ISSN 13303651 (Print), ISSN 18486339 (Online) https://doi.org/10.17559/TV-20250705002797
Original scientific paper

Research on Intelligent Monitoring of E-Government Using LLM, RAG, and Hybrid Retrieval Technologies

Huaiyu WEN*, Mengxuan HE

Abstract: Large language models are one of the core research contents of natural language processing and have been widely applied in many fields including government affairs. The problems that need to be solved in the research of large language models for government affairs are proposed, namely, data multimodality, correctly facing the trend of "model as a service", emphasizing high data security, and clarifying the boundaries of responsibility. In addition, the technical path for the research of large language models for government affairs has also been proposed. Then, the feasibility of the large language model in the process of industrial collaborative emergency preparedness is analyzed. The P-Tuning method is adopted for model fine-tuning. A local knowledge base is constructed based on the question-answering data related to the intelligent monitoring events of e-government. The content generated by the model is standardized, and the LangChain architecture is utilized to build the industrial collaborative emergency question-answering system. Based on the large model knowledge base, a hybrid retrieval dual-tower model was proposed. The model integrates the multi-path recall strategy to ensure the comprehensive retrieval results. The multi-level ranking of the retrieval results was achieved through the adoption of hybrid retrieval. The correlation of the retrieval results has been significantly improved. Further, the optimized information and the original query are input into the large language model to generate accurate answers. Experiments show that this method significantly improves the accuracy of retrieval.

Keywords: e-government Intelligent monitoring; hybrid retrieval; LLM; RAG

1 INTRODUCTION

In recent years, large-scale pre-trained language models have developed rapidly [1]. Foreign products such as the Generative Pre-trained transformer (GPT) series (GPT1, GPT2, The rapid iteration and update of GPT-3, ChatGPT/InstructGPT, and GPT-4 [2-5] have not only driven the rapid development of the entire industry but also sparked a wave of large language model research and development. Against this backdrop, domestic technology enterprises have also been actively laying out research on large language models. Huawei's "Pangu" series, Baidu's "Ernie Bot", SenseTime's "Shangliang", and Alibaba's "Tongyi Qianwen" have been released one after another. With the increasing volume of training data and the continuous update of training methods, the structure of large language models is becoming more and more complex. Many large language models have been widely applied in various scenarios such as language translation, abstract construction, named entity recognition, text classification and relation extraction, as well as in multiple industries including government affairs, finance and biomedicine.

In the ecosystem of the large model industry, its industrial chain structure is clear. According to the logic of upstream, midstream and downstream, it can be finely divided into underlying computing power service providers, large model builders and industry solution providers. In the field of underlying computing power services, service providers are committed to comprehensively integrating and efficiently providing a series of core resources such as data, networks, training tools, and chips to support the application and development of cutting-edge technologies in the industry. For instance, the single-card chip launched by NVIDIA, with its outstanding performance, has successfully completed the model training task at the level of tens of billions of parameters, providing a solid and reliable hardware foundation for the training of large-scale models. Meanwhile, industry leaders such as Huawei, Baidu, Alibaba, and Tencent are also actively investing in research and development, launching proprietary network protocols, and deeply integrating advanced hardware and chip technologies to provide high-bandwidth and low-latency network support for model training, ensuring the stability and efficiency of the training process. In terms of data resources, in addition to the widely adopted mainstream training datasets, proprietary data such as policies and official documents have also been specially incorporated to meet the application requirements of the government affairs large model, further enhancing the adaptability and practicality of the model in government affairs scenarios. At the level of general large-scale model construction, numerous technology enterprises and research institutions have been actively engaged in the research and development and optimization of training frameworks, model libraries, training sets, and tool platforms. By constantly improving algorithms and architectures, the training and deployment process of large models has been accelerated, and relatively complete solutions have been formed in basic functional areas such as intelligent search, robot question answering, and intelligent recommendation. AI large models have further integrated specific data and expert experience in the government affairs field, greatly enhancing the business practicality of the models in the government affairs field.

In terms of providing industry solutions, relevant vendors have demonstrated outstanding customized training capabilities, enabling them to deeply empower large model tools based on specific government affairs needs to address complex government affairs scenarios. For instance, in fields such as intelligent approval, intelligent robots, and automated decision-making, the application of large models has demonstrated significant advantages. It not only significantly enhances the satisfaction of the public in handling affairs, allowing them to experience a more convenient and efficient way when enjoying government services, but also greatly improves the work efficiency and service level of government personnel, injecting new vitality into the intelligent process

of government work. During the application of large models, by collecting and analyzing the generated data feedback, the training set can be continuously enriched and optimized. Through this continuous learning and optimization mechanism, the large model can better adapt to and meet the actual needs of government affairs work, achieving the continuous optimization and upgrading of government services. This paper reviews the research progress of large language models. Secondly, it focuses on reviewing the research progress of large language models. Finally, it introduces the application of large language models in the field of government affairs and discusses several problems that need to be solved in the research of government affairs large language models, with the aim of guiding the research of large language models and promoting their wide application in the field of government affairs. The first part of the article is the introduction, and the second part is an introduction to the relevant work. The third part is research methods of electronic intelligent monitoring of large language models rag and hybrid retrieval in the field of government affairs, part 4: simulation verification, part 5: conclusion.

2 RELATED WORK

In response to the call of the state, a large amount of resources has been invested in building integrated online government service platforms. These platforms simplify the government service process by integrating functions such as online payment and electronic certificates, and achieve one-stop services from inquiry, application, approval to payment. Take Guangdong Province as an example. Through the "Digital Government" reform, Guangdong Province has not only achieved "one-stop online processing" of government services, but also put forward the concept of "running at most once", enhancing the transparency and efficiency of government services. Furthermore, through cross-departmental data sharing and interconnection, the Guangdong Provincial Government Services Network has become a model for government information disclosure and service platforms [6]. These issues have restricted the efficiency of information sharing and resource integration, and hindered the full play of the potential of e-government services [7]. To overcome these bottlenecks, it is particularly important to introduce large models and knowledge base technologies. Especially based on large language model technology, they have the ability to process and analyze large-scale datasets and can perform complex natural language understanding and generation tasks, providing a technical basis for e-government [8]. By integrating large language models, e-government is expected to achieve intelligent upgrades of service functions at lower human and financial costs, thereby providing more accurate and personalized services and meeting the diverse needs of the public. This technological application can not only enhance the speed and accuracy of service response, but also promote smooth collaboration among departments, break down data silos, and achieve efficient utilization of resources.

When generating responses, large language models rely on the parameterized knowledge stored internally, but they may give unreasonable or false answers for highly specialized knowledge in specific fields. In order to enhance the answering ability of large language models for domain problems, a retrieval enhancement generation method is proposed [9]. First of all, input the user query into the pre-trained language model to generate the text embedding vector; Secondly, use the retrieval model to search for relevant knowledge documents from the vector knowledge base; Finally, by combining relevant knowledge documents and user queries and using the prompt engineering method, a large language model for prompt input is formed to generate answers. To enhance the performance of the retrieval-augmented Generation method, the self-reflective Retrieval-augmented Generation (self-rag) framework was proposed [10]. Self-rag is a new type of retrieval-enhanced generation framework, which improves the quality of text generated by large language models through retrieval-enhanced and self-reflection. However, RAG requires efficient retrieval strategies and technologies related to large databases. Additionally, it is necessary to maintain the integration of external data sources and data updates. The "Retrieve, COMmpress, Prepend" (RECOMP) method was proposed [11]. This method introduces an extraction compressor and a generation compressor. Through these compressors, relevant sentences are selected or document information is synthesized to create summaries suitable for multi-document queries. The noise information contained in the documents retrieved by the retriever can affect the performance of retrieval enhancement generation. When the retrieved documents are irrelevant to the input or do not provide additional information, RECOMP can return an empty string, thereby avoiding adding useless information. This selective enhancement ensures that the model utilizes the retrieved information only when necessary. The overall performance and efficiency of the model have been improved. However, the RECOMP method is trained and evaluated based on English Question Answering datasets such as Natural Questions (NQ) and Trivia Question Answering (TrivialQA) [12]. Further research is needed for domain knowledge question answering tasks in Chinese.

There are mainly three alignment methods for the semantic space of user queries and the semantic space of related knowledge in the knowledge base, namely the traditional alignment method, the document domain alignment method, and the query domain alignment method, as shown in Fig. 1 specifically. Traditional alignment methods map document fragments and queries to the same encoding space. For example, the Query Rewrite method proposed in reference [13] adds a rewriter before the traditional retrieval-reading method. The traditional retrieval-reading method first recalls the relevant documents from the retrievaler through query statements, and then sends the relevant documents into the reader to generate the correct answers. The rewriter rewrites the user query to narrow the distance between the query and the relevant documents. However, the rewritten query is prone to semantic drift, resulting in poor retrieval results. The document domain alignment method generates hypothetical answers for queries and uses the generated hypothetical answers to recall relevant knowledge. For example, the Hypothetical Document Embeddings (HyDE) method proposed in reference [14]. HyDE is an improved retrieval method that can be used to create an effective zero-shot dense retrieval system in the absence of relevant

labels. By generating hypothetical documents that can be used to answer user input questions, relevant documents are retrieved from the vector database after vectorization. The core idea of the HyDE method is to use hypothetical documents to narrow the semantic distance between queries and knowledge related to the knowledge base. However, the accuracy of generating documents using large language models in the HyDE method is a problem. When the generated hypothetical documents contain incorrect details, it will affect the relevance of the retrieval results. Furthermore, the HyDEmethod involves generating documents and vectorization processing [15], which may lead to high computational and storage costs when dealing with large-scale document libraries. The query domain alignment method first generates some hypothetical questions for each knowledge text fragment stored in the knowledge base, then maps these questions to the query space, and finally retrieves the hypothetical questions that are closest to the query

The problem and the corresponding knowledge text fragment. For example, the Prompt Aligned Retrieval Generation (PARG) method based on prompt engineering is proposed [16]. PARG generates hypothetical question-answer pairs for knowledge text fragments through large language models and uses vector representation models to combine and embed documents with related queries to align the semantic Spaces of queries and related knowledge. However, due to the diversity of user query methods, the generated hypothetical questions are difficult to cover all user queries, and generating hypothetical questions and answers has a relatively high computational cost. Studies have pointed out that combining RAG (Retrieval-Augmented Generation) with large models can effectively improve accuracy [17, 18] and reduce factual errors, but it still faces challenges in the comprehensiveness of knowledge and the accuracy of retrieval. At present, RAG has not been trained to efficiently utilize the retrieved information [19, 20], resulting in inaccuracy when selecting and generating content, and the output does not fully match the retrieved information. A self-reflective RAG strategy was proposed [21], which improves the quality of generated content through dynamic retrieval and feedback. However, this process increases the computational cost. Although the application of RAG (retrieval-Enhanced Generation) and hybrid retrieval in the field of government affairs has significantly improved the efficiency of information processing; the deep coupling of its technical characteristics with government affairs scenarios still faces multiple challenges [22]. Firstly, there are shortcomings in the data integration and verification mechanism. Government data is scattered across multiple format documents such as PDF (Portable Document Format), DOC (Document), and PPT (Power Point), and the cross-departmental data standards are not unified, resulting in a document parsing error rate as high as 15% to 20%. For instance, a district-level government affairs platform failed to parse a PDF form and mistakenly identified the "annual turnover" field of the enterprise subsidy policy as "years of establishment", which led to a complaint from the enterprise. Secondly, the timeliness guarantee is insufficient. The update frequency of policy documents is high (such as the newly issued "Digital Government Construction Guidelines" in 2025), but the knowledge base update cycle of the RAG system is generally lagging behind by 3 to 7 days, which easily leads to the risk of "outdated policy answers". In addition, security and privacy risks are prominent. Government data involves personal sensitive information (such as ID numbers and social security records). If mixed retrieval does not adopt federated learning or homomorphic encryption technology, it may lead to data leakage. The government service hotline system of a certain province once had 5000 citizen consultation records leaked due to the search logs not being desensitized.

The application of advanced artificial intelligence technologies such as deep learning, natural language processing, and virtual digital humans to establish a large AI model for e-government aims to achieve the intelligence and automation of government services [23, 24]. The model has a powerful natural language processing capability and can accurately understand and efficiently handle various natural language texts closely related to government affairs, such as policy documents and notice announcements. The large AI model for government affairs can respond promptly to various information query requests and provide accurate and detailed information support for government staff. Meanwhile, the model can also provide auxiliary decision-making functions, helping staff to grasp policy trends and actual situations more scientifically and accurately when facing complex decisions. Although e-government has made remarkable progress in improving the efficiency and quality of government services, its development is still constrained by several challenges, manifested in the insufficiency of intelligent and refined services, obstacles to cross-departmental collaboration, and the problem of data silos. These problems have restricted the efficiency of information sharing and resource integration, and hindered the full play of the potential of e-government services [25, 26]. To overcome these bottlenecks, it is particularly important to introduce large models and knowledge base technologies. Especially based on large language model technology, they have the ability to process and analyze large-scale datasets and can perform complex natural language understanding and generation tasks, providing a technical basis for e-government [27].

3 RESEARCH METHODS OF ELECTRONIC INTELLIGENT MONITORING OF LARGE LANGUAGE MODELS, RAG AND HYBRID RETRIEVAL IN THE FIELD OF GOVERNMENT AFFAIRS

3.1 Application Construction Architecture

From a technical perspective, a modular and hierarchical AI large model application platform can be constructed through layering and abstraction. Its overall architecture (Fig. 1) can be represented as follows:

1) AI infrastructure layer: It provides infrastructure support such as GPU (Graphics Processing Unit) hardware, networks, and storage to meet the needs of model training, inference, and application.

2) Model layer: It includes the underlying large model, the model pool for storing fine-tuned government affairs large models, and related tools for model training, evaluation, and optimization.

3) Service Development layer: It provides middleware and tool support for application development, including various Agent components, RAG components, vector databases, development toolchains, and related application interface services.

4) Application layer: Various applications developed based on the capabilities of AI large models.

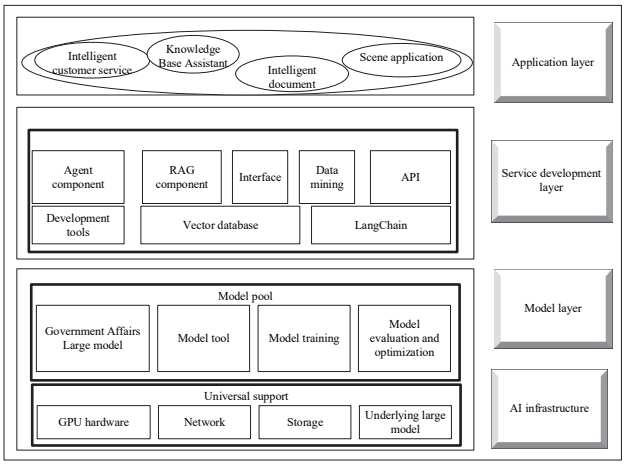

Figure 1 Architecture of the AI large model application platform construction

Abbreviations in the text is shown in Tab. 1.

Table 1 Abbreviations

Abbreviations	Full name
RAG	Retrieval-Augmented Generation
NQ	Natural Questions
TrivialQA	Trivia Question Answering
HyDE	Hypothetical Document Embeddings
KEHRAO	Keyword Extraction and Hybrid Retrieval

3.2 Application Development Technology

The application development technologies of AI government affairs large models can be classified into the following categories. 1) Prompt Engineering: It is a way to directly use large models to perform tasks. By inputting effective prompts, it guides the large model to understand the user's intention and generate responses or perform operations. The advantage of this method is that it is simple and easy to use, and interacts directly with the large model through precisely designed prompt words. However, its drawback is that it may not be able to handle complex task scenarios independently.

2) Function Calling: It is a way to connect large models with the external world and business systems. The large model itself has some "flaws" (such as mathematical calculation, timeliness, etc.), which can be improved through the Function Calling function. For example, functions such as accessing the Internet to query information and calling the interfaces of external business systems to access specific data, thereby overcoming the limitations of the knowledge and capabilities of large models. Function Calling uses the output of the large model as the input to trigger API calls to specific external systems or tools, enabling the large model to interact with the outside world to achieve more complex tasks.

3) Retrieval-Augmented Generation (RAG): It is a technical architecture that combines the two processes of Retrieval and Generation, and it integrates the capabilities of Retrieval systems and large model generation. This system uses a retrieval system (such as a vector database) to retrieve information from a large amount of external data (such as the knowledge base of the government affairs system), and then fuses the retrieved information into a large model. Through the capabilities of the large model, detailed answers based on the retrieved information are generated to produce more information-rich and accurate result outputs. Retrievy-enhanced generation is a solution to the insufficiency of "illusion and private domain knowledge" in large models.

4) Artificial Intelligence Agent (AI Agent): An Agent refers to an autonomous software program that can perceive its environment and take actions based on these perceptions to achieve specific goals. In the application development of AI large models, an artificial intelligence agent is an agent that can perceive the environment and take actions (for various tasks) to achieve the goal [28].

Specifically, technologies such as text classification, hybrid retrieval, and information extraction in large language models can serve as core technical support, as shown in Fig. 2. These technologies can not only help government agencies to quickly identify and classify massive text data, but also understand the emotional tendencies of the public, automatically extract feedback on the implementation effects of policies, as well as information such as key events and trend analyses. Enhance the intelligence and precision of government services, thereby improving the efficiency of government services, optimizing the public experience, and promoting the social governance system to move towards a smarter and more efficient direction.

Intelligent consultation and Q&A system

The intelligent consultation and question-answering system constructed with large language models can provide the public with fast and accurate government services. This system is capable of understanding and analyzing users' natural language questions, and retrieving answers or suggestions from a vast amount of government affairs information. It realizes knowledge-based intelligent self-service inquiries and business processing, reduces the working pressure of government affairs personnel, and improves the efficiency and quality of government services.

Policy Analysis and suggestions

By studying and analyzing a vast amount of policy literature, reports, news, etc., it provides precise policy suggestions for the government, helps the government understand the historical evolution, development trends and potential impacts of policies, so as to formulate more scientific and reasonable policies. In addition, large language models can also evaluate and provide feedback on the implementation effect of policies, offering a basis for the government to improve policies.

Business process Automation

By leveraging the standardized business processes in the knowledge base and integrating the data processing capabilities of large models, the automation of business processes can be achieved, repetitive labor can be reduced, and service efficiency can be enhanced.

Personalized service customization

Analyze public behavior data and combine it with business knowledge in the knowledge base to provide users with personalized government services. For example, it can

recommend government affairs information and services to users based on their browsing history, search records and consultation records. In this way, large language models can enhance the user experience and satisfaction of government services.

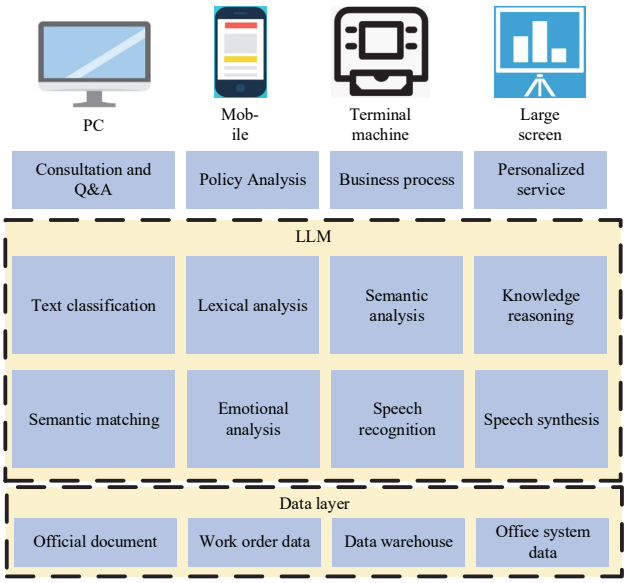

Figure 2 Architecture of the e-government system

3.3 Hybrid enhanced retrieval

The core of the alignment Optimization method based on Keyword Extraction and Hybrid Retrieval (KEHRAO) proposed in this paper lies in enhancing the performance in the retrieval stage. It makes full use of the language understanding and generation capabilities of large language models through prompt engineering. Firstly, the large language model is utilized to extract the key information of user queries. The main extracted key information includes time, place nouns, proper nouns, etc. Secondly, the combined query composed of the keyword list extracted by the large language model after concatenating the user query is input into the sparse retrieval model to recall the relevant documents, and the combined query is input into the dense retrieval model to recall the relevant documents. Then, the relevant documents recalled by sparse retrieval and dense retrieval are processed as a union, and the relevant texts are sorted through the resorter, thereby enhancing the sorting ability in the retrieval stage. Finally, input the relevant documents of the user's query and the output of the resorter into the text filter, extract the key information text that can answer the questions, merge the extracted key information text with the user's query, and input the large model through the prompt engineering method to generate the response and return it to the user.

The KEHRAO method jointly processes the documents recalled by sparse retrieval and dense retrieval in the reordering stage, and uses the reordering model to reorder the relevant texts of the retrieval recall. As shown in Fig. 3, in the key information extraction stage of KEHRAO, the user's questions and the relevant text output by the resorter are input into the text filter. The goal is to extract the key information text that can answer the questions and filter out the irrelevant parts in the relevant text, thereby reducing the noise of the input text of the large model and improving the accuracy of the large model in answering questions.

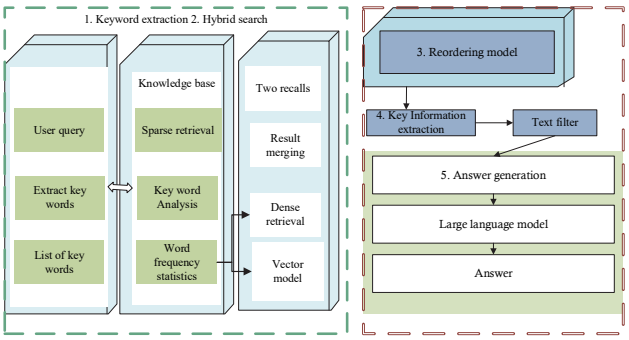

Figure 3 KEHRAO method framework

3.4 Keyword Extraction

When using the BM25 method for sparse retrieval, due to the low frequency of keyword occurrence in the query, the effect of sparse retrieval is not good in practical applications. To solve the problem of low matching degree between user queries and documents in the knowledge base in sparse retrieval, the KEHRAO method adopts query expansion and enhancement technology. It concatenates the keyword list extracted by the large language model with user queries to increase the word frequency of keywords. When conducting sparse retrieval, the matching of documents in the knowledge base can focus on corresponding keywords, thereby enhancing the recall effect of sparse retrieval. Keyword extraction uses a large language model and extracts the keywords queried by users through the method of prompt engineering. The prompt template for keyword extraction is shown in Tab. 2.

Table 2 Keyword extraction prompt template

Prompt	Prompt text
Keyword extraction prompt template	You are an excellent keyword extractor and your task is to accurately extract keywords. Each keyword in the user's question. Your reply must be a list of key extracted keywords. Each element in the list corresponds to the relevant template key words you extracted respectively. Do not reply with anything else not mentioned above. The user's question is: {The question entered by the user}

3.5 Hybrid Retrieval

The input of sparse retrieval is the combined query composed of query Q and keyword list concatenation and the document set D in the knowledge base, and the output is the K documents most relevant to the query. Sparse retrieval usually indexes and retrieves data based on some form of discrete representation, such as keywords or phrases. This method emphasizes the selection of a small number of highly relevant features (such as vocabulary or labels) from the document set for indexing. Sparse retrieval has advantages in terms of convenience and retrieval response speed. For precise queries with known exact terms, this method is both fast and effective. The sparse retrieval model in this paper adopts the BM25 algorithm. The core idea is to measure the correlation between Document set D and query Q based on Term Frequency

(*TF*) and Inverse Document Frequency (*IDF*), while considering the influence of document length information on the correlation. The correlation score between the query and the document calculated by BM25 is shown in Eq. (1) as follows:

$$score(D, Q) = \sum_{i=1}^{n} IDF(q_i) \cdot f(q_i, D) \quad (1)$$

Among them: score (D, Q) is the correlation score between document D and query Q, q_i is the i-th word in the query, and f (q_i, D) is the frequency of the word q_i in document D. The calculation method of *IDF* is shown as Eq. (2):

$$IDF(q_i) = \log \frac{N - n(q_i)}{n(q_i) + 0.5} + 1 \quad (2)$$

Here: N is the total number of documents in the document collection, and $n(q_i)$ is the number of documents containing the word q_i.

In order to enhance the Similarity matching ability of massive text vectors, Facebook Artificial Intelligence Similarity Search (FAISS) is used as the vector database, and cosine similarity is adopted to measure the correlation between queries and documents. Among them, the calculation of similarity is shown as Eq. (3):

$$S(Q, K) = \frac{Q \cdot K}{|Q||K|} \quad (3)$$

Among them: Q is the query vector, k is the knowledge vector, and · is the dot product of the vector.

Table 3 Pseudo-code in the retrieval stage

Input: Query q, document set D
Output: The K documents most relevant to query q
① query-keywords = LLM (q) ‖ Use large language models Extract the key words in the query q
② query = q + query_keywords ‖ In the original query Add keywords to expand queries to enhance sparse retrieval
③ sparse_results = retrieve_bm25 (query, D) ‖ Use the BM25 method for sparse retrieval to obtain the documents related to the query
④ q_embed = Embedding (q) ‖ use gte-large The zh model vectorizes q
⑤ dense-results = retrieve_faiss (q-embed, D) ‖ Calculate the distance between q_embed and all document vectors in the vector library Obtain the documents related to the query
⑥ convex_results = result-merge (sparse-results, dense-results) ‖ Obtain from sparse search and dense search respectively Merge the relevant document results

The hybrid retrieval method can fully integrate the advantages of sparse retrieval and dense retrieval, make up for their respective deficiencies, and thereby improve the overall retrieval performance. In the hybrid retrieval method of this study, keyword extraction is adopted to expand the query to enhance sparse retrieval. The documents recalled by sparse retrieval and dense retrieval are processed as a union and used as the candidate documents for hybrid retrieval. The pseudo-algorithm of the entire retrieval stage is shown in Algorithm 1 of Tab. 3. Hybrid retrieval is helpful for obtaining relevant documents, but it cannot effectively re-rank the texts recalled by sparse retrieval and dense retrieval.

After merging the documents and questions output by the resorter, they are input into the text filter for text filtering to extract the key information texts that can answer the questions. The text filter uses a large language model and inputs the relevant documents obtained from queries and retrieves through the prompt engineering method to extract key information. The prompt template for key information extraction is shown in Tab. 4.

Table 4 Key information extraction prompt template

Prompt	Prompt text
Prompt template for extracting key information	Please extract the key information text fragments that can answer the question based on the provided materials and the user's question. If you find the key information text fragments that can answer the question, please directly output the key information text; otherwise, output [the information in the material is insufficient], and do not generate the answer yourself. The following is the material: {Reference passage content}, the user question is: {The question input by the user}

The key information text output by the key information extraction module and the user query are filled into the predefined answer generation template to form prompts, as shown in Tab. 5 specifically, and then input into the large language model. The final answer is generated by using the understanding and reasoning capabilities of the large language model.

Table 5 Answer generation prompt template

Prompt	Prompt text
Answer generation prompt template	You are an expert in reading comprehension and are good at answering questions based on the materials. Please answer users' questions according to the provided materials. Do not repeat the prompt content within the question. The answer should be clear and accurate, including the correct key words. Don't fabricate randomly. The following is the material: {Reference passage content}, the user question is: {The question input by the user}

4 SIMULATION VERIFICATION

To verify the effectiveness of the KEHRAO method proposed in this paper, three experiments were designed in this chapter, and BERT scores (P, R, F1), RougEL, and VRecall@K values were used as the evaluation criteria of the method. Experiment 1 aims to verify the influence of the number of recalled knowledge on the answers generated by the KEHRAO method, and to determine the optimal number of recalled knowledge for subsequent comparative experiments. Experiment 2 aims to verify the feasibility of the KEHRAO method on public question-answering datasets. This experiment uses a government question-answering dataset. Experiment 3 aims to verify the influence of the base language model on the KEHRAO method. In Experiment 2 The methods in this paper are respectively related to the Dense Passage

Experiment 1 verified the influence of the recall quantity on the question-answering results by adjusting the

recall quantity of knowledge, thereby determining the appropriate recall quantity of knowledge. The experimental dataset is the e-government dataset. By changing the recall quantity m of knowledge, the variations of F1 and ROUGe-L values of the KEHRAO method with the recall quantity of knowledge are shown in Fig. 4. It can be known from the results of Experiment 1 that the number of knowledge recalls has a certain impact on the quality of the answers generated by the method. However, too many knowledge vectors may increase the computational burden and may introduce noise information, making it difficult for large language models to give correct answers. On the other hand, too few knowledge vectors may miss some important information, thereby leading to a decline in the quality of the answers. On the e-government question-answering dataset, the performance of the method reaches the best when *m* is 4 and 2 respectively. Therefore, subsequent experiments are all conducted under this setting.

Experiment 2 verifies the feasibility of the method on public datasets. To fully verify the feasibility of the proposed method on public datasets, the e-government dataset and the CMRC2018 dataset were taken as the experimental datasets. Experiment 1 determined that the knowledge recall quantities m of the two datasets were 4 and 2 respectively. The feasibility of the proposed method was verified through comparative experiments with the DPR-LLM method, HyDE method, QRO-LLM method, PARG method and BM25 method. The results of the *P*, *R*, F1 values and VRecall@K values of each method are shown in Tab. 6.

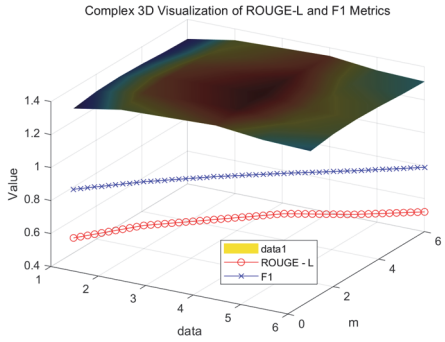

Figure 4 e-government question-answering dataset

Table 6 Comparative experiment results on the e-government question-answering dataset

Method	ROUGE-L	*P*	*R*	F1	VRecall@K
DPR-LLM	0.426	0.729	0.789	0.751	0.453
PARG	0.405	0.726	0.794	0.753	0.448
HyDE	0.433	0.734	0.792	0.757	0.458
QRO-LLM	0.466	0.752	0.803	0.772	0.484
BM25	0.463	0.747	0. 801	0.767	0.481
KEHRAO	0.565	0. 794	0. 843	0.813	0.506

It can be known from Tab. 7 that, compared with the single retrieval method, the KEHRAO method in this paper adopts multi-strategy retrieval. The alignment optimization effect of the KEHRAO method in this paper is better. On the two datasets, the F1 index reaches 0.813 and 0.798 respectively, and the ROUGE-L index reaches 0.565 and 0.581 respectively. Compared with the QRO-LLM method, the ROUGE-L indicator increased by 9.9 percentage points and 2.3 percentage points respectively on the two datasets, and the F1 indicator increased by 4.1 percentage points and 1.7 percentage points respectively.

The reasons why the KEHRAO method in this paper can achieve a better question-answering effect are as follows: Firstly, by extracting the key words in the user's query through the large language model and adopting the query expansion technology, the key word information extracted by the large language model is concatenated into the user's query to increase the word frequency of the key information. When conducting sparse retrieval, the matching of the query documents can focus on corresponding to the key information, thereby achieving a better recall effect. Secondly, hybrid retrieval is adopted to fully utilize the advantages of these two retrieval methods, make up for the respective deficiencies of sparse retrieval and dense retrieval, thereby improving the overall retrieval performance and further enhancing the accuracy of generated answers. Thirdly, use the reordering model to place the most relevant documents at the very beginning of the input of the large model, thereby improving the accuracy of the answers generated by the large language model; Fourth, use a text filter to extract the key information text, effectively eliminating the interference of noise on the answers generated by large language models.

Experiment 3: Impact Analysis of the Base Language Model

Table 7 Performance comparison of different base large language models

Data set	Base LLM	ROUGE-L	*P*	*R*	F1
Question-answering dataset	ChatGLM-6B	0.326	0.657	0.792	0.713
	Qwen-7B-Chat	0.565	0.794	0.843	0.813

To further verify the effectiveness of each component of the method, this chapter continued to conduct two sets of ablation experiments on two datasets. The experimental results are shown in Tab. 8. In the ablation experiment, N-mix indicates that no mixed search was used, only dense search was used, and other conditions were consistent with the KEHRAO method; N-KEY indicates that the query expansion technique is not used to enhance sparse retrieval, and other conditions are consistent with the KEHRAO method; "N-Filter" indicates that the key information extraction module is not used, and other conditions are consistent with the KEHRAO method. It can be known from the experimental results in Tab. 8 that when the mixed search module is removed and only dense search is used during the search, the decline in the search indicators compared with those in the generation stage is relatively large. This might be because most of the questions are precise matching of keywords, which is more suitable for sparse search. When only dense search is used, the search quality decreases, thereby affecting the accuracy of the answers generated by the large model. When the keyword extraction module is removed and the query expansion technology is not used to enhance sparse retrieval, the retrieval metrics and the metrics in the generation stage decrease slightly.

Table 8 Ablation experiment results on the e-government question-answering dataset

Method	ROUGE-L	*P*	*R*	F1	VRecall@K
N-MIX	0.426	0.729	0. 789	0.751	0.751
N-KEY	0.544	0.783	0. 838	0.804	0.804
N-FILTER	0.554	0. 787	0. 840	0.808	0.506
KEHRAO	0.565	0.794	0.843	0.813	0.813

In the real world, the forms of false information are diverse and constantly changing. In order to evaluate the processing ability of the hybrid RAG large language model for unknown categories without any labeled samples, that is, whether the model can make judgments only based on existing knowledge and language understanding. Four comparative experiments with a small number of sample demonstrations, namely 0-shot, 1-shot, 2-shot and 3-shot, were set up, and the experiments were conducted by setting a small number of prompt samples in the large language model. The specific experimental results are shown in Fig. 5.

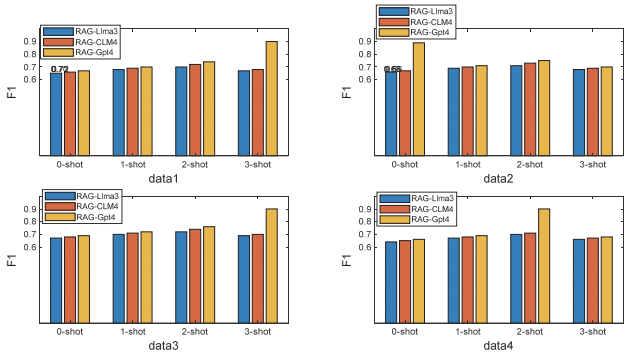

Figure 5 Indicates the influence of quantity on the identification of false information

According to the data analysis shown in Fig. 5, on the four datasets, with the increase in the number of demonstration samples, the prediction performance of the model generally shows an upward trend, especially in the tasks that performed poorly in the initial zero-sample situation, this trend is more obvious. Specifically, in the 0-shot scenario, the F1 scores of the RAG-ChatGLM4 model for the four datasets were 0.782, 0.795, 0.803, and 0.793 respectively. In the 3-shot scenario, the F1 scores of the RAG ChatGLM4 model increased to 0.835, 0.841, 0.832, and 0.827 respectively. Compared with the predicted values in the 0-shot scenario, they increased by 6.78%, 5.79%, 3.61%, and 4.29% respectively. This result indicates that by increasing the demonstration samples, the RAG-ChatGLM4 model can capture the features of false information more effectively, thereby improving the accuracy of detection. Similarly, the RAG-Llama3 and RAG-GPT4 models also observed performance improvements when the number of prompt samples increased. However, it is worth noting that for models with inherently superior performance, such as GPT-4, as the number of prompt samples increases, the extent of their performance improvement gradually decreases. Specifically, compared with the prediction results of 2-shot on the four datasets, the F1 scores of RAG-GPT4 in the 3-shot scenario increased by only 0.94%, 1.04%, 1.26%, and 1.07% respectively.

Furthermore, from the overall trend perspective, as the number of prompt samples increases, the performance differences among different large language models gradually narrow. For example, in the 2-shot scenario, the F1 score of RAG-ChatGLM4 on the PolitiFact and FakeNewsAMT datasets is close to 0.85, and it also shows a high degree of closeness to the prediction results of RAG-Llama3. This phenomenon indicates that in few-shot learning scenarios, by providing a limited number of demonstration samples, large language models can effectively enhance their adaptability and generalization ability in new domains. Although there are initial performance differences among different models, these differences gradually decrease as the number of samples increases.

Given the input text and the list of events it contains, accurately extract the relationships between event pairs and correctly classify their relationship types as causal or temporal. Comparative experiments were also conducted on six types of large language models, and the results of each benchmark experiment are shown in Fig. 6.

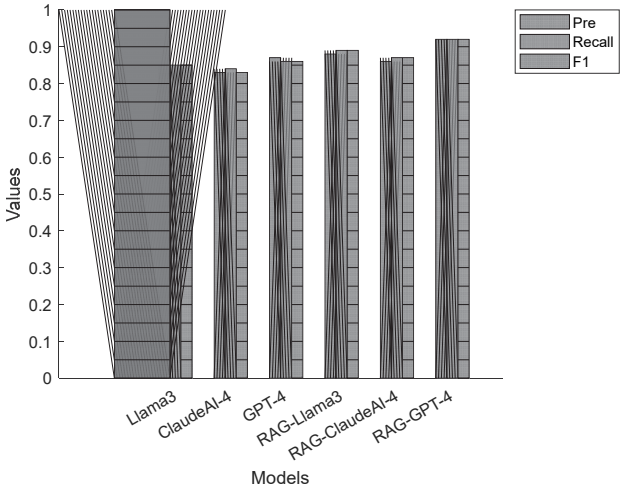

Figure 6 Results of the information extraction comparison experiment

It can be clearly seen from Fig. 6 that in the information extraction task, the three large language models that introduce the hybrid RAG strategy significantly outperform the corresponding models that do not adopt the hybrid RAG strategy in terms of prediction performance. Specifically, without adopting the hybrid RAG strategy, the F1 scores of the Llama3, ChatGLM4, and GPT4 models were 0.854, 0.840, and 0.868, respectively. The RAG-Llama3, RAG-ChatGLM4 and RAG-GPT4 models achieved F1 values of 0.891, 0.870 and 0.918 respectively. Compared with the benchmark model, their performances improved by 4.33%, 3.57% and 5.76% respectively. This is because the hybrid RAG strategy allows the model to retrieve and integrate relevant local knowledge in real time when answering questions, enabling the model to provide more accurate and comprehensive information extraction results when dealing with complex problems in specific domains. Secondly, the hybrid RAG strategy significantly improves the response speed and accuracy of the model to dynamically changing network information by enhancing the model's access ability to the latest knowledge.

As can be seen from Fig. 7, similar to the performance changes in the false information detection task, the increase in the number of prompt samples also brings about performance improvements. Compared with the prediction performance under 0-shot conditions, under 3-shot conditions, the F1 values of the Llama3, ChatGLM4, GPT4, RAG-Llama3, RAG-ChatGLM4, and RAG-GPT4 models increased by 6.21%, 5.48%, 5.65%, 5.17%, and 4.94% respectively And 3.92%. Among them, the F1 value of the RAG-GPT4 model under the 3-shot condition reached

0.954, far exceeding that of other benchmark models. The results show that few-shot prompts have a significant effect on improving the model performance. However, this improvement is not effective in all cases. Specifically, few-shot prompts show a relatively significant performance improvement for the same type of input events, but the prediction performance improvement for different types of events is not obvious and may even decline.

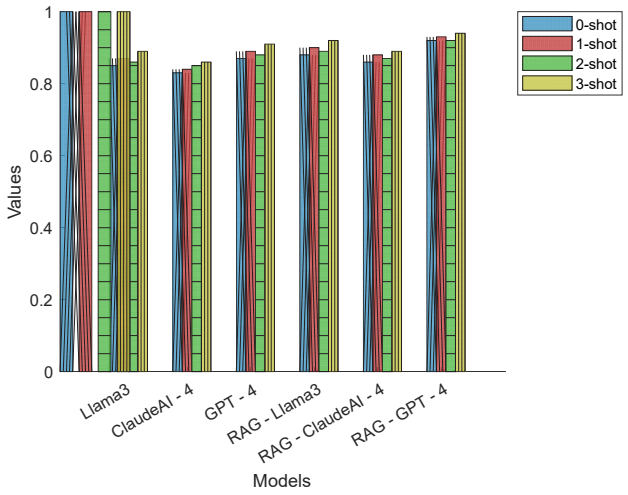

Figure 7 Results of the few-shot experiment for information extraction

In addition, the early detection of hybrid retrieval can enable a deep understanding of public sentiment changes before issues evolve into larger-scale social public opinions, and timely formulation and adjustment of response strategies can be carried out. To this end, this study simulates the process of e-government comments increasing from few to many, and tests the mixed retrieval performance of the model under different data volumes. Fig. 8 shows the prediction accuracy of the model when the quantity varies on the ChnSentiCorp_htl_all, on-line_shopping_10_CATS, and waimai_10k data.

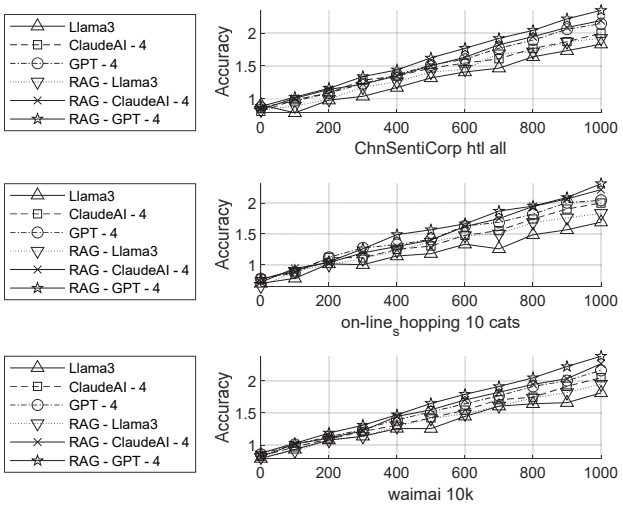

Figure 8 Early discovery results of mixed retrieval

According to the data analysis results in Fig. 8, the large language model based on the hybrid RAG strategy shows obvious advantages in the early analysis of e-government sentiment. When the number of online comments is relatively small, the prediction accuracy of most traditional benchmark models remains at around 0.75, while large language models that introduce the hybrid RAG strategy can achieve an accuracy of 0.775. This difference indicates that the hybrid RAG strategy enhances the model's understanding and reasoning ability in the data scarcity stage by combining the information of the external knowledge base, and guides the model to adapt to the task background by combining the context information and prompt samples, thereby significantly improving the prediction accuracy of the model.

Fig. 9 and Fig. 10 are the PR curves of partially relevant and fully relevant results in the retrieval list, which can more intuitively compare the impact of different strategies on the representation methods of government affairs features. It was found that OPSCM-UC, which used both strategies simultaneously, achieved the best results compared to the other combination of processing methods, with the P@5 metric reaching 0.54 and 0.81 in the fully correlated and partially correlated evaluation of search results. Furthermore, compared with the methods that do not use the two strategies, both OPSCM-UC-UW and OPSCM-UC-UN that separately introduce normalization processing or the weight strategy have improved in formula retrieval performance, and the improvement of OPSCM-UC-UN that uses the weight strategy is more obvious.

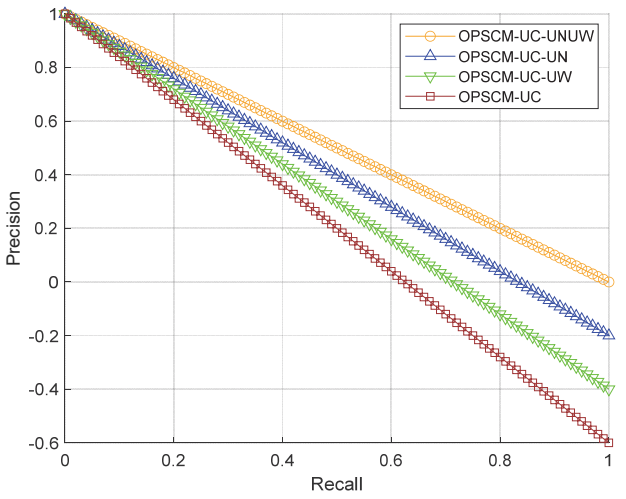

Figure 9 Relevant PR curves in part

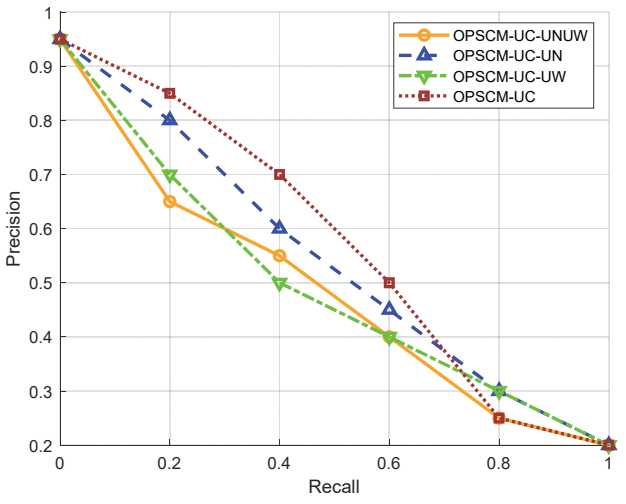

Figure 10 Fully correlated PR curves

In order to demonstrate more intuitively the impact of context semantic supplementation on the results of government affairs retrieval, we randomly selected 5 pieces of data within each topic and visualized the topic distribution of the top 500 retrieval results. Fig. 11 and Fig. 12 respectively show the topic distribution of the retrieval results of OPSCM-UC and OPSCM. The horizontal axis represents the topic to which the formula to be queried belongs, and the vertical axis represents the distribution of the topics through different colors. Among them, the part highlighted by the box is the total number of retrieval results corresponding to the current query topic.

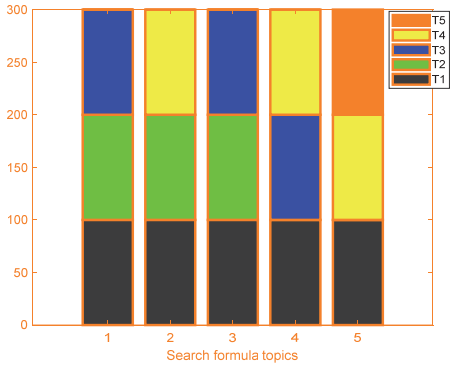

Figure 11 Topic distribution corresponding to the top 500 search results of OPSCM-UC

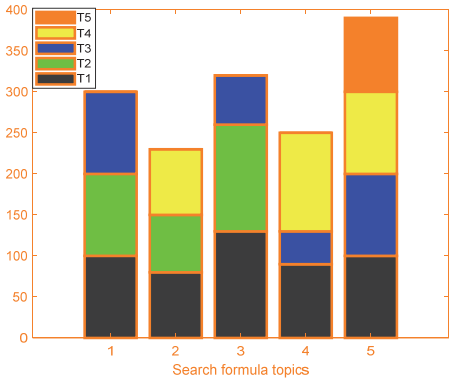

Figure 12 Topic distribution corresponding to the top 500 search results of OPSCM

With the increase in the number of comments, the prediction performance of all models has improved. This is because more text data provides the model with richer context information, enabling it to learn more complex emotional features and semantic patterns. For example, after obtaining more than 2500 comment data, the prediction accuracy of RAG-GPT4 increased to approximately 0.875. Meanwhile, as the amount of data further increases, although the improvement rate of the model performance gradually slows down, the introduction of the hybrid RAG strategy keeps the model capability at an optimal level all the time. This is because when the data volume reaches a certain scale, the model has been able to fully capture most of the emotional features. Therefore, with the help of additional data, its performance improvement tends to be stable. Secondly, when the data volume reaches a certain threshold, the marginal utility of the newly added data decreases, and the model no longer significantly relies on additional data input to improve its accuracy. In conclusion, the large language model based on the hybrid RAG strategy has significant performance advantages in the hybrid retrieval of e-government, especially in the early stage when the data volume is insufficient.

5 CONCLUSION

This study proposes a large language model improvement framework with a hybrid RAG strategy. This method retrifies high-quality target domain knowledge by constructing a RAG knowledge base and combines context information and prompt samples as input content. The effectiveness of the proposed method has been verified in multiple downstream tasks of e-government analysis. A domain question-answering alignment optimization method based on keyword extraction and hybrid retrieval is proposed. Through prompt engineering, the information extraction ability of large language models is fully utilized. Query expansion technology and hybrid retrieval strategies are adopted to retrieve relevant texts, and the retrieved relevant texts are reordered and filtered. The experimental results show that the KEHRAO method proposed in this paper has significant advantages in both F1 and ROUGE-L indicators. In addition, the method proposed in this paper achieves full-process adaptability without human intervention, optimizes the training and deployment costs of domain question answering based on large language models, improves the accuracy of domain question answering tasks based on retrieval enhancement generation technology, and has strong practicability. In the future, the development of large language models will also need to take into account issues such as the balance between model scale and economic cost, as well as technological ethics, and continuously improve the performance and usability of the models. Meanwhile, the application scenarios of large language models in the field of government affairs will continue to be enriched, and their application scope will also keep expanding. This will also provide more ideas for the iterative update of large language models.

Acknowledgments

The research was supported by the Sichuan Provincial Center for Educational Informatization Application and Development Research (No. JYXX22-002, "Research on Intelligent Monitoring of University Digital Content Security Based on Deep Learning"); the Guangyuan Federation of Social Sciences (No. GYSD23YB015, "Calligraphy Art Excavation and Digital Research of Jianmen Ancient Shudao"); and the Sichuan Provincial Key Research Base of Humanities and Social Sciences, Sichuan Research Center for Public Security and Social Governance Innovation (No. SCZA22A01, "Intelligent Governance of Internet Content Security Based on Big Data in the New Era").

6 REFERENCES

[1] He, C., He, W., & Liu, M. (2025). Enriched Construction Regulation Inquiry Responses: A Hybrid Search Approach for Large Language Models. *Journal of Management in Engineering*, *41*(3), 6444-6457.

https://doi.org/10.1061/JMENEA.MEENG-6444
[2] Ray, P. P. (2024). Timely need for navigating the potential and downsides of LLMs in healthcare and biomedicine. *Briefings in Bioinformatics*, *25*(3), 1-14. https://doi.org/10.1093/bib/bbae214
[3] Song, Y., Fan, H., & Liu, J. (2025). A goal-oriented document-grounded dialogue based on evidence generation. *Data & Knowledge Engineering*, 155, 102378. https://doi.org/10.1016/j.datak.2024.102378
[4] Wu, J., Jiang, M., & Fan, J. (2025). Arch-Eval benchmark for assessing chinese architectural domain knowledge in large language models. *Scientific Reports*, *15*, 13485. https://doi.org/10.1038/s41598-025-98236-0
[5] Sequeda, J., Allemang, D., & Jacob, B. (2025). Knowledge Graphs as a source of trust for LLM-powered enterprise question answering. *Journal of Web Semantics*, *85*, 100858. https://doi.org/10.1016/j.websem.2024.100858
[6] Iaroshev, I., Pillai, R., & Vaglietti, L. (2024). Evaluating Retrieval-Augmented Generation Models for Financial Report Question and Answering. *Applied Sciences*, *14*(20), 9318. https://doi.org/10.3390/app14209318
[7] Yu, J. (2025). SDD-LawLLM: Advancing Intelligent Legal Systems Through Synthetic Data-Driven Fine-Tuning of Large Language Models. *Electronics*, *14*, 742. https://doi.org/10.3390/electronics14040742
[8] Chen, A., Tian, Y., & Zhang, J. (2025). LLM-based intelligent Q&A system for railway locomotive maintenance standardization. *Scientific Reports*, *15*(1), 12953. https://doi.org/10.1038/s41598-025-96130-3
[9] Yang, J. (2024). An Enhanced Retrieval Scheme for a Large Language Model with a Joint Strategy of Probabilistic Relevance and Semantic Association in the Vertical Domain. *Applied Sciences*, *14*, 11529. https://doi.org/10.3390/app142411529
[10] Loevenich, J., Adler, E., Hürten, T., & Lopes, R., R., F. (2025). Design and evaluation of an Autonomous Cyber Defence agent using DRL and an augmented LLM. *Computer Networks*, *262*, 111162. https://doi.org/10.1016/j.comnet.2025.111162
[11] Irkovi, S., Mladenovi, V., & Tomi, S. (2025). Utilizing Fine-Tuning of Large Language Models for Generating Synthetic Payloads: Enhancing Web Application Cybersecurity through Innovative Penetration Testing Techniques. *Computers, Materials & Continua*, *82*(3), 4409-4430. https://doi.org/10.32604/cmc.2025.059696
[12] Yan, Z. & Xu, Y. (2024). Real-Time Optimal Power Flow With Linguistic Stipulations: Integrating GPT-Agent and Deep Reinforcement Learning. *Power Systems*, *39*(2), 4747-4750. https://doi.org/10.1109/TPWRS.2023.3338961
[13] Niewiadomska-Szynkiewicz, E. (2024). Large Language Models and the Elliott Wave Principle: A Multi-Agent Deep Learning Approach to Big Data Analysis in Financial Markets. *Applied Sciences*, *14*(24), 11897-906. https://doi.org/10.3390/app142411897
[14] Wan, Y., Chen, Z., & Liu, Y. (2025). Empowering LLMs by hybrid retrieval-augmented generation for domain-centric Q&A in smart manufacturing. *Advanced Engineering Informatics*, *65*, 103212. https://doi.org/10.1016/j.aei.2025.103212
[15] Ge, J., Sun, S., & Owens, J. (2024). Development of a liver disease-specific large language model chat interface using retrieval-augmented generation. *Hepatology*, *80*(5), 1158-1168. https://doi.org/10.1097/HEP.0000000000000834
[16] Delleani, M., D, Amico, S., & Sauta, E. (2024). The "David Vs Goliath" Study: Application of Large Language Models (LLM) for Automatic Medical Information Retrieval from Multiple Data Sources to Accelerate Clinical and Translational Research in Hematology. *Blood*, *144*(Supplement 1), 3597. https://doi.org/10.1182/blood-2024-205621
[17] Castellanos-Nieves, D. & García-Forte, L. (2025). Human-Centered AI for Migrant Integration Through LLM and RAG Optimization. *Applied Sciences*, *15*(1), 325. https://doi.org/10.3390/app15010325
[18] Han, B., Susnjak, T., & Mathrani, A. (2024). Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview. *Applied Sciences* , *14*(19), 9103. https://doi.org/10.3390/app14199103
[19] Castellanos-Nieves, D. & García-Forte, L. (2025). Human-Centered AI for Migrant Integration Through LLM and RAG Optimization. *APPLIED SCIENCES*, *15*(1), 325. https://doi.org/10.3390/app15010325
[20] Gokcimen, T. & Das, B. (2025). A novel system for strengthening security in large language models against hallucination and injection attacks with effective strategies. *Alexandria Engineering Journal*, *123*, 71-90. https://doi.org/10.1016/j.aej.2025.03.030
[21] Kumar, J. (2023). Large Language Models for Text Summarization: A Comprehensive Study. *Pranjana: The Journal of Management Awareness*, *26*(1/2):113-124. https://doi.org/10.5958/0974-0945.2023.00011.0
[22] Zhang, Y., Ren, S., & Wang, J. (2025). Aligning Large Language Models with Humans: A Comprehensive Survey of ChatGPT's Aptitude in Pharmacology. *Drugs*, *85*(2), 231-254. https://doi.org/10.1007/s40265-024-02124-2
[23] Xu, X., Yao, B., & Dong, Y. (2024). Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *8*(1), 31. https://doi.org/10.1145/3643540
[24] Thanasi-Boe, M. & Hoxha, J. (2024). From ideas to ventures: building entrepreneurship knowledge with LLM, prompt engineering, and conversational agents. *Education and Information Technologies*, *29*(18), 24309-24365. https://doi.org/10.1007/s10639-024-12775-z
[25] Caffaro, F. & Rizzo, G. (2024). Knowledge-Enhanced Conversational Agents. *Journal of Computer Science and Technology*, *39*(3), 585-609. https://doi.org/10.1007/s11390-024-2883-4
[26] Yang, F, C., Duque, K., & Mousas, C. (2024). The Effects of Depth of Knowledge of a Virtual Agent. *Visualization and Computer Graphics*, *30*(11), 7140-7151. https://doi.org/10.1109/TVCG.2024.3456148
[27] Vitale, M., Youssef, A., & Mishra, P. (2024). Harnessing Generative AI for Interactive System Failure Diagnostics: A User-Centric Approach to Streamlined Problem Solving and Maintenance. *SPE-Society of Petroleum Engineer*, *3*, 8-33. https://doi.org/10.2118/222033-MS
[28] Wan, Y., Chen, Z., & Liu Y. (2025). Empowering LLMs by hybrid retrieval-augmented generation for domain-centric Q&A in smart manufacturing. *Advanced Engineering Informatics*, *65*, 103212. https://doi.org/10.1016/j.aei.2025.103212

Contact information:

Huaiyu WEN
(Corresponding author)
College of Computer Science,
Chengdu University, Sichuan Chengdu, 610106, China
E-mail: wenhuaiyu@163.com

Mengxuan HE
College of Computer Science,
Chengdu University, Sichuan Chengdu, 610106, China