2024 2nd International Conference on Computer Network Technology and Electronic and Information Engineering (CNTEIE) | 979-8-3315-2443-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/CNTEIE66268.2024.00032

Application and evaluation of RAG technology in civil aviation policy question answering

Yinhui Luo[a], Wenhao Xu[b], Changchang Zeng[c*], Qiang Fu[d], Ziyu Xiang[e]
Civil Aviation Flight University of China, Guanghan, China
[a]luoyinhui@cafuc.edu.cn, [b]1410657527@qq.com, [c*]Corresponding author: zengchangchang@cafuc.edu.cn, [d]cafczdm@163.com, [e]978632700@qq.com

***Abstract*—Civil aviation policies are crucial for practitioners, but the current method of manually searching for relevant policies is cumbersome and inefficient. Existing question-answering systems based on knowledge graphs or large language models are not ideal due to the lack of high-quality policy question-answering datasets and the inability to update in real-time. To this end, this paper builds a civil aviation policy question-answering system based on Retrieval-Augmented Generation (RAG). The system obtains civil aviation policy-related knowledge from multiple data sources and segments it, uses ZhipuEmbedding-3 and Chroma to create a civil aviation policy text block vector library, combines vector retrieval technology to complete the retrieval of Query text blocks, obtains the context most relevant to the Query, and then combines the prompt word function of the large language model to integrate the Query and context information in real time to generate accurate answers. Finally, we use the Ragas evaluation framework to conduct a comprehensive evaluation of the system. The results show that Qwen-max has the best overall performance in the indicators related to Faithfulness, Answer Relevancy, and Context Recall, which are 0.89, 0.75, and 0.90, respectively. The reliability and effectiveness of the system have been verified.**

Keywords: Retrieval-Augmented Generation; Large Language Model; LangChain; Prompt Engineering

I. Introduction

In recent years, with the rapid development of the civil aviation industry, the Civil Aviation Administration has released a series of policies to ensure aviation safety, safeguard the rights of passengers, and improve service quality. Therefore, civil aviation practitioners need to be aware of current policies in a timely manner. In the early days, obtaining policy information mainly relied on manual review of documents, which was time-consuming, labor-intensive and inefficient.

To solve this problem, current solutions are mainly divided into two categories: one is to build a Q&A system by combining knowledge graphs[1], although this method is structured to process domain-specific knowledge to improve Q&A accuracy, it is still less applied in civil aviation field, meanwhile, the users often need to rely on the traditional search engine, and the quality of the construction of knowledge graphs as well as the maintenance of the post sequence will affect the system's results. The other is to use a general large model for question answering directly[2]. However, the accuracy of large general models in specific fields is limited, mainly due to the lack of high-quality civil aviation policy question-answering data sets and real-time update capabilities. In addition, the deployment and training costs of such systems are high, and they have strict hardware requirements [3].

The research presented above offers different solutions for quickly and accurately retrieving relevant policy documents, achieving good results. However, there are still some shortcomings. To address these issues, based on a retrieval-augmented generation (RAG) method, this paper uses external knowledge sources, such as the official website of the Civil Aviation Administration of China to construct a civil aviation policy question-answering system through vector retrieval, prompt engineering, and large-scale language model generation.

Our Contribution:

i) We have developed a civil aviation policy question and answer system based on RAG to improve the system's professionalism and accuracy in the field of policy question.

ii) We implemented an experimental evaluation process for the policy field knowledge question answering system, using three major evaluation indicators to comprehensively evaluate the system performance to ensure that the evaluation results are comprehensive and in-depth.

II. Related Work

Although RAG [4] technology was proposed relatively late, it has indeed developed rapidly in recent years, and its development trajectory in the era of large models shows several obvious stage characteristics. Initially, the emergence of the Transformer [5] architecture provided a direction for RAG, focusing on incorporating additional knowledge bases through pre-trained language models [6]. Therefore, during this period, the research focused on how to make the pre-training effect better. However, since the data cannot be updated in real-time, the impact of RAG in vertical fields is not ideal.

Subsequently, the emergence of ChatGPT pointed out a new direction [7]. Since LLM has a very powerful contextual learning ability, RAG research has gradually shifted to providing LLM with richer information in the reasoning stage to cope with more complex and knowledge-intensive tasks, thus promoting the rapid development of RAG research. However, as the research deepens, the enhancement of RAG is no longer limited to the reasoning stage. it is becoming more closely integrated with LLM technology, especially in the retrieval of additional knowledge bases. This combination enables RAG to dynamically extract relevant information from external knowledge bases when generating answers, thereby improving the model's understanding and ability to respond to problems in specific fields [8].

979-8-3315-2443-2/24/$31.00 ©2024 IEEE
DOI 10.1109/CNTEIE66268.2024.00032

This progress not only broadens the scope of knowledge that LLM can acquire, but also enables it to handle more complex and specialized tasks [9]. Meanwhile, to facilitate the development of language model-driven applications, the LangChain framework has emerged [10]. The core functionalities of this framework include calling language models, integrating different data sources, and interacting with the operating environment. LangChain consists of six main modules: agents, chains, indexes, memory storage, models, and prompt engineering. By integrating these components, LangChain not only enhances the construction efficiency of civil aviation policy Q&A systems but also significantly improves their performance and reliability. This integrated approach allows policy Q&A systems to better serve civil aviation professionals by providing accurate and timely technical support and decision-making assistance, thereby promoting the development of information technology in civil aviation.

III. System Architecture Analysis

A. System Structure

The question-answering system constructed in this article is mainly developed based on LangChain, and the basic process is shown in Figure 1.

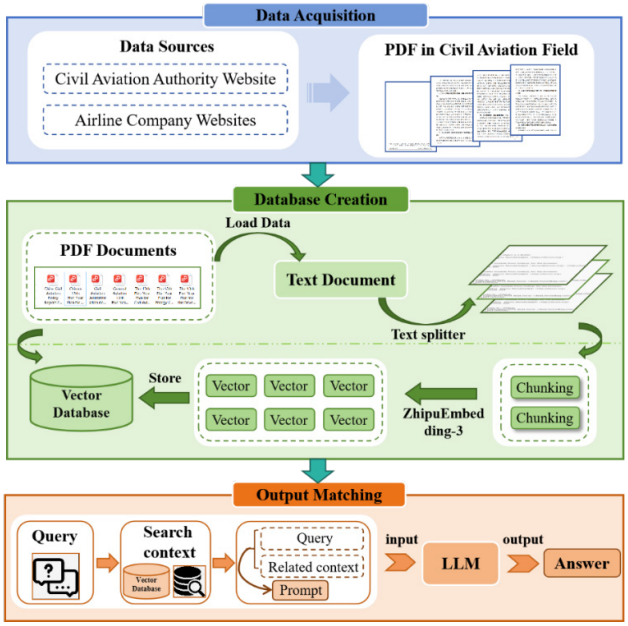

Figure 1: Policy Question and Answer System Process Framework

The general process is as follows: First, obtain relevant PDF documents in the field, build a knowledge base from them, and perform text segmentation to divide a large number of foundational documents into appropriately sized text blocks for subsequent indexing. Next, generate the knowledge base embedding index by converting the segmented knowledge pieces into high-dimensional vectors using an embedding model. These vectors are associated with their corresponding text segments and stored in a vector database. Knowledge retrieval is then performed, the query question entered by the user is vectorized, and a retriever is used to rank and extract relevant text segments that are most similar to the question. Finally, by integrating prompt engineering, combine the retrieved relevant context segments with the user's input question and input them into the LLM using a custom prompt template to obtain answers tailored to the specific context.

B. Selection of large language models

Currently, mainstream LLMs can be divided into two categories: Closed-source models led by ChatGPT [11] and open-source models represented by LLaMA [12]. Among closed-source large models, OpenAI's ChatGPT-4 often ranks at the top of various evaluation leaderboards. In the realm of open-source large models, notable examples include Qwen from Alibaba [13], and LLaMA from Meta, with specific parameter scales and descriptions provided in Table 1. This paper requires the participation of large models in two different stages. Considering that all documents processed by this system are Chinese corpus, we selected a large model optimized based on the Chinese context, so that it can more accurately understand the complex Chinese context and grammatical structure. To improve the accuracy and practicality of generated answers. In the dataset generation stage, we use the Kimi large model, using its powerful reasoning ability, combined with the provided context, to develop the expected question and Ground-truth for the subsequent RAG system evaluation. In the RAG generation stage, the large language model needs to answer the question through its powerful language understanding and generation capabilities, combined with the retrieved information, so we selected Qwen and ERNIESpeed, compared and analyzed these models, and finally selected the optimal model as the generation model of our system.

Table 1: Open-Source And Closed-Source Foundation Models

Model Name	Publishing Organization	Release time	Operating language
Llama	Meta	2023-02	English
Gpt-3.5-turbo	Openai	2023-08	English
Qwen	Aliyun	2023-08	Chinese
ERNIESpeed	Baidu	2024-03	Chinese

IV. Experimental Process And Analysis

A. Experimental environment

The environment configuration is shown in Table 2.

Table 2: Experimental Environment Parameters

project	Specific configuration
Operating system	Linux
GPU	Tesla T4
Programming language	Python .3.10
LangChain version	0.2.16
RAGAs version	0.1.21

B. Local vector knowledge base construction

At this stage, we prepare offline data. Since policy updates are relatively slow, we manually retrieve civil aviation policy-related documents from the official website of the Civil Aviation Administration of China, vectorize them, create indexes, and store them in the database. The process is divided

into four steps: document loading, segmentation, vectorization, and data storage.

First, we specify the knowledge base for civil aviation policy documents and import documents through a loader that supports multiple formats, such as pdf or txt. Next, we use a text segmentation tool (Text Splitters) to segment the loaded documents, employing recursive character segmentation technology to break long texts into smaller blocks. This improves processing efficiency while ensuring that semantic information remains unaffected. Then we use the most advanced text vector model ZhipuEmbedding-3 for vectorization, whose vector dimension can reach 2048. This model is trained with a special Chinese corpus, which can better understand the syntactic structure and semantic relationship in Chinese, improve the ability to process Chinese text, and thus improve the accuracy of similarity matching. Finally, these vectors will be stored in the open-source vector database Chroma, where indexes will be automatically created to ensure the effective storage, retrieval, and operation of high-dimensional vector data, laying the foundation for the excellent performance of the subsequent question-answering system.

C. Evaluation dataset construction

To comprehensively and accurately evaluate the RAG-based civil aviation policy Q&A system designed in this paper, we adopted RAGAs [14] for a comprehensive evaluation. The inputs required by this evaluation framework include the user's input question, the answer generated by the Q&A system, the contexts retrieved by the system related to the question, and the Ground-truth answer. The evaluation dataset is shown in table 3.

TABLE 3: SAMPLE DATASET

Label	Context
Query	How do airlines ensure cabin order during flights?
Contexts	Good cabin order is fundamental to ensuring cabin safety. Airlines should implement comprehensive management across the entire service chain, addressing bottlenecks in various service stages and striving to resolve common issues that passengers may encounter during their journeys, thereby minimizing the chances of passengers boarding with negative emotions. It is essential to establish collaborative response plans and procedures for the flight crew, cabin crew, and in-flight security team while strengthening specialized training and joint drills for personnel in key positions to effectively enhance the response capabilities of frontline staff.
Answer	Airlines need to implement comprehensive management to address bottlenecks in the service chain and reduce negative emotions among passengers. Additionally, it is important to strengthen the training and collaboration of personnel in key positions to ensure the maintenance of cabin order.
Ground-truth	Each airline should implement comprehensive management to address bottlenecks in the service chain, avoid passengers boarding with negative emotions, and enhance the response capabilities of frontline staff through specialized training and joint drills to ensure cabin order.

Considering the lack of a specialized question-related dataset for this task, we will obtain them from the documents, but manually creating a large number of question-answer pairs from documents is tedious and time-consuming, and the generated questions may lack the complexity needed for a comprehensive evaluation, ultimately affecting the quality of the assessment. Therefore, we directly use Kimi to extract various issues encountered in the production process from civil aviation-related documents and answer them. The construction process is shown in the Figure 2, and ultimately, we obtain a dataset containing queries and corresponding Ground-truths. Additionally, we also used manual methods to generate some datasets, with automatically generated data accounting for 90% and manually generated data accounting for 10%. Finally, we invited three civil aviation professionals to evaluate the generated Ground-truth, discarding some of the lower-rated ones. The remaining high-quality questions were input into the established civil aviation policy question-and-answer system to retrieve the context most relevant to the question. Finally, we combined the large language model to generate the corresponding answers and manually organized them into the JSON format required by RAGA. In the end, we obtained more than 400 evaluation data.

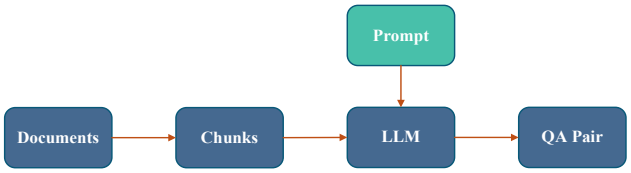

Figure 2: Dataset construction process

D. Evaluation Metrics

We used the three-evaluation metrics provided by RAGAs to assess the answers generated by different large languages models.

1) Faithfulness measures the extent to which the answer is faithful to the contexts, primarily calculated based on the answer and the contexts. The Faithfulness Score ranges from 0 to 1, where a higher value indicates better fidelity. Its calculation formula is shown in Equation 1.

$$F = \frac{|V|}{|S|} \tag{1}$$

Where S represents the number of key points extracted from the answer by the LLM, V represents the number of those key points that can be inferred from the contexts.

2) Answer Relevancy measures the degree of association between the user's question and the answer generated by the RAG system. Its calculation formula is shown in Equation 2. The score ranges from 0 to 1, with a higher value indicating a greater degree of relevance.

$$A_R = \frac{1}{n}\sum_{i=1}^{n} \mathrm{sim}(\mathbf{q}, \mathbf{q}_i) \tag{2}$$

Where $\mathbf{q}$ represents the original question, $\mathbf{q}_i$ represents the question generated by prompting the LLM based on that

answer, sim ($\mathbf{q}$, $\mathbf{q}_i$) represents the cosine similarity between the original question $\mathbf{q}$ and the generated question $\mathbf{q}_i$.

3) Context Recall measures the extent to which the retrieved document contexts contain the information necessary for the truths. The score ranges from 0 to 1, with values closer to 1 indicating better performance. Its calculation formula is shown in Equation 3:

$$C_R = \frac{|a|}{|b|} \tag{3}$$

Where a represents the number of key points identified by the LLM that can be found in the contexts, b represents the total number of key points extracted from Ground-truths using the LLM.

E. Experimental Results

In this study, we employed the three dimensions of Faithfulness, Answer Relevancy, and Context Recall from the Ragas evaluation framework to comprehensively evaluate the model output. Specifically, we input the Query, Contexts, Answer, and Ground-truth from the constructed JSON-format dataset into the Ragas framework and systematically analyzed the evaluation results (see Table 4). To further illustrate the performance differences across models in these dimensions, we also presented the results visually in Figure 3.

TABLE 4: RAG EVALUATION RESULTS

Model	Faithfulness	Answer Relevance	Context Recall
Qwen-max	0.89	0.75	0.90
Qweb-turbo	0.77	0.58	0.88
ERNIESpeed	0.84	0.68	0.91

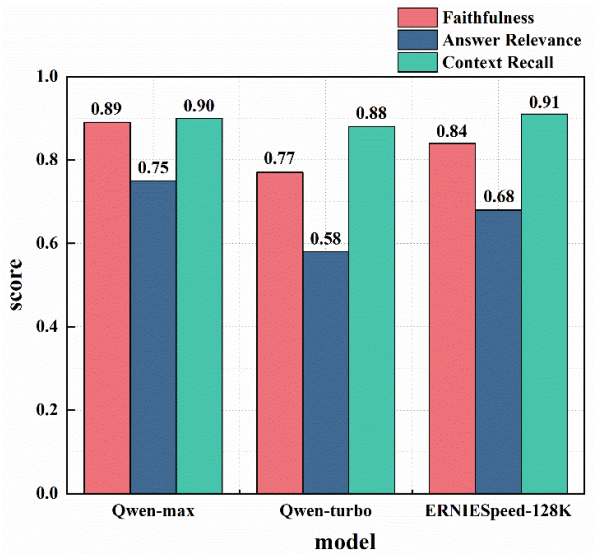

Figure 3: RAG evaluation results

The evaluation results reveal that the Qwen-max model demonstrated the best performance, achieving scores of 0.89, 0.75, and 0.90 on the dimensions of Faithfulness, Answer Relevancy, and Context Recall, respectively. In comparison, the ERNIESpeed-128K model scored 0.84, 0.68, and 0.91 on the same metrics. While its performance is acceptable, it remains slightly inferior to that of the Qwen-max model overall. In contrast, the Qwen-turbo model exhibited significantly lower scores of 0.77, 0.58, and 0.88, highlighting substantial room for improvement.

Therefore, this system uses the Qwen-max model as the generation module of the system to support the implementation of the civil aviation policy question-answering system.

V. CONCLUSION

This paper constructs a question-and-answer system specifically targeting aviation policy, enabling aviation professionals to more easily access the most accurate information related to aviation policies. Compared to traditional knowledge graphs or general large models, the use of RAG technology has significantly improved the accuracy of the generated answers. Additionally, the evaluation dataset used in this paper is comprehensive and sourced directly from relevant domain documents, effectively addressing the issues of limited and low-quality datasets in the aviation sector. This approach is also applicable to RAG systems in other fields, demonstrating high transferability.

However, building a more reliable aviation policy Q&A system still faces many challenges, particularly due to the increased complexity of language understanding in the Chinese language, which makes the model more difficult to comprehend. Furthermore, the development of multimodal Q&A systems is an important area for future work, as user inputs may not be limited to natural language text. Therefore, in future research, we will continue to optimize RAG to further enhance the accuracy and usability of the Q&A system, ensuring it better meets user needs.

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisors, Professor Yinhui Luo and Dr. Changchang Zeng, for their guidance and support. This research was funded in part by the Civil Aviation Administration of China Flight Technology and Flight Safety Key Laboratory Project: No. FZ2022ZZ01; in part by the Fundamental Research Funds for the Central Universities: No. PHD2023-028; in part by the Civil Aviation Administration of China Flight Technology and Flight Safety Key Laboratory Project: No. FZ2022KF03.

REFERENCES

[1] Huang, X., Zhang, J., Li, D., and Li, P. 2019. Knowledge graph embedding based question answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 105–113. DOI: https://doi.org/10.1145/3289600.3290956.

[2] Yunxiang, L., Zihan, L., Kai, Z., and others. 2023. Chatdoctor: A medical chat model fine-tuned on LLaMA model using medical domain knowledge. *arXiv preprint* arXiv:2303.14070, 2(5), 6.

[3] Sakib, M. N., Islam, M. A., Pathak, R., Arifin, M. M., and others. 2024. Risks, causes, and mitigations of widespread deployments of large language models (LLMs): A survey. *arXiv preprint*, arXiv:2408.04643. Available at: https://arxiv.org/abs/2408.04643.

[4] Lewis, P., Perez, E., Piktus, A., and others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.

[5] Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

[6] Radford, A. 2018. Improving language understanding by generative pre-training.

[7] Zhao, W. X., Zhou, K., Li, J., and others. 2023. A survey of large language models. arXiv:2303.18223.

[8] Gao, Y., Xiong, Y., Gao, X., and others. 2023. Retrieval-augmented generation for large language models: A survey. arXiv:2312.10997.

[9] Shuster, K., Poff, S., Chen, M., and others. 2021. Retrieval augmentation reduces hallucination in conversation. arXiv:2104.07567.

[10] Topsakal, O. and Akinci, T. C. 2023. Creating large language model applications utilizing LangChain: A primer on developing LLM apps fast. In *Proceedings of the International Conference on Applied Engineering and Natural Sciences*, 1(1), 1050–1056.

[11] Brown, T. B. 2020. Language models are few-shot learners. arXiv:2005.14165.

[12] Touvron, H., Lavril, T., Izacard, G., and others. 2023. Llama: Open and efficient foundation language models. arXiv:2302.13971.

[13] Bai, J., Bai, S., Chu, Y., and others. 2023. Qwen technical report. arXiv:2309.16609.

[14] Es, S., James, J., Espinosa-Anke, L., and others. 2023. Ragas: Automated evaluation of retrieval augmented generation. arXiv:2309.15217.