Design and Validation of an Intelligent Query System for Operating Procedures of a Certain Molten Salt Reactor Based on RAG

Wen Yiqi, Song Guanlin, Han Lixin, Hou Jie

2026-07-28

ChinaRxiv

Abstract

Operating procedures are the "laws and regulations" for the safe operation of reactors. To address the pain points of low query efficiency and inaccurate information acquisition in the paper-based operating procedures of a molten salt reactor, as well as the insufficient retrieval accuracy caused by the homogenization of their format and content, an intelligent query system for operating procedures adapted to this molten salt reactor was constructed by combining retrieval-augmented generation with large language model technology. The PRF pseudo-relevance feedback mechanism and reranking were introduced into hybrid retrieval to achieve fast and accurate retrieval of procedure content, and the system was tested using a self-built test question-answer pair set. The results show that the proposed retrieval strategy maintains a stable retrieval time of 3s, achieves a Hit of 0.96 and an MRR@3 of 0.86, with an average end-to-end system latency of 15.9s and an answer generation accuracy of 0.79. The proposed retrieval strategy can significantly improve the retrieval accuracy of the operating procedures for this molten salt reactor, and the constructed system demonstrates good application performance.

Full Text

Design and Validation of an Intelligent Query System for Operating Procedures of a Certain Molten Salt Reactor Based on RAG

Wen Yiqi[1,2] Song Guanlin[1,2] Han Lixin[1] Hou Jie[1]

1 (Shanghai Institute of Applied Physics, Chinese Academy of Sciences, Shanghai 201800)
2 (University of Chinese Academy of Sciences, Beijing 100049)

Abstract Operating procedures are the "laws and regulations" governing the safe operation of reactors. To address the pain points of low query efficiency and inaccurate information acquisition in paper-based operating procedures for a certain molten salt reactor, as well as the problem that the homogenization of their formats and content easily leads to insufficient retrieval precision, an intelligent query system for operating procedures adapted to this molten salt reactor was constructed by combining retrieval-augmented generation and large language model technologies. PRF was used as a relevance feedback mechanism and reranking was introduced into hybrid retrieval, enabling rapid and accurate retrieval of procedural content. A self-built test question–answer set was used to test the system. The results show that the proposed retrieval strategy has a stable retrieval time of 3 s, a Hit of 0.96, an MRR@3 of 0.86, an average end-to-end system latency of 15.9 s, and an answer-generation accuracy of 0.79. The proposed retrieval strategy can significantly improve the retrieval accuracy of operating procedures for this molten salt reactor, and the constructed system

has good application performance.

Keywords RAG; intelligent query system; molten salt reactor; information retrieval; intelligent operation and maintenance

CLC number TP3

The Design and Validation of An Intelligent Query System for Operating Procedures of a Certain Molten Salt Reactor Based on RAG

WenYiQi[1,2] SongGuanJin[1,2] HanLiXin[1] HouJie[1]

[1](Shanghai Institute of Applied Physics, Chinese Academy of Sciences, Shanghai 201800, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract [Background]: Operating procedures are the "laws and regulations" governing the safe operation of reactors. However, they are mainly stored in paper. In experimental scenarios based on the reactor, it is difficult for operators and others to quickly locate and obtain relevant information through manual queries. Long and intense manual queries can easily lead to errors in parameter reading and other issues. Traditional manual queries are inefficient and highly dependent on experience. [Purpose]: To address the issues of low efficiency in traditional manual query methods, this solution aims to enable operators and other relevant personnel to efficiently obtain the required procedural information, thereby reducing their workload and enhancing their operational efficiency. [Methods]: By combining RAG and LLM, a Intelligent Query system for the operation procedures of a certain molten salt reactor was constructed, and an exclusive search strategy was designed for the operation procedures. The parameters were optimized using Optuna to determine the optimal parameters. [Results]: Through testing of the system, the retrieval time of the proposed dedicated retrieval strategy remained stable at 3 seconds, the Hit rate reached 0.96, the MRR@3 reached 0.86, the average end-to-end delay of the system was 15.9 seconds, and the accuracy of answer generation was 0.79. [Conclusions]: In conclusion, the proposed exclusive retrieval strategy not only ensures high accuracy while

---

No funding.

First author: Wen Yiqi, male, born March 2, 2003, received a bachelor' s degree in engineering from South China University of Technology in 2024; research direction: intelligent operation and maintenance of molten salt reactors

Corresponding author: Hou Jie, E-mail: houjie@sinap.ac.cn

Received date: 20XX-00-00; revised date: 20XX-00-00

reducing retrieval time, but also ensures stable system operation and excellent performance.

Key words RAG, Intelligent Query System, Molten Salt Reactor, Information Retrieval, Intelligent Operations and Maintenance

Introduction

During the operation of nuclear power, human error is one of the core inducements of nuclear safety accidents. The Three Mile Island accident was essentially a cognitive error caused by defects in human–machine interface design and insufficient operator training: unreasonable instrument design misled the operators, causing them to make erroneous judgments during the critical window of accident handling[^1], which led to the deterioration of an initial equipment failure into a core-melting accident. The Chernobyl accident was the result of insufficient safety awareness at the management level, delayed decision-making during the emergency response phase, and cognitive blind spots regarding beyond-design-basis natural disasters[^2]. The Fukushima nuclear accident was caused by the combined effects of regulatory negligence, delayed emergency-response decision-making, and cognitive blind spots regarding beyond-design-basis natural disasters[^3], ultimately evolving into a major nuclear safety accident. Lessons from these accidents show that, without an effective human-factors management system, true nuclear safety cannot be achieved.

Because of the safety shortcomings exposed by the three major nuclear accidents, the digital transformation of nuclear power should not focus solely on technological updates, but should instead reconstruct human–machine collaboration models, improve organizational learning mechanisms, enhance system resilience to disturbances, and remedy structural defects in traditional human-factors management systems at the source[^4][^5]. In this transformation process, digital procedures and verification and validation of nuclear-safety-grade software constitute two major technical supports for improving human reliability[^6]. Chiuhsiang Joe Lin et al. evaluated the impact of digital operating procedures on operators through design experiments, and the experimental results showed that they can improve performance, strengthen situational awareness, and enhance attitude perception[^7]. O' Hara, Zhang Li, and others pointed out that, compared with traditional paper procedures, digital operating procedures can shorten the time required to locate key information and reduce the risk in typical operation error types by an order of magnitude[^6][^8]. Gao Chao and others proposed that, although digital technology has become a general trend in the development of nuclear instrumentation and control technology, the verification and validation process for nuclear-grade software remains a key issue[^9]. Therefore, rigorous safety verification is a prerequisite for applying digital systems in the nuclear power field. Nuclear-grade software development must take both reliability and stability into account, and standardized verification testing

should be used to ensure compliance with relevant standards and prevent safety accidents induced by software design defects.

The Thorium Molten Salt Reactor Nuclear Advanced Energy System (TMSR) is a fourth-generation advanced nuclear energy system independently developed in China[^10]. The relevant operating procedures are the "laws and regulations" that ensure its safe, stable, and economical operation. Their contents are highly specialized, and their document structures are complex. At present, they are still mainly stored as paper documents or ordinary electronic documents. In experimental scenarios based on molten salt reactors, it is difficult for operators and others conducting manual queries to quickly locate and obtain the information they need. Under prolonged, high-intensity work, operators are vulnerable to factors such as operational fatigue and environmental interference, which can lead to frequent problems such as parameter misreading and version confusion of documents. Traditional manual lookup and retrieval are time-consuming, inefficient, and highly dependent on experience.

Although current search engines, SQL queries, and electronic documents have alleviated retrieval needs to a certain extent, querying molten salt reactor operating procedures still has certain limitations[^11]. Traditional search engines are constrained by keyword-rule matching mechanisms and are prone to ambiguity and mismatches when faced with colloquial, vague, or specialized queries. Although SQL queries are suitable for structured data retrieval, operating procedures are mostly unstructured or semi-structured texts; converting them into structured databases would not only require high system reconstruction and maintenance costs, but also increase the learning burden on personnel. In addition, existing electronic procedure documents still essentially rely on manual browsing and shallow searches. When faced with large volumes of procedures of different types and for multiple scenarios, information-location efficiency is low, and omissions of key parameters and operating conditions are very likely. Such deficiencies in existing technologies make it difficult to meet the requirements of molten salt reactor operating procedures for highly accurate, strongly real-time, and traceable real-time responses.

In recent years, generative large language models (LLMs), represented by ChatGPT[^12] and DeepSeek[^13], have emerged, making it possible for computers to "understand" natural language. However, because large language models lack pretraining on specialized domain knowledge, when directly asked questions about operating procedures they are prone to output answers that do not conform to the facts, i.e., hallucinations. To address these issues, Lewis et al. proposed retrieval-augmented generation (RAG)[^14], enabling large models to generate answers based on specialized knowledge bases so as to alleviate the "hallucination" problem of large language models, thereby providing a feasible path for reliable application of large language models in specialized domains.

Bochu Li et al. used databases to implement vectorized storage of safety-regulation clauses, integrated RAG with safety-regulation clauses, and constructed a Q&A system for power-system regulations, confirming the

applicability of RAG in safety-regulation documents[^15]. Junde Wu et al. designed the MedGraphRAG query framework, with knowledge graphs as the core; they designed a new-structure knowledge-graph triplet graph and a new retrieval strategy and applied them to RAG, improving the accuracy of medical knowledge queries[^16]. Jian Cao et al. designed CPEQA, an intelligent Q&A system applied to national secrecy, publicity, and education. The system integrates keyword embedding, word-vector similarity comparison, and prompt engineering, improving the precision of knowledge-base question answering and proving that RAG is suitable for confidential-knowledge queries[^17]. In summary, RAG has already been applied in various professional fields and has performed well.

The nuclear energy field imposes extremely high requirements on the determinacy, interpretability, and fault tolerance of system operation. When applying RAG and LLMs to molten-salt-reactor operating-procedure query scenarios, because their inherent probabilistic generation mechanisms and hallucination phenomena can easily lead to insufficient confidence in outputs, it is difficult to meet the nuclear field' s compliance requirements for interpretability, verifiability, and traceability. In practical scenarios, high standards are also imposed on the system' s rapid response capability and retrieval accuracy. These contradictions have become one of the key issues urgently to be solved in the current field of intelligent operation and maintenance for nuclear power.

Nevertheless, RAG has already demonstrated application value in non-safety-level or auxiliary scenarios in the nuclear field. For example, Chang D et al. proposed a multi-agent retrieval-augmented generation system for the nuclear-waste management field, integrating Agents with RAG to improve decision-making accuracy. Case studies showed that, while ensuring compliance and safety requirements, the system can answer questions based on documents and achieve efficient retrieval and semantic expression[^18].

In summary, for the special operating-procedure system of molten-salt reactors and the extremely high requirements for nuclear-safety assurance, targeted research and practical implementation are still lacking. An intelligent query solution meeting nuclear-safety-level requirements is urgently needed to solve the low efficiency of traditional manual retrieval. Therefore, this paper proposes a design scheme for an intelligent query system for molten-salt-reactor operating procedures that is adapted to nuclear-safety-level requirements. Under the premise of meeting nuclear-safety requirements, it realizes efficient and accurate querying of procedure information and provides support for intelligent operation and maintenance of molten-salt reactors.

1. System Design

1.1 File Characteristic Analysis

The operating-procedure documents described in this paper are the original operating procedures for a certain molten-salt reactor. These operating procedures

are mainly divided into seven major categories, including AOP, SOP, and others. There are 274 documents in total. The average number of pages per document is about 20, and the average number of characters is about 8,000. The number of operation steps varies according to the difficulty of the task; for example, a certain procedure contains about 60 operation steps. The document format is complex, containing multimodal data such as cross-page tables, text, and images. The format is highly consistent, generally adopting a three-part structure of "reference document–process limits–operation-step rules," among which the operation rules are all in the form of cross-page tables. The document contents are relatively long and highly specialized, containing many professional terms and abbreviations such as system names, equipment names such as valves, and operating parameters such as temperature. There are highly similar operating procedures; for example, some procedures involve similar equipment and similar operations, differing only in the number of devices to be inspected or in operations on different devices.

1.2 System Requirements Analysis

In view of the characteristics of the operating-procedure documents analyzed above, the system must meet the following design objectives: the knowledge base should have independent iterative updating capability; it should support natural-language understanding and be able to handle colloquial and generalized state queries; it should realize efficient and accurate retrieval of procedure contents and mitigate retrieval interference caused by document homogenization; output results should include source file names and preview links to ensure content traceability; and it should also support multimodal interaction, adapting to scenarios in which manual input of questions is not possible because operations are being performed.

1.3 System Design

The intelligent query system for molten-salt-reactor operating procedures takes a traditional RAG architecture as its core framework. It consists of a knowledge base, retrieval module, and interaction module, collaboratively realizing the full-process question-answering capability of "knowledge storage → precise retrieval → intelligent generation." The system workflow is shown in Figure 1. The knowledge base is responsible for storing documents together with their vectors and related information. The retrieval module is responsible for retrieving chunks and documents related to the question from the knowledge base. The interaction module is responsible for receiving the user' s question and sending it to the retrieval module, and then sending the retrieved relevant chunks or documents into the LLM to generate the answer most relevant to the question.

Fig. 1 Workflow Diagram of the Intelligent Query System for Operating Procedures of Molten Salt Reactors

The system interaction page is shown in Fig. 2.

Figure 1 workflow diagram of the intelligent query system for operating procedures of molten salt reactors

Figure 1: Figure 1 workflow diagram of the intelligent query system for operating procedures of molten salt reactors

Fig. 2 Interactive page display diagram of the intelligent query system for operating procedures of molten salt reactors

Figure 2: Fig. 2 Interactive page display diagram of the intelligent query system for operating procedures of molten salt reactors

Fig.2 The interactive page display diagram of the intelligent query system for operating procedures of molten salt reactors

1.3.1 Knowledge Base

The knowledge base constructed in this paper is built according to file categories, with independent storage sub-databases established; each single file corresponds to a dedicated sub-database. Each storage sub-database contains three types of core files: chunk vector files, index files, and the original document. The knowledge base should have independent iterative updating capability; that is, when a single file is updated, a global update is not required. Version iteration can be completed by adjusting only the corresponding dedicated sub-database, while the metadata are completely retained, thereby ensuring update efficiency and data integrity.

The specific construction process of the knowledge base is shown in Fig. 3. First, the document is split by adopting a recursive character-splitting strategy, and an overlapping window is set to ensure consistency of the semantic context. Next, the split chunks are vectorized and stored locally. Their metadata are then extracted, including the chunk sequence number, source-document name, high-frequency keywords, etc. Each chunk is subsequently integrated and stored together with its corresponding metadata. On this basis, preprocessing of the entire file is completed. Then the attributes of each chunk—such as the content summary and document length—are collected. Together with the chunks and metadata, these are used to construct index files. The above process is repeated until preprocessing of all procedure files is completed, and the results are persistently saved locally. The knowledge-base structure is shown in Fig. 3.

Fig.3 Flowchart for the Construction of Operating Procedures Knowledge Base

Fig. 3 Flowchart for the Construction of Operating Procedures Knowledge Base

Figure 3: Fig. 3 Flowchart for the Construction of Operating Procedures Knowledge Base

1.3.2 Retrieval Module

During construction of the module, it was found that homogenization of the formats and contents of procedure files can easily lead to deviations in retrieval matching; keyword retrieval has insufficient adaptability to synonymous expressions; vector retrieval tends to produce results that are semantically similar but have low matching accuracy for professional terminology; and the ranking stability of results from multiple retrieval paths is insufficient. These factors can cause highly relevant procedure fragments to rank lower, thereby affecting the accuracy of LLM answers. Therefore, how to balance retrieval accuracy, response efficiency, and source traceability is a core issue that must be addressed in system design.

Mansi Garg et al.' s biomedical literature question-answering system relies on MiniLM semantic embeddings and FAISS vector search, achieving good performance[^19]. Hei Yu Chan et al. introduced query-rewriting and technical-term semantic alignment into a power-domain customer-support system, and combined them with knowledge-graph matching and fused reranking of retrieval results to improve accuracy[^20]. Fangshu Chen et al.' s IFQA-LLM financial knowledge question-answering system proposed a multi-path retrieval strategy, using RRF ranking to merge vector and keyword retrieval results; combined with reranking, this improves accuracy and can handle complex queries[^21]. ZHANG Jinying et al. introduced the BM25 inverted-index algorithm into RAG retrieval, reducing the inaccuracy of vector-database retrieval[^22].

To solve the above problems, this paper proposes a dedicated retrieval strategy adapted to molten-salt-reactor operating procedures. By introducing PRF pseudo-relevance feedback and a reranking mechanism into traditional hybrid retrieval, and by replacing weighted fusion with RRF ranking fusion, high-precision retrieval under low latency is achieved, while also providing technical support for system traceability.

The retrieval process is as follows: after the system receives a natural-language question entered by the user, it first invokes keyword retrieval to recall relevant chunks. The keyword matching formula is as follows:

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot (1 - b + b \cdot \frac{len(D)}{avg\_len})} \quad (1)$$

In the formula, N is the total number of documents; n is the number of documents containing the keyword; qi is the i-th term in the query; $IDF(qi)$ is the inverse document frequency; $f(qi, D)$ is the number of occurrences of term qi in document D; $len(D)$ is the length of the document; avg_len is the average length of all documents; and $k1$ and b are tuning parameters, which control the

effects of term-frequency saturation and document-length normalization, respectively.

Next, the top-k chunks returned by keyword retrieval are read, and their vectors are averaged. This average vector is then weighted and fused with the question vector to generate the revised query vector. PRF constructs the revised query vector through the relevance mechanism as follows:

$$q_{prf} = \alpha \cdot q + \beta \cdot d \qquad (2)$$

In the formula, $qprf$ is the revised query vector; α is the weight of the original query vector; q is the original query vector; β is the weight of the average vector of the feedback chunks; and d is the average vector of the feedback chunks.

Vector retrieval is then performed: the revised query vector and all vectors returned by keyword retrieval are used to compute cosine similarity, thereby quantifying semantic consistency. The cosine-similarity formula is as follows:

$$cosine\ similarity(A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \qquad (3)$$

In the formula, A and B are respectively the vectors of text block A and text block B, and $\|A\|$ and $\|B\|$ are respectively the magnitudes of the corresponding vectors.

To address the problem that the rankings returned by the two retrieval methods are based on different criteria and have inconsistent scoring scales, the system introduces the RRF rank-fusion algorithm. It uniformly fuses and ranks the keyword-retrieval results and vector-retrieval results, thereby improving the ranking positions of highly relevant fragments in the final candidate results. The RRF ranking formula is as follows:

$$RRF\_score(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)} \qquad (4)$$

In the formula, $RRFscore$ is the final RRF fusion score of document d; $r(d)$ is the ranking of document d in the i-th retrieval algorithm; and k is a smoothing constant, used to avoid division by zero and to adjust the weight differences among document rankings. Here, k is set empirically to 60.

Finally, the ranked chunks are sent to the reranking model to generate question-chunk pairs for final verification, ensuring their relevance to the question. The documents to which the relevant chunks belong are then located and deduplicated.

For the characteristics of regulation texts—large volume, complex content, and high similarity of expression—this retrieval strategy can ensure retrieval accuracy under low time consumption, while satisfying the dual requirements of

Search strategy flowchart

Figure 4: Search strategy flowchart

boiler-code queries for precise matching and semantic understanding. It remedies the information omission and semantic deviation defects that exist in a single retrieval method, provides high-quality candidate content for subsequent answering, and the retrieval process is shown in Figure 4.

Figure 4 Search strategy flowchart

Fig.4 Search Strategy Flowchart

1.3.3 Interaction Module

The interaction module in this paper is responsible for receiving user input and forwarding it to the retrieval module, and then receiving the chunks output by the retrieval module to construct prompts suitable for LLM generation, which are input into the LLM to generate relevant answers. However, LLMs are stochastic when generating answers and may summarize, rewrite, or otherwise process the required information. Because of the special nature of regulations, the LLM is required not to perform any processing on the retrieved original text content, but to directly output the original text content relevant to the question.

This paper adjusts the temperature parameter of the LLM, while using the designed prompt-optimization engineering to construct answer templates for the LLM, and combines the settings of system instructions to effectively constrain the randomness of the LLM when generating answers, thereby ensuring that the final generated content is highly relevant and strictly unified in format. Because of nuclear-safety characteristics, the LLM is forced to output the original regulatory text in its answers; therefore, the temperature parameter is set to 0.

2 Retrieval Strategy Verification and Optimization

To evaluate the influence of different retrieval strategies on the retrieval performance of the regulatory-query system, this paper constructs a multi-strategy comparative experiment. By comparing the core performance indicators of each scheme, the optimal retrieval strategy is selected, ensuring that the experimental results can objectively reflect the practical application effects of each retrieval strategy.

2.1 Experimental Grouping

According to different strategies, the experiment is divided into five groups: Group 1 uses only keyword retrieval; Group 2 uses only vector retrieval; Group 3 uses hybrid retrieval combining keywords and vectors, with weighted fusion used for scoring; Group 4 adds a PRF pseudo-relevance mechanism on the basis

of Group 3; Group 5 adds RRF fusion ranking and reranking mechanisms on the basis of Group 4. The parameter settings of the above strategies are the same when testing their indicators.

2.2 Performance Testing Method

The experiment uses retrieval time, hit rate Hit, and precision MRR@3 as evaluation indicators for retrieval strategies. Retrieval time refers to the time from when the user inputs a question to completion of reranking; the hit rate Hit indicates whether the most relevant document is retrieved among the retrieved documents: if retrieved, it is recorded as 1, and if not retrieved, it is recorded as 0; precision MRR@K indicates the rank of the document most relevant to the question. For example, for MRR@3, if the most relevant document ranks first, then MRR@3 = 1; if it ranks third, then MRR@3 = 1/3, and so on.

2.3 Data Preparation

The regulatory documents described in this paper are all original documents of operating procedures for a molten-salt reactor, and the files are in PDF format.

To verify whether the system can accurately retrieve and return operating procedures related to user questions, this paper constructs a test question-answer dataset. The questions are all obtained from investigations with operators and relevant operation-and-maintenance personnel, and are combined with daily duty logs and records of frequently used procedures. More than 100 high-frequency query requests are collected; after deduplication, merging, and semantic-expression normalization, they are condensed into 70 representative question-answer pairs. Due to nuclear-safety requirements, all reference answers are the original texts of the relevant regulatory documents. Based on actual usage scenarios, five categories of questions are set: the first category is used to check whether the system can output complete standardized answers; the second category focuses on questions developed around partial content of the procedures; the third category consists of representative questions raised in combination with the system operating status; the fourth category is used to query all operations covering the system operating status; and the fifth category is used to check whether the system can accurately understand colloquial expressions and provide standardized answers. Table 1 shows specific examples of each type of question.

Table 1 Demonstration of specific problem instances

Type of Question	Example of specific question
1	Please introduce the x operating procedure.
2	What are the safety restrictions for the x operating procedure?
3	What should be done if the x cabinet fault light is on?
4	How should dust on the filter screen of the cabinet be handled?
5	What should be done if the control rod is stuck?

Type of Question	Example of specific question

2.4 Environment Setup

The system is developed based on the Langchain platform. Knowledge-base construction relies on Chroma. The LLM is DeepSeek:R1-Qwen-Distill, the text-embedding model is BGE-M3, and the reranking model is BGE-Reranker. None of the models undergoes any preprocessing or fine-tuning. The GPU is a single RTX 4090.

2.5 Test Results and Analysis

Table 2 compares the retrieval time, Hit, and MRR@3 indicators corresponding to different retrieval strategies. The results show that keyword retrieval is the fastest, but its retrieval accuracy is insufficient; the MRR@3 of vector retrieval improves, but the retrieval time is as high as 30.03 s; hybrid retrieval combining keyword and vector retrieval ···

The improvement in MRR@3 for search is limited, because the protocol content contains a large number of formats and content-similar samples, which constrain retrieval performance. After introducing PRF pseudo-relevance feedback, MRR@3 rises to 0.80, indicating that PRF can alleviate the above pain points. Considering that keyword retrieval and vector retrieval have different scoring scales, the RRF ranked-fusion algorithm is adopted, and a reranking mechanism is introduced to further reduce retrieval errors caused by textual similarity. The experimental results show that Hit increases to 0.90, MRR@3 reaches 0.83, and retrieval time is stabilized at around 3 s, verifying that the reranking mechanism can optimize retrieval effectiveness. The dedicated retrieval strategy designed in this paper— "keywords + PRF + vectors + RRF + reranking" —performs well.

Table 2 Results of evaluation indicators for different search strategies

Strategies	Search time consumption / s	Hit	MRR@3
Keywords	0.05	0.83	0.67
Vectors	30.03	0.81	0.73
Keywords + vectors + weighted fusion	0.34	0.85	0.75
Keywords + PRF + vectors + weighted fusion	0.33	0.88	0.80
Keywords + PRF + vectors + RRF + reranking	3.07	0.90	0.83

2.6 Parameter Settings and Optimization

To further optimize strategy performance and improve system retrieval efficiency and reliability, this paper conducts hyperparameter optimization for the above dedicated retrieval strategy. The parameters that affect retrieval accuracy and

Fig. 5 Changes in Hit and MRR@3 with the number of iterations

Figure 5: Fig. 5 Changes in Hit and MRR@3 with the number of iterations

time consumption mainly include the following: keyword retrieval $k1$ and b; the number of chunks returned by keyword retrieval, keyword-top-k; the number of chunks for PRF pseudo-relevance feedback, prf-top-n; the query-vector weight of the PRF pseudo-relevance mechanism, prf-α; the average vector weight of feedback chunks, prf-β; the number of chunks sent to the Reranker, rerank-top-k; and the final number of returned documents, final-top-k.

Retrieval strategy performance is jointly affected by multiple parameters. If manual stepwise parameter tuning and validation are adopted, the workload is large and tuning efficiency is low. Therefore, this paper introduces the Optuna hyperparameter optimization framework to carry out automatic parameter search. The framework incorporates sampling algorithms such as a tree-structured Parzen estimator, can adaptively search the hyperparameter space based on historical evaluation results, and can efficiently approach the optimal parameter combination within a limited number of trial rounds.

The optimization experiment proceeds as follows: first, a set of parameters is specified as the baseline configuration, and the QA pairs are used to run the full-chain retrieval program, obtaining the three baseline indicators—Hit@3, MRR@3, and retrieval time—as performance references. Optuna iterative optimization is then started. In each iteration, the TPE algorithm generates, based on historical evaluation results, a set of parameter combinations satisfying the constraints; the complete retrieval chain is executed on all test QA pairs and the objective-function value is calculated. By continuously fitting the probability distributions of high-quality and low-quality parameters, the algorithm gradually converges toward a better result region.

Figure 5 shows the results of Hit and MRR@3 as the number of Optuna iterations changes. It can be seen from the figure that, during 30 iterations of parameter optimization, the 28th iteration achieves the highest indicators, with Hit reaching 0.96 and MRR@3 reaching 0.86. The corresponding optimal parameter combination is: $k1 = 1.02$, $b = 0.37$, the number returned by keyword retrieval top-k is 20, prf-top-n is 17, prf-$\alpha = 0.87$, prf-$\beta = 0.13$, rerank-top-k=17, and final-top-k=6. To ensure the consistency and comparability of evaluation results among different parameter combinations, in the performance evaluation stage this paper fixed Hit@3 and MRR@3 as evaluation indicators; that is, only the hit rate and mean reciprocal rank of the top 3 documents in the ranked results are calculated.

Fig. 5 Changes in Hit and MRR@3 with the number of iterations

Fig.6 The Boxplot of average end-to-end latency

Figure 6: Fig.6 The Boxplot of average end-to-end latency

3 System Test Results Display

To verify the comprehensive performance of the system, this paper conducts end-to-end performance evaluation experiments based on the test question-answer set. The retrieval module adopts the optimal parameter configuration obtained through the hyperparameter search described above, and two new evaluation metrics are added: end-to-end latency and answer accuracy. Among them, end-to-end latency is defined as the total time consumed from the user submitting a query request to the system beginning to output the response result; answer accuracy is the cosine similarity between the generated answer and the reference answer.

After all test question-answer pairs were cyclically tested in the system for 100 rounds, statistics showed that the system' s average end-to-end latency was 15.9 s, the average answer accuracy was 0.79, the average Hit value was 0.96, and the average MRR@3 was 0.86. The system performance was good. By analyzing the test results of the question-answer pairs, it was found that for 3 different types of questions, the relevant regulatory documents could not be successfully retrieved; the Hit failed to reach 1. The reason may be that keyword retrieval is sensitive to synonym rewriting and word-segmentation errors, and failed to recall the expected chunks; MRR@3 did not reach 1, which may also be caused by the inherent limitations of keyword retrieval.

Figure 6 is a boxplot of the end-to-end latency for all test question-answer pairs. As can be seen from the figure, the median is 16.03 s and the mean is 15.93 s, which are basically consistent, indicating that the latency distribution is approximately symmetric. The lower quartile Q1 and upper quartile Q3 corresponding to the box are 15.29 s and 16.69 s, respectively, and the interquartile range (IQR) is 1.40 s, indicating that 75% of the queries can be completed within 16.69 s. In addition, outliers in the figure appear only below the box, meaning that some are obviously lower than the main body— "fast" —while there are no outliers above, that is, no occasional high-time-consumption "slow" cases appear.

Fig.6 The Boxplot of average end-to-end latency

Figure 7 is a bar chart of the cosine similarity between the system-generated answers and the reference answers for each test question-answer pair. The average semantic similarity of all question-answer pairs is 0.79. Because the system adopts a generation strategy that directly outputs the original regulatory text, and the answers do not involve content summarization or semantic rewriting, typical LLM hallucination and semantic deviation issues can be excluded. The reason the similarity did not reach a high level is that the answers contain redundant irrelevant regulatory content, and the output format differs from the format specification of the reference answers, which has a certain impact on the

Fig.7 Bar chart of the accuracy of each Q&A

Figure 7: Fig.7 Bar chart of the accuracy of each Q&A

similarity calculation results.

Fig.7 Bar chart of the accuracy of each Q&A

In summary, the intelligent query system for molten-salt heap operation procedures designed and constructed in this paper performs well, and the proposed dedicated hybrid retrieval strategy ···

···while reducing retrieval time, the accuracy of retrieval is also ensured. It is demonstrated that, for such documents with highly similar content and formats, PRF can be introduced as a relevance-feedback mechanism during retrieval to enhance retrieval accuracy. The system' s end-to-end latency is stable at 13–18 s. On the designed test question–answer set, the system' s answer accuracy reaches 0.79. Among the five types of questions given, it performs well overall on tasks such as general queries and status queries; for status-query questions, Hit reaches 0.94 and MRR@3 reaches 0.83.

4 Conclusion

This paper integrates retrieval-augmented generation (RAG) and large language model (LLM) technologies to construct an intelligent query system for molten-salt-reactor operating procedures. In view of the homogenized wording and high semantic overlap of operating-procedure texts, an adaptive hybrid retrieval strategy is designed, and a PRF pseudo-relevance-feedback mechanism is introduced, effectively alleviating the bias problems that arise in conventional hybrid retrieval during procedure queries and improving retrieval accuracy. Optuna is used for hyperparameter optimization, and system performance is tested using the optimal parameters. The results show that the constructed system operates stably and achieves good retrieval accuracy and response performance. This study successfully realizes intelligent and natural-language querying of molten-salt-reactor operating procedures, providing a feasible technical solution and practical example for intelligent retrieval of procedure materials and human–machine interaction applications in the nuclear power field, and also offering a reference basis for the subsequent development and optimization of intelligent nuclear-power operation and maintenance systems.

4.1 System Limitations and Engineering Application Challenges

This system is built on RAG and LLM. The operating mechanism of RAG is interpretable, but the LLM responsible for answer-generation tasks is a typical black-box model; that is, because the generation mechanism has not yet received a complete and clear theoretical explanation, the reliability of its output content and the quantification of its credibility still require further in-depth exploration

and verification. Considering the extremely high safety and rigor requirements of nuclear engineering operating procedures and nuclear safety culture, the intelligent query system for molten-salt-reactor operating procedures developed in this paper must be used in conjunction with a manual review mechanism in actual application scenarios. The limitations and engineering application challenges are analyzed as follows:

Because nuclear-reactor operation and maintenance require extremely high safety and rigor, when applied in real scenarios this system can serve only as an auxiliary tool. Although the generated answers are all based on the original procedure texts, they must be used as reference operational steps only after manual review has confirmed their correctness. During a query, the system provides the names of the documents cited by the current answer and preview links to the originals; users may conduct manual rechecking and verification according to actual conditions, and only after confirming correctness should the results be used as a basis for auxiliary decision-making.

For unknown conditions, to ensure the credibility and reliability of answers, this paper strictly restricts the LLM through refined prompt engineering, requiring it to summarize, integrate, and standardize the retrieved relevant content. It is prohibited from answering independently outside the retrieved content. For unknown conditions not recorded in the knowledge base or system states that have not appeared, the system uniformly returns: "No relevant original text was found in the retrieved procedure content." This prevents the LLM from generating unsupported guidance, and the handling of unknown conditions is left to human judgment.

Throughout the lifecycle operation of a reactor, operating-procedure documents are continuously iterated and optimized in combination with experimental progress and daily operating data, thereby realizing the standardization and refinement of operating workflows. Therefore, operating procedures are characterized by frequent iteration and detailed updates; full updates are time-consuming and computationally costly, and may suffer from delayed updating and low efficiency. To address the above pain points, this paper adopts an incremental update strategy: only the new-version procedure documents are used independently to build an incremental vector database, and knowledge-base updating can be achieved by replacing database modules. However, this strategy also imposes higher technical requirements for consistency verification between old and new data.

4.2 Future Work

Regarding the problem that the system' s Hit and MRR@3 do not reach 1, further improvement can be achieved by introducing query rewriting or synonym expansion.

The current system supports only text-based question–answer interaction. It does not yet support image input for retrieving procedures, and the model also

cannot output visual image materials corresponding to query questions. Therefore, subsequent research will focus on exploring multimodal input and output capabilities for operating procedures, investigating the processing of complex procedure documents (such as those containing text, images, and cross-page tables), and introducing multimodal large models such as Qwen-VL to realize recognition of process images. Based on the recognition results, the corresponding operating procedures can be retrieved, and multimodal content such as images and tables matching the query scenario can be generated.

In addition, the next stage of work plans to conduct structured organization of operating procedures, abstract each operating step in the procedures into knowledge nodes to construct a knowledge graph, use it as supporting evidence for retrieving procedures and generating answers, and further suppress hallucinations in LLMs.

References

1. Kemeny J, Babbitt B, Haggerty P E, et al. Report of the President' s Commission on the Accident at Three Mile Island[J]. Pergamon, 1979. DOI:10.2172/6986994

2. Jr H T P. Summary report on the post-accident review meeting on the Chernobyl accident[J]. Journal of Environmental Radioactivity, 1987, 5(5):403-404. DOI:10.1016/0265-931X(87)90015-4.

3. Kurokawa K, Ishibashi K, Oshima K, et al. The National Diet of Japan. The Official Report of the Fukushima Nuclear Accident Independent Investigation Commission[R/OL]. Tokyo: The National Diet of Japan, 2012[2026-07-16]. https://www.nirs.org/wp-content/uploads/fukushima/naiic_report.pdf

4. Fleger, J O. Use of HFE Guidance for the Review of Nuclear Power Plants[J]. Proceedings of the Human Factors and Ergonomics Society Annual Meeting[2026-07-17]. DOI:10.1177/1071181320641001.

5. DAVID R D, STEPHEN F. IEEE Human Factors Standards for Nuclear Facilities: The Development Process, Available Standards, Current Activities, and the Future[J/OL]. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2019, 63(1):587-591[2026-07-17]. DOI:10.1177/1071181319631374

6. O' HARA J M, STUBLER W F, KRAMER J, et al. Computer-Based Procedure Systems: Technical Basis and Human Factors Review Guidance[R/OL]. Washington, DC: U.S. Nuclear Regulatory Commission, 2000[2026-07-16]. https://adams.nrc.gov.

7. LIN C J, HSIEH T L, YANG C W, et al. The impact of computer-based procedures on team performance, communication, and situation awareness[J]. International Journal of Industrial Ergonomics, 2016, 51:21-29.

DOI:10.1016/j.ergon.2014.12.001.

8. A Y Z, A L Z, B L D, et al. Human Reliability Analysis for Digitized Nuclear Power Plants: Case Study on the LingAo II Nuclear Power Plant - ScienceDirect[J]. Nuclear Engineering and Technology, 2017, 49(2):335-341. DOI:10.1016/j.net.2017.01.011.

9. Gao Chao, Hu Lisheng. Research on Verification and Validation Techniques for Nuclear Software[J]. Microcomputer Applications, 2010, 26(4):3. DOI:10.3969/j.issn.1007-757X.2010.04.002. Gao Chao, Hu Lisheng. Research on Verification and Validation Techniques for Nuclear Software [J]. Microcomputer Applications, 2010, 26(4): 4-5, 11. DOI: 10.3969/j.issn.1007-757X.2010.04.002.

10. Cai Xiangzhou, Dai Zhimin, Xu Hongjie. Thorium-based molten salt reactor nuclear energy system[J]. Physics, 2016, 45(9):13. DOI:10.7693/wl20160904. Cai Xiangzhou, Dai Zhimin, Xu Hongjie. Thorium-based molten salt reactor nuclear energy system [J]. Physics, 2016, 45(9):13. DOI:10.7693/wl20160904.

11. KEYVAN K, HUANG J X. How to Approach Ambiguous Queries in Conversational Search: A Survey of Techniques, Approaches, Tools, and Challenges[J/OL]. ACM Computing Surveys, 2022, 55(6): 1-40. DOI:10.1145/3534965.

12. Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023. DOI:10.48550/arXiv.2303.08774

13. Liu A, Feng B, Xue B, et al. Deepseek-v3 technical report[J]. arXiv preprint arXiv:2412.19437, 2024. DOI:10.48550/arXiv.2412.19437

14. Lewis P, Perez E, Piktus A, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks[J]. 2020. DOI:10.48550/arXiv.2005.11401.

15. Li B, Jiang Y, Xu J, et al. The Design and Implementation of an Intelligent Q&A System for Electric Power Safety Regulations Based on Large Language Model Technology[EB/OL]. 2024 6th International Conference on Energy Systems and Electrical Power (ICESEP). (2024). DOI:10.1109/iceseP62218.2024.10652122

16. Wu J, Zhu J, Qi Y, et al. Medical Graph RAG: Towards Safe Medical Large Language Model via Graph Retrieval-Augmented Generation[J]. 2024. DOI:10.48550/arXiv.2408.04187.

17 Cao J, Cao J. CPEQA: A Large Language Model Based Knowledge Base Retrieval System for Chinese Confidentiality Knowledge Question Answering[J]. *Electronics* (2079-9292), 2024, 13(21). DOI:10.3390/electronics13214195.

18 Chang D, Kim S, Park Y S. AI-Supported Platform for System Monitoring and Decision-Making in Nuclear Waste Management with Large Language Models[J]. 2025. DOI:10.48550/arXiv.2505.21741.

19 GARG M, WANG L C, GHANCHI B, et al. Biomedical Literature Q&A System Using Retrieval-Augmented Generation (RAG)[J/OL]. 2025. https://arxiv.org/abs/2509.05505. DOI:10.48550/arxiv.2509.05505.

20 CHAN H Y, HO K T, MA C, et al. Enhancing Retrieval-Augmented Generation for Electric Power Industry Customer Support[J/OL]. 2025. https://arxiv.org/abs/2508.05664. DOI:10.48550/arxiv.2508.05664.

21 Chen F, Huang Y, Wang J, et al. Ifqa-llm: intelligent intention-driven financial question-answering with large language models[J]. *Journal of Supercomputing*, 2025, 81(13). DOI:10.1007/s11227-025-07726-5.

22 Zhang Jinying, Wang Tiankun, Yao Changying, Xie Hua, Chai Linzheng, Liu Shukai, Li Tongliang, Li Zhoujun. Construction and Evaluation of an Intelligent Question Answering System for an Electric Power Knowledge Base Based on a Large Language Model[J]. *Computer Science*, 2024, 51(12): 286-292. DOI:10.11896/jsjkx.240300104. ZHANG Jinying, WANG Tiankun, YAO Changying, XIE Hua, CHAI Linzheng, LIU Shukai, LI Tongliang, LI Zhoujun. Construction and Evaluation of Intelligent Question Answering System for Electric Power Knowledge Base Based on Large Language Model[J]. *Computer Science*, 2024, 51(12): 286–292. DOI:10.11896/jsjkx.240300104.

Note: Figure translations are in progress. See original paper for figures.