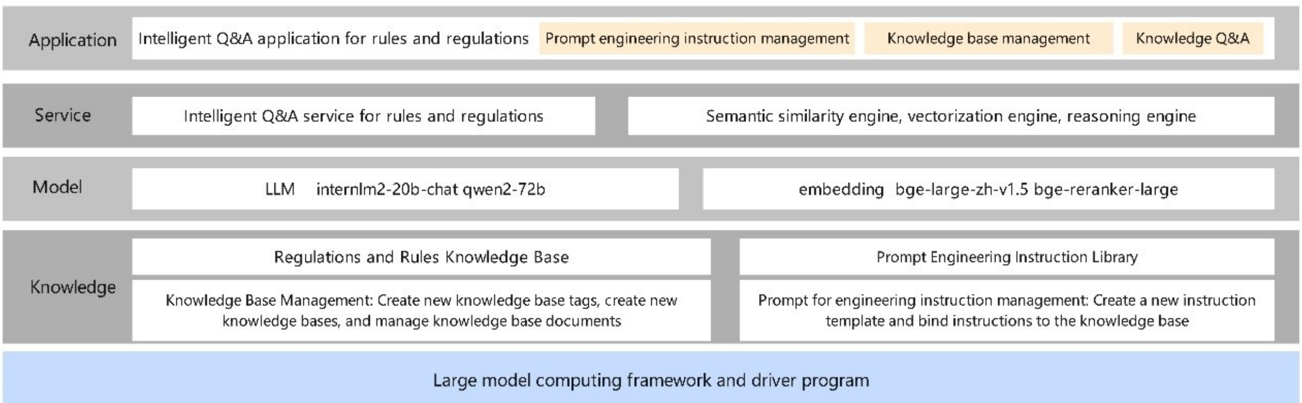

Figure 1. The overall architecture of intelligent Q&A

The knowledge layer realizes the management of knowledge and instructions through the knowledge base and the prompt engineering instruction base[6]; The model layer integrates LLM and vectorization models to provide semantic understanding, similarity calculation, and Q&A reasoning capabilities. The service layer is responsible for providing standardized API interfaces, intelligent Q&A, and general services. The application layer provides users with a convenient and intuitive experience through an interactive Q&A window, instruction management, and knowledge base maintenance interface. The architecture combines advanced technologies to support accurate Q&A and knowledge traceability, and meet the needs of intelligent Q&A of rules and regulations.

The overall architecture of intelligent Q&A is shown in the figure 1.

4.2. Technical architecture

The technical architecture of intelligent Q&A consists of an infrastructure layer, a data storage layer, a model layer, an inference and service deployment layer, a network layer, a presentation layer, and an application layer. The infrastructure layer provides computing power and resource management support through CPUs, GPUs, Docker, and private clouds; The data storage layer[7] uses MySQL, Milvus, and MinIO to achieve efficient storage and access of structured, high-dimensional vectors and data. The model layer integrates vectorized models and LLMs, supports text vectorization and intelligent question answering functions; The inference and service deployment layer relies on Transformers, vLLM, and Xinference to achieve efficient deployment and inference of models. The gateway layer uses API gateway and authentication mechanisms to ensure request management and resource security. The presentation layer uses HTML5, CSS3, Vue and Element-Plus to build an intuitive user interface. The application layer provides intelligent Q&A, prompt project management, and knowledge base management functions.

4.3. Functional architecture

According to the technical route, the rules and regulations intelligent question and answer system is divided into two parts: front-end and back-end, the front-end function is oriented to the user's application stage, and the relevant questions of the rules and regulations are input and answered intelligently in the form of an interactive window; The main functions of the back-end include knowledge base management, instruction management, and system management.

Knowledge base management includes creating knowledge base topic tags, creating new knowledge bases, and managing documents in knowledge bases. Instruction management includes creating a new instruction template, managing instruction versions, and binding instructions to the knowledge base.

System management includes user management, role management, system journal, etc.

5. Application example validation

This article verifies this article with specific application examples. Instance verification Based on Intel 26-core CPU, 512G memory high-performance computing power, equipped with NVIDIA A40 graphics card, CUDA version 12.4, operating system Centos 7.6, Python version 3.8. Among them, Internlm2-20B-chat is used for LLM, Bge-large-zh-v1.5 is used for vectorization model, Milvus is selected for vector database, and Minio is selected for file storage.

5.1. Construction of knowledge base of rules and regulations

According to the characteristics of the knowledge base of rules and regulations, the knowledge base of rules and regulations is classified and graded, and it is specifically divided into knowledge of rules and regulations, knowledge of responsibility regulations, knowledge of organization and management, and knowledge of discipline and regulations.

5.2. Rules and regulations, intelligent Q&A, prompt words, and instruction construction

The prompt word template is based on the CO-STAR framework and the following is shown in the business scenario of rules and regulations.

Context (Background)

Please refer to the following as known information.

Text content:<Document insertion point>

User issue:<Issue insertion point>

Objective

Clear and explicit resolution of user issues.

Use concise and professional language, otherwise avoid redundant information.

Avoiding direct copying of text requires language organization and structural optimization.

Structure

Please output using Markdown syntax!

Audience (audience)

The main audience is users who wish to gain a deeper understanding of relevant rules and regulations.

Relevance

Ensure that the answer is highly relevant and avoid introducing irrelevant information.

5.3. Q&A validation

In this section, a set of validation schemes is designed to verify the performance of the system in the following dimensions.

Accuracy (40%): Whether the answer content highly matches the user's question.

Integrity (30%): Answer whether the content comprehensively covers the relevant information in the knowledge base.

Logical (30%): Answer whether the logic is clear and organized, and whether it can provide users with clear knowledge points.

Validation datasets and scenarios: A total of 50 test questions covering typical scenarios were selected, covering rules and regulations, responsibilities, organizational management, discipline and other fields. There are three types of problems: simple query, medium complexity and multi-dimensional problem.

Validation process: Test questions are grouped and fed into the Q&A system. Based on the input questions, the system searches for the most similar content from the knowledge base. RAG technology is used to combine search results with a prompt word template to generate responses. Five experts were invited to rate the responses output by the system (out of 10 points), and the overall score was calculated using a weighted average.

The final score is the average score of each expert on the 50 questions, with the five experts scoring 8.5, 8.8, 9.0, 8.3 and 8.9 respectively. Judging from the evaluation results, the overall score is 8.7 points. Through the verification of examples, the intelligent question and answer system of rules and regulations designed in this paper can meet the requirements.

6. Epilogue

In this paper, an intelligent question and answer system for rules and regulations based on LLM and RAG technology is studied and implemented, and a knowledge base of rules and regulations is systematically constructed through OCR and vectorization technology. Through LLM and RAG technology, efficient, accurate, and explainable Q&A services are realized. The case verification shows that the system performs well in the accuracy, completeness and logic of the answers, and provides strong technical support for enterprise employees to learn rules and regulations. In the future, the research will further optimize the performance of the model, expand the application scenarios, and contribute more to the development of enterprise informatization and intelligence.

References

[1] Guan Dianxi, Huang Kun, Cui Nianzhi Research on Survey Geotechnical Q&A Robot Based on Large Model, RAG, and Intelligent Agent Technology [J]. China Survey and Design, 2024 (8): 101-104.

[2] Liu Qian, Yang Xiutong Research on the Application of Large Model Technology in Enterprise Research Project Management: Taking Intelligent Question Answering as an Example [J]. Project Management Technology, 2024, 22 (9): 5-11.

[3] Yan Xiaotong, Tang Xiaobin, Shen Tong, etc Overview of the Development of Large Language Models [J]. Journal of Statistics, 2024, 5 (4): 13-18.

[4] VIDIVELLI S,RAMACHANDRAN M,DHARUNBALAJI A. Efficiency-driven custom chatbot development:Unleashing langchain, RAG, and performance-optimized LLM fusion[J].Computers, Materials & Continua, 2024, 80(2): 2423-2442.

[5] WANG G ,HE J ,LI H , et al. RAG-leaks: difficulty-calibrated membership inference attacks on retrieval-augmented generation [J].Science China(Information Sciences), 2025, 68(06):23-40.

[6] LEIYB, CAOY, ZHOUTY, etal.Corpus-steered query expansion with large language models [DB/OL]. https: // arxiv.org / abs/2402.18031.2024-2-28.

[7] LEWISP, REREZE, PIKTUSA, etal. Retrieval-augmented generation for knowledge-intensive NLP Tasks [DB/OL]. https://arxiv.org/abs/2005.11401.2020-05-22.