Social Medicine and Health Management (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/socmhm.2024.050208
ISSN 2616-2210 Vol. 5 Num. 2

# *Integrated Application of LLM Model and Knowledge Graph in Medical Text Mining and Knowledge Extraction*

**Jinzhu Yang***

*Dyania Health Inc, 525 Washington Blvd Suite 300, Jersey City, 07310, NJ, USA*
*jinzhu.yang0625@yahoo.com*
**Corresponding author*

***Keywords:*** LLM, Knowledge Base Q&A System, Professional Knowledge Background, Medical Knowledge Services, Triple Data Structures, Model Output Validation

***Abstract:*** This article proposes an innovative comprehensive framework that deeply integrates Large Language Models (LLM) with Knowledge Graphs (KG) to meet the urgent need for high-quality professional knowledge in medical question answering systems. We have fully utilized the triplet data structure of the knowledge graph, which effectively enhances the professional knowledge foundation of LLM in the medical field and significantly enhances its explanatory power. By accurately aligning the output of LLM with relevant information in KG, this method achieves dual verification and enhancement of model output accuracy and consistency, greatly improving the security and reliability of medical question answering systems. The experimental results show that the method proposed in this article exhibits significant advantages in accuracy and reliability compared to traditional knowledge base question answering (KBQA) systems and single LLM methods. This achievement provides a more efficient and accurate solution for the field of medical knowledge services, and this study also demonstrates the enormous potential and prospects of integrating LLM and KG in medical text mining and knowledge extraction.

## 1. Introduction

In recent years, the combined application of large language model (LLM) and knowledge graph (KG) has aroused wide attention, and they also have a certain position in the medical field. Medical industry as one of the key fields of artificial intelligence technology application, medical question answering system is the core tool to provide medical information and knowledge services, it is currently facing the challenge of understanding and explaining complex medical concepts. When dealing with professional problems in the medical field, the traditional question-and-answer system based on knowledge base has the problems of incomplete information and limited reasoning ability. The LLM has a strong understanding of language, but it lacks a deep understanding of what is actually happening in the medical field. Therefore, how to effectively integrate structured medical knowledge in LLM and knowledge graph has become the key to improve the accuracy and credibility of medical intelligent question answering system. This paper aims to explore and propose a new method to improve the performance of medical question answering system in information retrieval and answer reasoning by combining the advantages of LLM in semantic understanding and detailed entity relationship in knowledge graph, so as to open up a new research direction for intelligent application in medical field.

## 2. Relevant Literature

Large language models such as ChatGPT perform well in natural language question answering, but have limited application in specific domains and high training and deployment costs. Taking traditional Chinese medicine prescriptions as an example, this study designed a question and answer system combining LLM and knowledge graph, which can accurately answer domain questions, generate professional answers, and deploy them quickly. By enhancing the information extraction capability of LLM, the system converts natural language answers into structured knowledge and verifies them through knowledge graph. [1] LLM has attracted wide attention in the field of natural language processing, and has certain applications in finance, medical treatment, education and other fields. However, LLM has the problems of insufficient interpretation, delayed knowledge updating and generating false information. The researchers combined the Knowledge graph with the LLM to address these challenges and improve its explanatory and reasoning capabilities [2].

By simulating the human-computer interaction process, AIChain breaks down the user's interaction with a large language model (LLM) into multiple steps, and utilizes logically connected AI and non-AI units to effectively guide the LLM to generate natural language and rule language prompt to clarify queries and recommend applicable apis [3].

The application of inference to knowledge graphs is aimed at supplementing missing triples, refining structured knowledge, and supporting a variety of subsequent tasks. The inference method based on rule extraction provides a more explainable reasoning method by extracting first-order logic rules. In order to overcome the difference between discrete symbol space and continuous embedded space, the DRaM method combines a large-scale pre-trained language model and fuses discrete rules with continuous vector space. Experiments on multiple knowledge graph datasets show that DRaM performs well in link prediction tasks, especially for the best inference results under the Hits@10 metric. This method can not only solve the challenge of inference effectively, but also extract the first-order logic rules with confidence, which improves the interpretability and inference effect [4].

This technology introduces a method to build a mold professional question and answer system based on LLM model. The system combines the advantages of retrieval and generative question answering, solves the challenge of complex problem processing, and reduces the occurrence of model illusion. In order to cope with the demand for high-quality data in the mold field, LoRA technology is used to solve the problem of high hardware requirements. Methods Combining knowledge base and fine-tuning large language model, ChatGPT was used to deal with the data requirements in the field of mold, unstructured data was processed through training and fine-tuning, and finally a professional question and answer model was constructed [5].

With the intensification of global environmental problems, low-carbon power market has attracted much attention. This paper analyzes the current situation and challenges of its development, and discusses the potential of large language model (LLMs) in promoting its development. Through fine-tuning, prompt design, and semantic embedding, LLMs shows potential for power supply structure adjustment, power demand forecasting, and risk warning. The agent and chain of thought approach based on LLMs can solve complex problems, facilitate the construction of low-carbon power market, and support the low-carbon transition and "dual carbon" goal of China's power system. However, it is necessary to pay attention to the application limitations and potential risks of LLMs to ensure its safe and orderly development [6].

With the rise of large language model (LLM) technology, AI bots like ChatGPT are widely used. Despite the multiple security mechanisms built into these bots, attackers can still bypass the protections and use them to generate phishing emails for cyberattacks. Therefore, how to recognize AI-generated text has become a hot research topic. We collected question and answer data from social platforms and ChatGPT, proposed several detection strategies: online similarity analysis, offline statistical difference analysis, adversarial generation analysis and fine-tuned LLM model classifier, and compared their detection effects. At the same time, from the point of view of network attack and defense, it puts forward some techniques to combat AI text detection [7].

Varatharajah, Y explored the application of reinforcement learning based human-machine collaborative recommendation systems in COVID-19 clinical management, emphasizing the integration of multi-dimensional information through LLI models and knowledge graphs to predict disease severity and complication risk, in order to guide individualized clinical decision-making [8]. Chen, H's study analyzed clinical data of 10123 COVID-19 patients and proposed a staging and prognostic model that combines LLI model and knowledge graph to identify high-risk patients who require mechanical ventilation and optimize medical resource management [9].

In order to explore the influence of artificial intelligence generation technology such as large Language Model (LLMs) on user information retrieval behavior, this paper analyzes the influence of LLMs such as ChatGPT on information retrieval system and user behavior. LLMs brings unique advantages in information retrieval, but its reliability and accuracy are still challenging, and it is difficult to completely replace traditional methods. It is suggested to make full use of LLMs technology and information service intelligence in the construction of information resources to cope with the change of information demand in the future. [10] Design a medical question answering system based on knowledge graph to improve the level of medical service. The medical data knowledge graph is constructed and the question template is generated. The similarity algorithm and intention recognition technology are used to extract the user's intention of the question, and finally the intention is transmitted to the knowledge graph to query the answer and return to realize the intelligent medical question and answer [11].

In the past ten years, the hot spots, emerging fields and dynamic trends of the research of medical consortiums (medical consortiums) have been explored through literature analysis, so as to provide references for the future development of medical consortiums. We analyzed 1,218 literatures collected by CNKI from 2009 to 2019, described the cooperative relationship between authors and institutions, and identified the research hotspots and frontiers through keyword analysis. The study found that medical association is mostly carried out in the form of team cooperation, focusing on information, performance appraisal and specialized medical association. Emerging areas of research include Internet + healthcare, management mechanisms and rehabilitation. It is concluded that information construction, financial management, performance appraisal, rehabilitation and specialized medical consortium cooperation are the frontiers of medical consortium research, but more empirical research is needed [12].

## 3. Integrated Application of LLM Model and Knowledge Graph in Medical Text Mining and Knowledge Extraction

### 3.1. Methods and Technical Basis

### 3.1.1. Overview of LLM Model and Its Training Methods

At present, large language model (LLM), as an important artificial intelligence technology, has shown its powerful application potential in the field of natural language processing. LLM is able to understand and generate natural language through pre-training in large-scale corpora, and has outstanding performance in multiple tasks such as machine translation, text mining and dialogue system, such as the medical question answering framework model of large language model integrating knowledge graph (see Figure 1). Although LLM still faces challenges in terms of knowledge acquisition, especially in terms of knowledge cleaning and refining, it has great potential for solving complex tasks.

The LLM's construction approach covers key techniques such as prompt word engineering and thought chains. Prompt word engineering designs prompt word templates adapted to specific tasks, such as cloze forms and prefixes, to guide LLM to produce accurate output. The thought chain imparts intermediate reasoning steps into prompt words, which helps LLM to solve complex problems step by step, and ensures the logic and traceability of reasoning process. The combination of these technologies enables LLM to better adapt to diverse application needs and provide strong support for further development in the field of intelligent information processing.

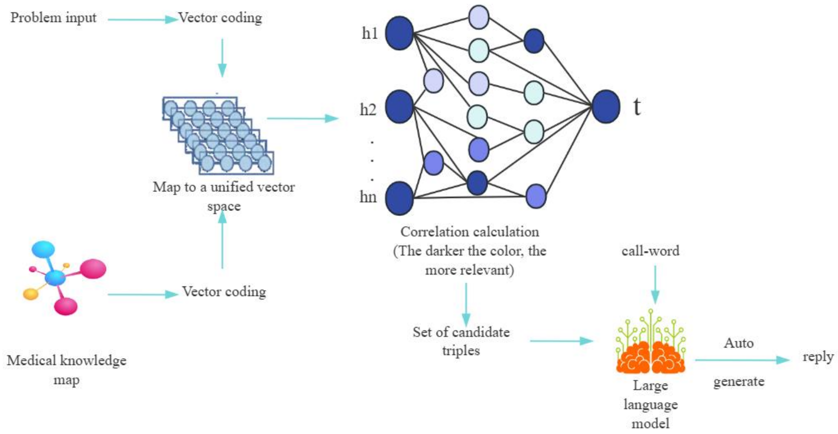
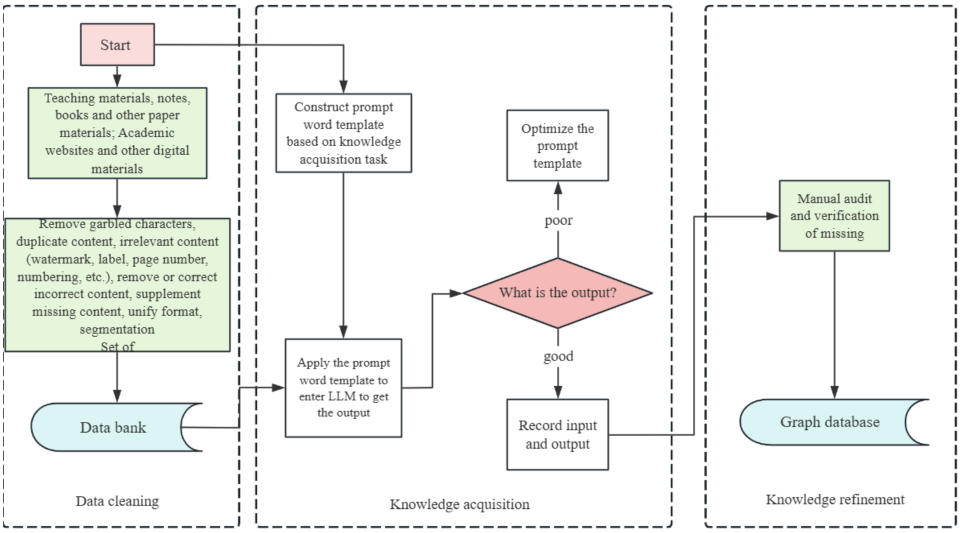


Figure 1: Medical question and answer framework of knowledge graph combined with large language model

### 3.1.2. Knowledge Graph Construction and Maintenance Technology

In current academic research, knowledge graph construction (Figure 2) and maintenance techniques are regarded as key tools to solve complex information management and inference tasks. A knowledge graph is defined as a multi-relationship graph G=(V, E), where V represents a set of entity nodes, E represents a set of edges connecting these nodes, and R represents a set of possible relationship types. Effectively building an accurate and comprehensive knowledge graph relies on the integration of large-scale entity and relational data, as well as efficient algorithms and techniques for processing this data. Advanced pre-trained language models, such as BERT or GPT, can effectively assess the degree of association between nodes in the knowledge graph to improve the accuracy of information retrieval and inference. Analyzing the entity information in question - answer can optimize the structure of knowledge graph and make it adapt to different application scenarios. Subgraph construction and scoring mechanism can optimize the local structure of knowledge graph and improve the relevance and practicability of information.

The key step to ensure the timeliness and accuracy of knowledge is the regular update and correction of knowledge graph. The combination of machine learning and natural language processing technology can automatically identify and process errors or outdated information in knowledge graph, and further improve the reliability and practicability of the system. In academic research and application, the construction and maintenance of knowledge graph is of great significance. Intelligent and automated means can support knowledge management and information retrieval needs in a variety of complex application scenarios, providing efficient and reliable solutions for academia and industry.



Start
Teaching materials, notes, books and other paper materials; Academic websites and other digital materials
Remove garbled characters, duplicate content, irrelevant content (watermark, label, page number, numbering, etc.), remove or correct incorrect content, supplement missing content, unify format, segmentation
Set of
Data bank
Data cleaning
Construct prompt word template based on knowledge acquisition task
Apply the prompt word template to enter LLM to get the output
Optimize the prompt template
poor
What is the output?
good
Record input and output
Knowledge acquisition
Manual audit and verification of missing
Graph database
Knowledge refinement

Figure 2: Flowchart of knowledge graph construction

### 3.1.3. Integration Method of LLM and Knowledge graph and Its Key Technologies

The integration of LLM and knowledge graph and its key technologies have become an important research direction in the field of artificial intelligence. LLM combines pre-trained language representation ability with structured data of knowledge graph to effectively improve information processing and reasoning ability. This integrated approach involves the comprehensive application and optimization of several key technologies: the construction and updating of knowledge graph is the foundation, and the knowledge graph containing entities, relationships and attributes is constructed by integrating and cleaning multi-source data to improve the timeliness and accuracy of information. Fine-tuning and adaptation of LLM models to domain-specific corpora enables better understanding and generation of domain-specific natural language, and significantly enhances the model's performance on knowledge acquisition and inference tasks, combined with structural and semantic information from knowledge graphs. The application of prompt word engineering and thought chain technology helps guide LLM to perform query and reasoning tasks on the knowledge graph, and improves the logic and reasoning efficiency of the model by designing effective prompt word templates and step-by-step reasoning of complex problems. The combination of relation extraction and semantic understanding technology enables the model to understand the entity relationship in text more deeply, and effectively enhance the ability to deal with complex problems. Finally, through the application verification of actual scenarios, the integration method of LLM and knowledge graph has shown remarkable effects and potential in many fields, providing important technical support and application prospects for the further development of intelligent systems.

## 3.2. The Key Link and Application Strategy of Medical Text Mining and Knowledge Extraction Framework Design

### 3.2.1. Data Acquisition and Preprocessing

The stage of data acquisition and preprocessing is particularly critical in the framework design of medical text mining and knowledge extraction. This phase covers a variety of data sources, including clinical records, hospital information systems, scientific literature, and Internet health platforms. An effective data collection strategy must cover different categories and domains of medical text data to ensure that diverse medical information is reflected. The data preprocessing phase involves several key steps, such as text cleaning, de-noising, standardization, word segmentation and entity recognition, which are mainly aimed at improving the quality and accuracy of the data. These techniques provide a solid foundation for the subsequent information extraction, knowledge graph construction and analysis tasks, which also directly affect the performance and application effect of the final model. The optimization of data processing strategies will help the system understand medical texts more precisely to support critical clinical decision making, disease prediction, and drug discovery.

### 3.2.2. Design of a Medical Text Mining Model Based on LLM

The deep pre-training and extensive language understanding ability of LLM are especially suitable for dealing with the complicated context and related issues of professional terms in the medical field. It can also quickly generate relevant content or make complex reasoning according to the input text prompts or questions, thus significantly improving the efficiency of text understanding and information extraction. In terms of knowledge extraction, LLM can accurately identify and extract medical entities, relationships and their attributes, providing strong support for the construction and maintenance of medical knowledge graph. Llm-based design can effectively cope with the diversity and complexity of medical texts, and it also promotes the deep application of medical intelligence technology in clinical decision making and health management.

### 3.2.3. Application Strategy of Knowledge Graph in Medical Knowledge Extraction

Knowledge graph can effectively express the complex knowledge system in the medical field in a structured form, and it can also establish a comprehensive entity relationship network by integrating and standardizing multi-source medical data, including clinical records, scientific research literature, hospital information system and other data channels, to ensure the integrity and accuracy of data. Using intelligent reasoning and query technology, knowledge graph can realize accurate analysis of medical entities and their associated information, and improve the efficiency and quality of information acquisition. It is important to continuously optimize the updating strategy of knowledge graph to reflect the latest developments of knowledge in the medical field in time, and to provide support for medical decision-making and intelligent inference of treatment plans. The integration of these technical means can bring deeper and more accurate data analysis and decision support capabilities for medical text mining and knowledge extraction, and promote the intelligent development of medical health management.

## 4. Results and Discussion

The in-depth discussion of integrated LLM and knowledge graph in medical text mining and knowledge extraction includes several key aspects. This technology integration not only improves the ability of medical information processing in theory, but also shows significant advantages in practical application. The language understanding and generation ability of LLM model combined with the medical entity relationship data rich in knowledge graph can realize the accurate understanding and information extraction of medical texts, which is of great significance to support clinical decision-making, disease prediction and personalized medical management. However, the challenges are focused on data diversity and quality assurance. Textual data in the medical field comes from a wide range of sources, such as clinical records, case reports, and scientific literature, and therefore requires effective integration and cleansing of these data to ensure their quality and consistency. In addition, privacy protection and data security issues are also factors that cannot be ignored when implementing the technology, and need to be fully considered in terms of technical design and legal compliance.

In terms of future research directions and development trends, emphasis should be placed on optimizing the domain adaptability and accuracy of LLM models, especially fine-tuning and optimizing for text and terminology specific to the medical field. In addition, how to realize the automatic construction of knowledge graph, real-time update and seamless integration with external data sources is also one of the key directions of future research. These efforts not only help to improve the intelligent level of medical information processing, but also promote the refinement and personalization of medical decision-making and treatment programs, bringing new possibilities and progress to medical health management.

## 5. Conclusion

In the field of medical text mining and knowledge extraction, the integrated application of LLM model and knowledge graph has shown remarkable academic value and potential. This integration not only effectively improves the accuracy and efficiency of information processing, but also shows a wide application prospect in clinical decision-making, disease prediction and personalized medical management. Its implementation still faces many challenges in practice, such as data quality assurance, privacy and security issues, and technical and legal compliance. Future research should focus on the fine-tuning of LLM model in the medical field and the dynamic updating of knowledge graph, and explore cutting-edge solutions for data integration and privacy protection. These efforts will promote the progress of medical intelligent technology, provide more intelligent and personalized support and services for health management, and then benefit all sectors of society.

## References

*[1] Zhang Heyi, Wang Xin, Han Lifan, et al. Big language model applying knowledge map of question answering system research. Journal of computer science and exploration, 2023, 17 (10): 2377-2388. The DOI: 10.3778 / j.i SSN. 1673-9418.2308070.*

*[2] Huang Bo, Wu Shenao, Wang Wenguang, et al. Complementation of graph and model: a review on fusion of Knowledge graph and large model. Journal of Wuhan University (Science Edition), 2024.*

*[3] Xing Zhenchang, Wan Zhenyu, Wang Changjing, et al. A method and system for querying and clarifying knowledge-guided large language model for API recommendation: CN202310661769.7. CN116776895A [2024-07-06].*

*[4] Pan Yudai, Zhang Lingling, CAI Zhongmin, et al. Knowledge map based on large-scale language model differentiable rules extraction. Journal of computer science and exploration, 2023, 17 (10): 2403-2412. The DOI: 10.3778 / j.i SSN. 1673-9418.2306049.*

*[5] Anonymous." Construction method of Mold Professional Question and answer System based on LLM model", CN117909458A.2024.*

*[6] Cao Yi, Zhang Li, Guo Jing, et al. The development and application prospect of low-carbon electricity market based on large language model. Smart Power, 2024, 52 (2): 8-16.*

*[7] Ye Luchen, Fan Yuan, Wang Xin, et al. Research on content detection algorithm and bypass mechanism of large language model. Information Security Research, 2023, 9 (6): 524-532. (in Chinese)*

*[8] Varatharajah, Y., Chen, H., Trotter, A., & Iyer, R. K. (2020). A Dynamic Human-in-the-loop Recommender System for Evidence-based Clinical Staging of COVID-19. In HealthRecSys@ RecSys (pp. 21-22).*

*[9] Chen, H., Varatharajah, Y., de Ramirez, S. S., Arnold, P., Frankenberger, C., Hota, B., & Iyer, R. (2020). A retrospective longitudinal study of COVID-19 as seen by a large urban hospital in Chicago. medRxiv, 2020-11.*

*[10] Guo Pengrui, Wen Tingxiao. Big language model of information retrieval system and user behavioral impact study. Journal of agricultural information, 2023, 35 (11): 13-22. DOI: 10.13998 / j.carol carroll nki issn1002-1248.23-0573.*

*[11] Yan Debiao. Design and implementation of medical question answering system based on Knowledge graph. Information and Computer, 2023, 35 (13): 123-125. (in Chinese)*

*[12] Bao Yong, Feng Yuanyuan, Xie Qing, et al. Discussion on hot spots and emerging fields of medical consortium research based on knowledge graph. Chinese Public Health Administration, 2022 (001): 038.*