ISSN 2616-5775 Vol. 9, Issue 4: 103-109, DOI: 10.25236/AJCIS.2026.090413

Coronary Heart Disease Medical Technology Innovation Knowledge Graph-Enhanced Large Model Intelligent Retrieval QA System

Shuo Wang[1,a], Min Wang[1,b,*], Haotian Wu[1,c], Hongyang Qi[1,d], Yuhang Zhou[1,e], Haozhe Li[1,f]

[1]Information School, Beijing City University, Beijing, China
[a]wang1350789090@163.com, [b]wangmin@bcu.edu.cn, [c]15910670460@163.com,
[d]18801204542@163.com, [e]ljwt3306@163.com, [f]lhzhym2004@163.com
**Corresponding author*

Abstract: *This paper designs and implements a large model intelligent retrieval question answering system enhanced by a knowledge graph, focusing on the field of coronary heart disease medical technology innovation. The system constructs a professional knowledge graph from public patent data and enhances a large language model using RAG (Retrieval-Augmented Generation) technology, effectively improving the accuracy and credibility of domain-specific question answering and alleviating model hallucination. The front end uses HTML, CSS, and JavaScript to build the user interface, while the back end is based on the Python Flask framework. Neo4j is used to store the knowledge graph, SQLite manages user data, and a locally deployed Ollama (qwen3:8b model) is integrated to support multi-mode intelligent question answering. This paper details the system design, key technologies, implementation process, and test results, verifying the system's effectiveness and practicality in coronary heart disease medical knowledge retrieval and intelligent QA.*

Keywords: *Knowledge Graph; Large Language Model; RAG; Coronary Heart Disease; Neo4j; Intelligent Question Answering*

1. Introduction

Coronary heart disease (CHD), as a common and severe cardiovascular disease, has always been a key research area in medicine, with continuous innovations and patent achievements emerging. Traditional large model intelligent QA systems perform well in general domains, but when faced with highly specialized medical questions, their generated results often lack accuracy and may produce "hallucinations," making it difficult to meet the precise information needs of medical researchers. In recent years, the integration of knowledge graphs (KGs) and large language models (LLMs) has become a key technical path to enhance professional domain QA. RAG (Retrieval-Augmented Generation) technology effectively improves the accuracy and credibility of LLM-generated answers by retrieving relevant information from external knowledge sources as context.

This paper aims to design and implement a large model intelligent retrieval QA system enhanced by a CHD medical technology innovation knowledge graph. By constructing a CHD patent knowledge graph and combining it with LLM capabilities using RAG technology, the system aims to provide medical researchers with a tool to quickly and accurately obtain cutting-edge CHD scientific and technological information, thereby enhancing research efficiency and knowledge sharing.

2. Research Background

In recent years, artificial intelligence and machine learning have developed rapidly, especially in natural language processing (NLP) and knowledge acquisition technologies [1]. With the rapid development of AI, existing general-purpose generative AI has introduced more domain-specific knowledge and data to enhance its cross-domain capabilities [2]. In the application of combining large language models (LLMs) and knowledge graphs (KGs), domestic research institutions have successively released open-source vision-language large models, including InternVL, Qwen-VL, CogVLM, etc., all achieving high scores in various multi-modal standard tests, comparable to

ISSN 2616-5775 Vol. 9, Issue 4: 103-109, DOI: 10.25236/AJCIS.2026.090413

international vision models such as GPT-4o and Gemini [3]. Domestic research institutions and enterprises have extensively explored the integration of LLMs and KGs, mainly focusing on knowledge-enhanced pre-trained language models (KEPLMs), knowledge graph construction and updating, multi-modal KG and LLM integration, and KG-enhanced intelligent QA systems. Professor Zhengren Li from Beijing University of Posts and Telecommunications proposed a telecom customer complaint adjudication method based on Linked-RAG and LLM, achieving precise and efficient complaint adjudication through hierarchical semantic decoupling and a dynamic historical knowledge retrieval mechanism. A review by a Peking University team systematically introduced the application of RAG in AI-generated content (AIGC). Domestic research emphasizes the Chinese language environment and application deployment. Tsinghua University has significant influence in domain-specific KG construction and reasoning rule mining, while Baidu and Alibaba have implemented intelligent search and QA services based on KGs [4]. However, existing systems still face challenges such as inaccurate intent recognition and low credibility of generated results in professional medical research applications. Based on the study of the domestic research background and multiple papers and reviews, this research decided to build an intelligent QA system enhanced by a KG and RAG for the CHD medical research domain. By crawling the latest CHD patent data and constructing a KG for visualization, the KG is used to enhance a general LLM in RAG mode to compensate for the shortcomings of existing general LLMs, providing ideas for further system development.

After 2024, international research in NLP has shown a distinct trend of deep integration and collaborative enhancement [5]. Researchers are no longer satisfied with simply combining LLMs and KGs but are committed to building more refined fusion architectures and dynamic interaction mechanisms. For example, Graph Language Models (GLM) directly process graph structure information by modifying the Transformer architecture to achieve joint reasoning. New frameworks aim to combine complex machine learning models (e.g., random forest, LSTM, and Transformer-based models) with RAG to better capture context and enable adaptive decision-making [6]. The Collaborative Enhancement Framework (CogMG) promotes bidirectional knowledge flow and dynamic updates between LLMs and KGs, effectively compensating for the shortcomings of static knowledge. Meanwhile, in improving model proactivity and reasoning ability, methods based on Bayesian Experimental Design (BED-LLM) significantly optimize the information acquisition efficiency of multi-turn QA. These technological advances have been effectively validated in professional fields such as healthcare, with research showing that strategies like few-shot learning can greatly reduce model "hallucination" and improve the credibility of tasks such as named entity recognition. Determining the native language of a document is called language identification [7]. The continuous evolution of foundational tools (such as the next-generation GloVe word vectors) also supports more accurate language understanding. Overall, international cutting-edge research is driving LLMs towards more reliable, efficient, and continuously learning professional applications through architectural innovation, collaborative mechanisms, and domain optimization.

3. Requirements Analysis of the CHD Medical Technology Innovation Knowledge Graph-Enhanced Large Model Intelligent Retrieval QA System

3.1. System Functional Requirements

First, by using efficient crawling technology to systematically search and crawl national patent databases, patent documents related to CHD are obtained. These documents cover drug development information, treatment methods, medical devices, and new technologies for CHD prevention. Through deep text analysis and processing of these patent documents, key terms and concepts are extracted to form a professional CHD-related thesaurus. This thesaurus serves as the basis for subsequent processing, containing not only professional medical terms but also key information such as drugs and treatment methods related to CHD. The thesaurus information is then converted into a knowledge graph, and VectorGraphRAG technology is used to enhance the large model intelligent QA retrieval system. The functions are as follows:

Admin user management: To enable administrators to manually add, delete, modify, and query system users through the front-end page, ensuring system management. First, an SQLite database is designed with a user table containing unique identifier (ID), username, hashed password, and user role fields. Administrators can view all user information through a specific interface, using an HTML table to display basic user information and provide corresponding operation buttons. The entire system also focuses on security, preventing unauthorized access through password hashing and input validation.

ISSN 2616-5775 Vol. 9, Issue 4: 103-109, DOI: 10.25236/AJCIS.2026.090413

Admin and user use of large model intelligent retrieval QA: Intelligent retrieval QA is one of the cores of the system, allowing administrators and users to interact with the locally deployed large model system by entering questions into a chat box. The front-end page has a chat box where users can enter questions related to their needs, and the system automatically calls the local qwen3:8b large model to process the user's query in real time. Three different modes are provided: Vector Retrieval-Augmented Generation (RAG) system (model1.py), AC automaton-based entity recognition system (model2.py), and local large model QA (model3.py).

3.2. System Non-Functional Requirements

The client uses VSCode as the development launcher and accesses the Flask-based backend service through a browser. The front end uses HTML+CSS+JavaScript to build the interactive interface. Flask, as a lightweight Python web framework, is responsible for receiving HTTP requests, processing business logic (such as calling model1.py for vector retrieval or model2.py for entity recognition), and returning JSON-formatted responses.

4. High-Level Design of the CHD Medical Technology Innovation Knowledge Graph-Enhanced Large Model Intelligent Retrieval QA System

The system architecture adopts a layered design, mainly including the data collection layer, data processing layer, data storage layer, and application layer.

Data Collection Layer: Responsible for crawling CHD-related patent information from public data sources such as the China National Intellectual Property Administration. Python and Chromedriver are used for web data crawling, with appropriate delays to avoid anti-crawling mechanisms.

Business Support Layer: Preprocesses the collected data, performs tokenization, and deep learning model analysis. The jieba library in Python is mainly used for tokenization, and the BERT model is used for text analysis to extract key terms and concepts, perform named entity recognition (NER), and finally build the knowledge graph.

Data Storage Layer: Uses a hybrid storage solution, including SQLite for user data and Neo4j for storing and managing the knowledge graph data.

Backend System Layer: The backend system layer is the core processing engine of the system, responsible for receiving front-end requests, coordinating various AI models and data processing workflows, and returning results to users. It intelligently schedules different functional modules to achieve end-to-end processing from question understanding to knowledge retrieval, ensuring efficient and accurate responses to user queries while maintaining system stability and scalability.

Application Layer: Provides user interaction interfaces, supporting functions such as data query, visualization display, knowledge graph browsing, and large model intelligent QA. Front-end technologies (such as HTML, CSS, JavaScript) are used to build the user interface, interacting with the backend through APIs.

5. System Testing of the CHD Medical Technology Innovation Knowledge Graph-Enhanced Large Model Intelligent Retrieval QA System

5.1. System Functional Testing

5.1.1. Login and Registration Module Test

The Registration and Login module serves as the system's entry point, responsible for user identity authentication and account management. The purpose of testing this module is to verify that users can successfully complete registration and login operations, and to ensure proper permission allocation for users with different roles. The tests simulate real-world usage scenarios for both regular users and administrators, sequentially performing registration and login procedures as detailed in Table 1.

ISSN 2616-5775 Vol. 9, Issue 4: 103-109, DOI: 10.25236/AJCIS.2026.090413

Table 1: Registering a Regular User

Test Content	Input	Expected Output	Actual Output	Test Result
Register Regular User	Username: 1234, Password: 1234, Confirm: 1234	Registration successful, please log in	Registration successful, please log in	Pass
Login Regular User	Username:1234, Password: 1234	Login successful, redirect to home page	Login successful, redirect to home page	Pass
Login Admin	Username: admin, Password: admin123	Login successful, redirect to home page	Login successful, redirect to home page	Pass

5.1.2. User Management Module Test

The test simulates the operational behaviors of users with different roles, sequentially conducting permission verification and user information operations (addition, deletion, modification, and query). The test details are presented in Table 2.

Table 2: User Management Module Test

Test Content	Input	Expected Output	Actual Output	Test Result
User Management (Regular User)	None	No access permission. Only admin can access user management page.	No access permission. Only admin can access user management page.	Pass
User Management (Admin)	None	Redirect to user management	Redirect to user management	Pass
Add User (Regular User)	Username: 12345, Password: 12345, Role: Regular User	User information added successfully	User information added successfully	Pass
Add User (Admin)	Username: master, Password: master, Role: Admin	User information added successfully	User information added successfully	Pass
Edit User (12345)	Username: 123456, Role: Regular User	User information edited successfully	User information edited successfully	Pass
Delete User (123456)	None	User information deleted successfully	User information deleted successfully	Pass

5.1.3. Knowledge Graph Module Test

The Knowledge Graph Module serves as the core data layer of the system, responsible for the structured storage and visual presentation of medical knowledge on coronary heart disease. The purpose of this module's testing is to verify that users can view node information in the knowledge graph through an interactive interface and quickly locate specific entities and their associated relationships via keyword searches. The test simulates actual user query behaviors by sequentially performing node information viewing and keyword search operations, as detailed in Table 3.

Table 3: Knowledge Graph Management Module Test

Test Content	Input	Expected Output	Actual Output	Test Result
View Node Information	None	Hover mouse over node to view node information	Hover mouse over node to view node information	Pass
Retrieve Node Information	Keyword: Rong Daoming	Information about companies and fields related to Mr. Rong Daoming	Information about companies and fields related to Mr. Rong Daoming	Pass

ISSN 2616-5775 Vol. 9, Issue 4: 103-109, DOI: 10.25236/AJCIS.2026.090413

5.1.4. Large Model Intelligent Retrieval QA Module Test

The large-scale model-based intelligent retrieval and Q&A module serves as the core functional component of this system, offering three distinct Q&A modes: Model 1 (intelligent vector retrieval), Model 2 (knowledge graph relationship inference), and Model 3 (direct Q&A using local large-scale models). The purpose of this module's testing is to evaluate the response quality and accuracy of the three modes across various scenarios, with particular emphasis on demonstrating how knowledge graph augmentation enhances model performance. The tests involve invoking each mode with the same query and comparing their output results, as detailed in Table 4.

Table 4: Large Model Intelligent Retrieval QA Module Test

Test Content	Input	Expected Output	Actual Output	Test Result
Intelligent Vector Retrieval	Question: Rong Daoming	Information about companies and fields related to Rong Daoming (including analysis results)	Information about companies and fields related to Rong Daoming (including analysis results)	Pass
Graph Relationship Reasoning	Question: Rong Daoming	Information about companies and fields related to Rong Daoming (only triple information)	Information about companies and fields related to Rong Daoming (only triple information)	Pass
Large Model	Question: Rong Daoming	Unable to obtain relevant information about Rong Daoming	Unable to obtain relevant information about Rong Daoming	Pass

5.2. System Performance Testing

This performance test evaluated the CHD knowledge graph-enhanced large model intelligent retrieval QA system. Two tests were conducted by inputting 19 different author names and measuring the response time of the large model outputting their academic achievements related to CHD. The goal was to assess the system's performance in real-world applications to ensure it can quickly and accurately respond to user retrieval needs.

5.2.1. Environmental Information

Hardware Environment:

Graphics Card: NVIDIA GeForce RTX 4070 (8G)

Database: Neo4j, SQLite

Large Model: Qwen3:8b

5.2.2. System Test Results

1) Model1

A total of 19 data sets were entered. From these data, it can be inferred that the average response time is 32.74 seconds. Most response times are concentrated between 28-31 seconds, but there are two larger outliers (55 and 61 seconds), possibly due to the higher complexity of the name or resource bottlenecks during processing. The standard deviation of response time is 11.11, indicating significant dispersion, as detailed in Table 5.

Table 5: Large Model Intelligent Retrieval QA Module Performance Test – Model1

count	19
mean	32.74
std	11.11
min	27
25%	28
50%	30
75%	31
max	61

2) Model2

ISSN 2616-5775 Vol. 9, Issue 4: 103-109, DOI: 10.25236/AJCIS.2026.090413

A total of 19 data sets were entered. From these data, it can be inferred that the average response time is 14.84 seconds, significantly lower than Model1. The data dispersion is small, with a standard deviation of only 2.32, indicating stable response times. Most response times are between 13-17 seconds, with a small range between max and min, demonstrating excellent overall performance, as detailed in Table 6.

Table 6: Large Model Intelligent Retrieval QA Module Performance Test – Model2

count	19
mean	14.84
std	2.32
min	12
25%	13
50%	15
75%	17
max	18

5.3. Test Conclusion

Both functional and performance tests met the expected requirements.

In functional testing, each module operated normally according to design specifications. The login and registration module tests showed that users could successfully register and log in with a smooth process and good user experience. The user management module tests also demonstrated system stability, with strict permission control for regular users and administrators, and normal CRUD operations. The knowledge graph module tests verified the correct display of node information and the effectiveness of the retrieval function, ensuring users could accurately obtain relevant information. The large model intelligent retrieval QA module performed well, quickly responding to user questions and returning accurate information and analysis results.

In performance testing, the response time difference between Model1 and Model2 was significant. Model2's average response time was only 14.84 seconds, indicating significantly better performance than Model1. This suggests that in subsequent development, system performance can be further optimized to improve user experience. Although Model1 experienced response delays in some scenarios, this provides potential stress test data for high-load situations, guiding future optimization and adjustments.

6. Conclusion and Future Work

6.1. Conclusion

The project provides medical researchers with a convenient and efficient platform for knowledge sharing and collaboration. Researchers can quickly understand the latest research progress and technological innovations related to CHD through the platform, promoting knowledge exchange and cooperation among different research teams, thereby advancing CHD research and clinical application. Through this systematic knowledge graph, the project aims to improve scientific decision-making ability in CHD research and treatment, supporting innovation and breakthroughs in the medical field.

6.2. Future research directions and prospects

Systematically fine-tune the large model to enhance its capabilities in the CHD technology innovation domain. Further optimize the RAG process and vector index structure to shorten response time and improve user experience.

Acknowledgement

This paper is supported by Beijing City University Student Innovation and Entrepreneurship Training Program Project.

ISSN 2616-5775 Vol. 9, Issue 4: 103-109, DOI: 10.25236/AJCIS.2026.090413

References

[1] Chen Binjie, Fang Zhongding. Design of a Mathematical Problem Solving System Based on Large Model and RAG. China Science and Technology Information, 2025, (17): 105-108.
[2] Qin Yifan, Fu Xiang, Zhang Zhixing, et al. Key Technologies of Coal Mine Professional Large Model Application Based on LoRA Fine-Tuning and RAG Fusion. Industry and Mine Automation, 1-10 [2025-09-05].
[3] Ji Xiaofei, Xue Huajie, Jiang Jinhai, et al. Application of Visual Large Models in Port Weed Identification. China Port Science and Technology, 2025, 7(08): 76-82.
[4] Xu Siyuan. Research and Application of Rice Pest and Disease Diagnosis Method Based on Knowledge Graph. Master's Thesis, Heilongjiang Bayi Agricultural University, 2025.
[5] Randolph C, Michaleas A, Ricke O D. Large language models for closed-library multi-document query, test generation, and evaluation. Frontiers in Artificial Intelligence, 2025, 8: 1592013.
[6] Revar A, Bhesaniya B, Wandra K. RAG-Enhanced Spam Detection Framework: A Synergy of Cloud Scalability and Adaptive AI Models. SN Computer Science, 2025, 6(5): 516.
[7] Vipin J, Lata K K. Improved word vector space with ensemble deep learning model for language identification. Sādhanā, 2024, 49(2).