

PLMD-QA: Intelligent Question-Answering System Integrating Large Language Models with Dynamic Medical Knowledge Graphs

Chunming Tang¹, Xinliang Cao²

¹Tianjin Key Laboratory of Intelligent Control for Electrical Equipment, Tiangong University, Tianjin, China

²School of Artificial Intelligence, Tiangong University, Tianjin, China

*Corresponding Author.

Abstract:

An intelligent question-answering system integrating large language models (LLMs) with a dynamic medical knowledge graph is designed to provide rapid and accurate question-answering services for healthcare practitioners and patients. To address the limitations of traditional medical knowledge graphs in terms of coverage and real-time updates, a large-scale dynamic medical knowledge graph is constructed by integrating multi-source data. This graph encompasses various categories, including diseases, medications, and symptoms. An advanced information filtering module, Med-BERT, is implemented to restrict the system's responses to the medical domain effectively, improving both efficiency and accuracy. The system employs the RBBSC named entity recognition model, an entity alignment module, and the RBTRC intent recognition model to ensure precise identification of key user query information. These components enable accurate extraction of critical elements from user questions and alignment with the medical knowledge graph. Additionally, a novel question-answering paradigm is proposed by incorporating P-Tuning techniques. This approach addresses the catastrophic forgetting problem common in traditional fine-tuning methods while enhancing the normative and interpretable nature of the content generated by the LLM. Experimental evaluations demonstrate that the proposed system outperforms mainstream models in entity extraction and intent recognition tasks, achieving higher precision (P) and F1 scores. Furthermore, performance assessments comparing this system to ChatGLM2 and ChatGPT indicate significant improvements in efficiency and response quality. The results highlight the system's ability to deliver specialized and high-quality answers tailored to the medical domain, making it a practical solution for intelligent medical question-answering applications.

Keywords: intelligent medical question-answering, dynamic knowledge graph, p-tuning, entity extraction, intent recognition

INTRODUCTION

The intelligent medical question-answering system based on knowledge graphs integrates natural language processing and graph knowledge reasoning, providing convenient and accurate medical question-answering services for patients and healthcare professionals. It has primarily undergone three developmental stages [1]: rule-based, information retrieval-based, and hybrid multi-model strategies.

Rule-based knowledge graph question-answering systems [2] parse user queries through predefined rules and templates to generate answers. These systems employ methods such as rule and template matching, semantic parsing, keyword matching, and logical reasoning, utilizing query languages like SPARQL and Cypher to define question patterns and answer templates. For example, Hu et al.[3] proposed optimizing natural language question-answering systems by combining keyword search with SPARQL queries. While these methods exhibit high interpretability, they are limited in handling complex queries and lack coverage of broad medical knowledge. Information retrieval-based knowledge graph question-answering systems [4] address some limitations of rule-based systems by generating answers through similarity matching between query vectors and relation vectors in the knowledge graph. These methods can be divided into feature engineering-based [5] and representation learning-based [6] approaches. The former relies on manually designed features, converting queries and topic entities into query graphs and topic graphs to generate answers. However, this process is labor-intensive, struggles to cover all semantic variations, and is prone to ambiguity. To address these challenges, representation learning methods leverage deep learning [7], using attention mechanisms and multi-column modules [8] to improve matching accuracy, automatically learn representations of queries and knowledge graphs, effectively capture semantic information, and handle complex associations. However, these methods require large annotated datasets and face challenges in addressing semantic sparsity [9].

Hybrid multi-model strategy-based knowledge graph question-answering systems [10,11] combine rule matching, information retrieval, and deep learning techniques to provide more accurate and personalized answers. For instance, Yang et al.[12] developed a disease question-answering system that uses BiLSTM-Attention-CRF for entity extraction and BERT-BiLSTM for intent recognition, achieving an F1 score of 85.24% for question recognition. Ji et al.[13] proposed a hybrid BiLSTM-CRF and Attention-BiLSTM-CRF model for named entity recognition in electronic medical records, achieving an F1 score of 90.82%. Xu et al.[14] constructed a RoBERTa-BiLSTM-CRF model in online health communities for diabetes patients, achieving an F1 score of 80.7%. In recent years, large language models such as InstructGPT [15], ChatGPT, and GPT-4 [16] have demonstrated exceptional semantic processing and emergent capabilities across various natural language processing tasks [17-19]. However, directly applying large language models to medical question-answering systems still faces numerous challenges, including high hardware resource requirements, issues with the authenticity and accuracy of generated responses, and the potential for "hallucinations" [20].

In summary, intelligent question-answering systems face the following challenges: (i) the vast volume of medical data requires continuous updates and improvements to public datasets to enhance their coverage, timeliness, and accuracy; (ii) medical data contains a large number of specialized terms that are prone to misrecognition during information retrieval, directly affecting the precision of answers; (iii) the professional nature of medical information output by knowledge graphs makes it essential to ensure that non-professionals can easily understand the content to improve system usability and enhance doctor-patient communication; and (iv) applying large models to the medical domain remains challenging, with performance in vertical domains still falling short.

To address these challenges, this study enriches the medical knowledge graph, constructs more accurate entity extraction and intent recognition models, and integrates multiple techniques, including P-Tuning of large language models [21], to advance research in medical knowledge graph question-answering systems. The specific contributions are as follows: (i) constructing a large-scale dynamic medical knowledge graph integrating multi-source data, encompassing 86,211 medical entities and 703,281 relationships, providing a comprehensive foundation of medical information; (ii) proposing high-accuracy named entity extraction (RBBSC) and intent recognition (RBTRC) models, and designing a Med-BERT information filtering module and an entity alignment unit to better identify and understand the core aspects of user queries; and (iii) introducing a novel question-answering paradigm based on P-Tuning that integrates large language models with medical knowledge graphs.

DESIGN AND IMPLEMENTATION OF THE INTELLIGENT MEDICAL QUESTION-ANSWERING SYSTEM

The framework of the intelligent question-answering system proposed in this study, named PLMD-QA, is illustrated in Figure 1. It is primarily divided into two major components: the construction of a dynamic medical knowledge graph and the generation of question-answering responses.

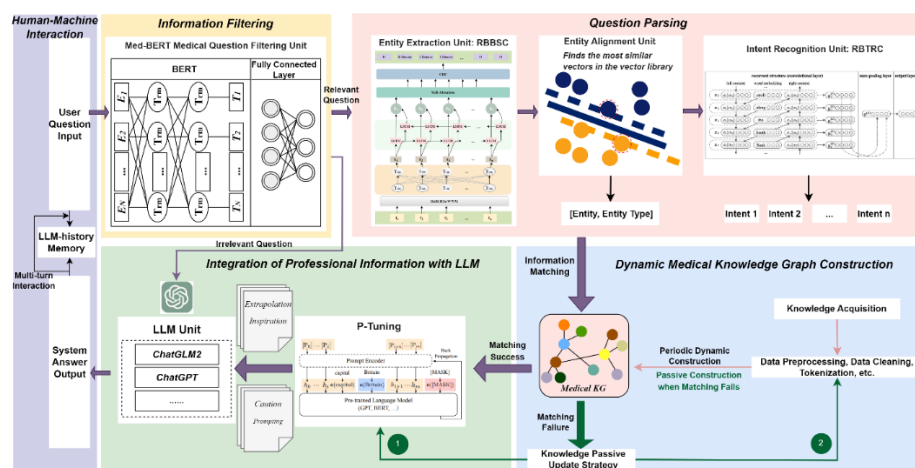


Figure 1. Framework of the PLMD-QA system

Construction of the Dynamic Medical Knowledge Graph

The construction of the dynamic medical knowledge graph consists of three key stages: knowledge acquisition, knowledge graph construction, and knowledge updates.

Acquisition of medical knowledge data

To ensure professionalism and accuracy, the medical knowledge graph data constructed in this study is primarily sourced from: (i) authoritative professional medical information websites, including XYWY.COM, healthoo.com, and wedoctor.com. These websites store semi-structured data, which was parsed and acquired using web crawling techniques; (ii) open-source electronic medical record datasets, consisting of unstructured medical text data, from which information was extracted using the joint extraction model TPLinker [22]; and (iii) the open-source medical knowledge graph dataset CmeKG2.0, comprising structured data that can be directly integrated. All extracted information underwent cleaning and normalization processes to unify it into structured data.

Construction of the medical knowledge graph

Knowledge in the graph is represented in two formats: entity-relation-entity and entity-attribute-attribute value. The cleaned data was categorized into nine types of entities and 11 types of entity relationships (as shown in Table 1) and further divided into eight major disease attributes, seven doctor attributes, and three hospital attributes (as shown in Table 2).

Table 1. Entity and relationship categories in the medical knowledge graph

Entity	Relationship
Disease	recommand_drug (Relationship between diseases and drugs)
Drug	has_symptom (Relationship between diseases and symptoms)
Symptom	belongs_to (Relationship between diseases and departments)
Department	acompany_with (Relationship between diseases and complications)
Check(Diagnostic Test)	need_check (Relationship between diseases and required diagnostic tests)
Food	recommand_recipes (Relationship between diseases and recipes)
Producer(Pharmaceutical Company)	do_eat (Relationship between diseases and recommended foods)
Doctor	not_eat (Relationship between diseases and forbidden foods)
Hospital	drugs_of (Relationship between drugs and producers)
/	specializes_in (Relationship between doctors and specialties)
/	affiliated_with (Relationship between doctors and affiliated hospitals)

Table 2. Attributes of diseases, doctors, and hospitals in the medical knowledge graph

Disease	Doctor	Hospital
name	name	name
Desc (Disease Description)	title	level
cause	desc (Doctor Description)	address
prevent (Prevention Measures)	gender	/
cure_lasttime (Treatment Duration)	reception_rate	/
cure_way (Treatment Method)	positive_review_rate	/
cured_prob (Cure Probability)	consultation_volume	/
easy_get (Susceptible Population)	/	/

The aforementioned entities, relationships, and attributes were imported into the Neo4j graph database, resulting in the construction of a knowledge graph comprising 86,211 entities and 703,281 relationships.

Knowledge graph updates

a) Periodic Automatic Updates: Using the APScheduler mechanism, a periodic web crawling process is implemented to update medical data regularly. Knowledge updates are scheduled for the first day of each week,

during which the latest medical data from platforms such as XYWY.COM, healthoo.com, and wedoctor.com is retrieved, processed, and integrated into the existing knowledge graph.

b) **Updates Triggered by Information Matching Failures:** When the dynamic medical knowledge graph fails to retrieve relevant content, indicating a matching failure, the system initiates a passive update mechanism. For unrecognized entities, the information matching unit performs keyword searches on "Baidu Health Encyclopedia" and conducts targeted web crawling. The retrieved data is formatted according to predefined templates, integrated into corresponding prompts, and sent to the large language model (LLM) for response generation. Once the response is generated, the extracted information is converted into triplets and stored in the knowledge graph, thereby completing the update process.

Question-Answer Generation

The question-answer generation process comprises four core modules: information filtering, query parsing, integration of professional information with the LLM, and human-machine interaction.

Med-BERT information filtering

As illustrated in Figure 1, the process begins with Med-BERT, which filters user queries to determine their relevance to the medical domain, thereby improving the system's response efficiency. Specifically, the query is first analyzed by BERT to extract its features, which are then passed through a fully connected layer to generate a binary classification result, indicating whether the query pertains to a medical issue. If the query is identified as relevant, the system forwards it to the query parsing module for detailed analysis. Conversely, if deemed irrelevant, the system utilizes the large language model (LLM) to generate a rejection response. During training, the input vector dimension for the fully connected layer is 768, with two hidden layers of dimensions 384 and 768, respectively, and an output dimension of 2 (corresponding to the binary classification task). Finally, the system selects the label with the highest classification probability as the final result.

Query parsing

Query parsing involves three components: the RBBSC named entity extraction model, entity alignment, and the RBTRC question intent recognition model.

a) **RBBSC Named Entity Extraction Model:** As shown in Figure 2, the RBBSC named entity extraction model is composed of three layers. First, the input query is processed through the RoBERTa-wwm embedding layer, which generates semantic feature vectors. These vectors are then passed to the feature extraction layer, where BiLSTM captures bidirectional contextual information and extracts deep sequence features. Subsequently, the Self-Attention mechanism globally weights these features, highlighting critical dependencies within the sequence. Finally, the output is processed by the CRF (Conditional Random Field) layer, which imposes constraints on label dependencies and selects the optimal output path, completing the entity extraction task.

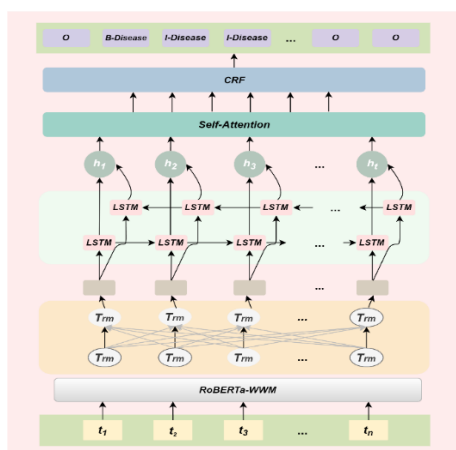


Figure 2. Diagram of the RBBSC named entity extraction model

(i) RoBERTa-wwm Embedding Layer

In this study, the RoBERTa-wwm model [23] is utilized as the embedding layer to extract semantic and positional features from the input text, transforming it into semantic vectors. The computation process is represented by (1) and (2).

$$e_i = W_p + W_t \quad (1)$$

$$x_\alpha = Tf(e_{\alpha-1}), \alpha \in [1, N] \quad (2)$$

Here, e_i represents the initial word embedding vector, which is initialized using the word embedding matrix W_p and the positional embedding matrix W_t . For single-sentence inputs, sentence-level embeddings are not considered. x_α denotes the output of the α -th Transformer layer, where $Tf(.)$ represents the Transformer function, and N is the maximum number of Transformer layers. After processing through the embedding layer, the final output is $X = \{x_1, x_2, \dots, x_n\}$.

(ii) Feature Extraction Layer

The feature extraction layer integrates BiLSTM and Self-Attention mechanisms. The output sequence X from the embedding layer serves as the input to the BiLSTM. Through bidirectional encoding, the forward and backward hidden states are computed for each time step, producing the hidden state output h_t . The feature matrix is expressed as $H = [h_1, h_2, \dots, h_t]$, as defined in (3):

$$H = \text{BiLSTM}(X) \quad (3)$$

Next, H is input into the Self-Attention layer, which calculates the relationships between elements in the sequence to further enhance feature representations and highlight dependencies within the sequence. This process involves transforming H into the query vector Q , key vector K , and value vector V through linear transformations, maintaining the input sequence dimension d . The attention weights α are then computed using scaled dot-product attention, as shown in (4):

$$\alpha = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \quad (4)$$

Finally, the weighted sum of the value vector V and attention weights α is calculated to obtain the Self-Attention output vector A , as defined in (5):

$$A = \sum_{n=1}^N \alpha_n V_n \quad (5)$$

Here, $V_n \in \mathbb{R}^{d \times d}$ represents the word vector of the n -th word in the input sequence. A denotes the final weighted feature output. The output of the Self-Attention layer is represented as $A = [V_1, V_2, \dots, V_n]$.

(iii) CRF Output Layer

The output layer employs a CRF (Conditional Random Field) to impose BIO constraints on the label sequence, thereby enhancing model performance. The BIO constraint rules for sequence labeling are outlined in Table 3.

Table 3. BIO constraint rules for sequence annotation

Annotation Sequence	Example	Description
Valid Annotation Sequence	B-type1, I-type1, I-type1	Correct annotation
Invalid Annotation Sequence 1	B-type1, I-type1, I-type2	Inconsistent labels within an entity
Invalid Annotation Sequence 2	B-type1, O, I-type1	Discontinuous labels within an entity
Invalid Annotation Sequence 3	I-type1, I-type1, I-type1	Entity label starting with "I"
Invalid Annotation Sequence 4	O, I-type1, I-type1	Entity label starting with "O"

B represents the beginning of an entity, I indicates the inside of an entity, and O denotes non-entity tokens. Type specifies the category of the entity.

b) **Entity Alignment:** In the medical domain, the abundance of specialized terminology and disease names often results in multiple representations for the same entity after extraction. Entity alignment utilizes entities from the knowledge graph to correct and standardize the extracted entities, ensuring the question-answering system can accurately interpret user queries.

The implementation involves first using the Word2Vec embedding method to convert the entities in the knowledge graph into corresponding vector representations [24]. For the user input, the extracted entities are converted into embedding vectors using the entity extraction model. All vectors are then normalized using the L2 norm, as shown in (6):

$$V_{norm} = \frac{V}{\|V\|_2} \quad (6)$$

Here, $\|V\|_2$ denotes the L2 norm of the vector, which represents its magnitude.

Next, the similarity between the extracted entities and the knowledge graph entities is computed using cosine similarity, as defined in (7):

$$\text{cosine_similarity}(V_{input}, V_{graph}) = \frac{V_{input} \cdot V_{graph}}{\|V_{input}\| \times \|V_{graph}\|} \quad (7)$$

In this equation, V_{input} and V_{graph} are the two input vectors, and $V_{input} \cdot V_{graph}$ represents their dot product. $\|V_{input}\|$ and $\|V_{graph}\|$ are the L2 norms of V_{input} and V_{graph} , respectively. If the similarity exceeds a predefined threshold, the extracted entity is replaced with the corresponding representation from the knowledge graph.

c) **RBTRC Question Intent Recognition Model:** The RBTRC question intent recognition model is structured from bottom to top into three layers: the vector representation layer, the TextRCNN layer [25], and the classification layer. First, the given text $T = [t_1, t_2, \dots, t_n]$ is processed through the RoBERTa-wwm pre-trained model, generating word embeddings $T = [t_1, t_2, \dots, t_n]$. These embeddings X are then passed to the TextRCNN layer, which extracts semantic vectors. Finally, the semantic vectors are input into the classification layer, producing intent recognition results across 29 categories (see Table 4).

Table 4. Intent categories in the question-answering system proposed in this paper

Intent Type	Intent Type	Intent Type
Symptoms of a Disease	Diseases Treatable by a Medication	Hospital Affiliation of a Doctor
Possible Diseases for a Symptom	Manufacturer of a Medication	Title of a Doctor
Overview of a Disease	Diagnostic Procedures for a Disease	Profile of a Doctor
Causes of a Disease	Prevention Measures for a Disease	Gender of a Doctor
Treatment Department for a Disease	Treatment Duration for a Disease	Reception Rate of a Doctor
Complications of a Disease	Treatment Methods for a Disease	Positive Rating of a Doctor
Foods to Avoid for a Disease	Cure Rate of a Disease	Consultation Volume of a Doctor
Recommended Foods for a Disease	Susceptible Population for a Disease	Rank of a Hospital
Suggested Recipes for a Disease	Specialists for a Disease	Address of a Hospital
Medications for a Disease	Diseases Specialized by a Doctor	/

Integration of professional information with LLM

Table 5. Examples of prompts (total of 29 categories)

Intent Type	Prompt Template
Symptoms of a Disease	[f"Acting as a medical expert, based on professional knowledge: {information}, classify the related symptoms by type and describe the relationship between each type and its corresponding symptoms in detail. Think step by step and output sequentially to answer the user's query: {query}."]
Possible Diseases for a	[f"Acting as a medical expert, based on professional knowledge: {information}, analyze

Symptom	the possible diseases indicated by the symptoms step by step, and output sequentially to answer the user's query: {query}."
Causes of a Disease	[f"Acting as a medical expert, based on professional knowledge: {information}, explain the causes of the disease mentioned in the user's query in detail. Please follow these steps to output step by step: 1. Primary Causes: List the main causes and explain each in detail; 2. Transmission Routes: Elaborate on how different routes of exposure or environmental factors influence the disease's occurrence; 3. Symptoms and Impacts: Describe common symptoms and their health impacts; 4. Prevention: Provide suggestions to reduce the risk of the disease. Answer the user's query: {query}."
Treatment Department for a Disease	[f"Acting as a medical expert, based on professional knowledge: {information}, specify the appropriate department for treating the user's queried disease, along with suggestions for managing severe symptoms and other relevant recommendations. Output sequentially to answer the user's query: {query}."
Medications for a Disease	[f"Acting as a medical expert, based on professional knowledge: {information}, analyze suitable medications for the disease step by step. Categorize them by mechanism of action, and discuss each category's specific effects, indications, and precautions in detail. Output sequentially to answer the user's query: {query}."
...	...
Susceptible Population for a Disease	[f"Acting as a medical expert, based on professional knowledge: {information}, specify the susceptible populations for the disease mentioned in the user's query, and provide different lifestyle prevention recommendations for each group. Output step by step to answer the user's query: {query}."
Specialists for a Disease	[f"Acting as a medical expert, based on professional knowledge: {information}, provide information on specialists experienced in treating this disease: {query}. Output step by step."
Profile of a Doctor	[f"Acting as a medical expert, based on professional knowledge: {information}, provide the doctor's profile. Please output step by step according to the following: 1. Basic Information: Name, gender, and date of birth; 2. Work Unit: Current position; 3. Education Background; 4. Career Development: Career history; 5. Academic and Research Information: Research projects and academic publications. Answer the user's query: {query}."

The performance of large language models (LLMs), such as their generalization capabilities, is influenced by the type of training data and alignment standards. For instance, ChatGPT-4, with training data up to September 2021, is unable to respond to subsequent events and may produce hallucinations. This limitation arises from the temporal sensitivity of data and the complexity of domain-specific information. Additionally, medical knowledge graph-based question-answering systems may generate outputs that, while rich in professional content, are challenging for patients to comprehend.

To address these issues, this study integrates the ChatGLM2 with specialized medical knowledge graph information, proposing a P-Tuning-based prompt learning method to enhance professional interpretability in question-answering. This method leverages the semantic understanding of LLMs to transform graph data into comprehensible model inputs, improving the accuracy and relevance of responses [26]. Compared to large-scale parameter tuning, P-Tuning optimizes only the prompt embedding parameters, significantly reducing resource consumption.

a) **Information Matching:** After extracting key entities and intents in the query parsing module, the information matching unit constructs Cypher queries to search the knowledge graph for relevant information. The results are then populated into the corresponding intent prompt templates in the {query} and {information} fields and passed to the LLM for response generation.

b) **P-Tuning:** P-Tuning is an innovative fine-tuning method for LLMs that introduces learnable prompt embeddings to enhance the model's understanding of downstream tasks, thereby generating high-quality outputs tailored to task requirements. Unlike traditional fine-tuning, P-Tuning optimizes only the prompt embeddings[27], retaining the model's generative capabilities while significantly reducing fine-tuning complexity. In this study, we combine prompts for each question category with ideal question-answer pairs to

fine-tune the model, using P-Tuning to optimize prompts for better comprehension of medical terminology and expressions. Questions are categorized into 29 types, with specific prompt templates designed for each (some examples are provided in Table 5) to guide the LLM in identifying question types and generating professional responses. Additionally, ideal output samples are constructed for each question type to train the model to produce optimal answers. These samples are derived from actual medical consultations and relevant literature to ensure the accuracy and professionalism of the generated outputs.

Human-machine interaction

The backend integrates the knowledge graph and various question-answering modules using the Flask framework, while the frontend employs the Vue framework to provide a user-friendly interface. The human-machine interaction module enables intuitive visualization of generated answers and supports multi-turn conversations, ensuring a consistent and coherent interactive experience. During multi-turn interactions, the LLM-history memory structurally stores the content of previous conversations, assisting the model in understanding the current context and user intent. By leveraging prior dialogue history, the system optimizes responses, enhancing personalization and interaction quality.

EXPERIMENTAL DESIGN AND RESULT ANALYSIS

Data Collection and Preprocessing

Table 6 provides detailed information about the various datasets utilized for the implementation of this system.

(i) The CMEE dataset, developed by institutions such as Peking University and Zhengzhou University, covers nine categories of medical entities, including diseases, symptoms, and medications. The CMID dataset includes 36 subcategories of medical intents, addressing aspects such as symptom descriptions, inquiries about causes, and treatment recommendations.

(ii) Samples for the Med-BERT filter, along with the custom-built medical entity extraction and dialogue datasets, were sourced from real conversations on medical websites.

(iii) Since the Med-BERT dataset of negative samples lacked comprehensive coverage, this study expanded it using LLM-based prompt methods. ChatGPT was utilized to generate additional negative samples, with the specific procedure detailed in the pseudocode below.

Table 6. Dataset sample sizes

Med-BERT Medical Question Filtering Training Dataset	Positive Samples (Existing Q&A Dataset): 1200	
	Negative Samples (Existing Q&A Dataset + ChatGPT-Generated Samples): 300	
Entity Extraction Experiment Dataset	CMeEE Medical Chinese Entity Extraction Dataset	15,000 samples (Split 7:1:2 into training, validation, and test sets)
	Custom Medical Chinese Entity Extraction Dataset	
Intent Recognition Experiment Dataset	CMID Chinese Medical Dialogue Dataset	12,000 samples (Split 7:1:2 into training, validation, and test sets)
	Custom Chinese Medical Dialogue Dataset	
P-Tuning Dataset	1,125 ideal Q&A pairs	
Question-Answering Effectiveness Test Dataset	Subjective Q&A Test: 100 questions	
	Objective Test: 200 questions	

Algorithm 1: Generating Negative Samples with GPT

Require: Negative sample text T, ChatGPT API key{API_KEY}.

Ensure: Negative sample data R generated by ChatGPT.

ChatGPT_connection ← create(API_KEY)

For (i = 1 to N) do

 t ← select(T) //Select one piece of negative sample data

 prompt ← P(t)//Generate a prompt based on the selected data


```

r ← chatgptresult(ChatGPT_connection, prompt)
R ← abstract(r)
End
Where P is defined as: [f"Now you are a medical expert. Please refer to the following data: {information},
and generate some related questions."]
```

Entity Extraction Experiment

Experimental setup

This experiment utilizes the CMeEE medical entity extraction dataset and a custom-built medical entity extraction dataset, comprising a total of 15,000 samples. The data is divided into training, validation, and test sets in a 7:1:2 ratio. The datasets are annotated using the BIO labeling scheme, where B represents the beginning of an entity, I represents the inside of an entity, and O denotes non-entity characters. The experiment employs the Adam optimizer with the following model parameters: a maximum sequence length (Max_len) of 100, a hidden dimension (Hidden_dim) of 768, 12 Transformer layers, a learning rate (Learning_rate) of 1e-5, a training batch size (Train_batch_size) of 32, and 8 attention heads in the self-attention mechanism. To mitigate overfitting, the dropout rate is set to 0.5, while the LSTM hidden layer contains 128 units. The number of training epochs (epoch) is set to 50.

Experimental results and analysis

We compared the performance of several mainstream entity extraction models, and the experimental results are presented in Table 7. In this table, P denotes precision, R denotes recall, and the F1-score is calculated as shown in (8):

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (8)$$

Table 7. Comparison of experimental results for entity extraction models

Model	P(%)	R(%)	F1(%)
BiLSTM-CRF[28]	64.4	66.2	65.3
BERT-BiLSTM-CRF[29]	75.1	76.8	75.9
RoBERTa-WWM-BiLSTM-CRF[30]	76.4	78.9	77.4
RBBSC (Ours)	77.3	79.1	78.2

The experimental results indicate that the proposed model outperforms BiLSTM-CRF, BERT-BiLSTM-CRF, and RoBERTa-WWM-BiLSTM-CRF models in terms of precision, recall, and F1-score.

Entity Alignment Experiment

After multiple tests, the entity alignment performed well when the cosine similarity threshold was set above 0.8. This allowed the system to replace incorrect entities extracted by the entity extraction model with entities from the knowledge graph, enabling a more accurate understanding of user queries. Partial results of the entity alignment are presented in Table 8.

Table 8. Examples of entity alignment results

RBBSC Output	Knowledge Graph Entity	Cosine Similarity	Threshold	Replaced	Replacement Result
Chronic Bronchitis	Bronchitis	0.75	0.8	No	/
Fungal Pneumonia	Bacterial Pneumonia	0.92	0.8	Yes	Bacterial Pneumonia
Elevated Blood Pressure	Hypertension	0.89	0.8	Yes	Hypertension
Arrhythmia	Atrial Fibrillation	0.86	0.8	Yes	Atrial Fibrillation

Cardiac Arrest	Heart Disease	0.65	0.8	No	/
Pulmonary Bullae	Pulmonary Bullae	1	0.8	Yes	Pulmonary Bullae
Pulmonary Tuberculosis	Tuberculosis	0.85	0.8	Yes	Tuberculosis

Intent Recognition Experiment

Experimental setup

This experiment utilizes the publicly available CMID medical intent dataset, along with a manually annotated dataset collected from real dialogue data. The dataset aligns with the system's 29 intent categories and contains a total of 12,000 records. The data is split into training, validation, and test sets in a 7:1:2 ratio.

Experimental results and analysis

In the intent recognition task, the performance of several models, including TextCNN, TextRNN, BERT-TextCNN, RoBERTa-WWM-TextCNN, and the proposed RBTRC model, was compared. The evaluation metrics used are consistent with those employed in the entity extraction task. The results are shown in Table 9.

Table 9. Comparison of experimental results for intent recognition models

Model	P(%)	R(%)	F1(%)
TextCNN[31]	81.18	80.99	81.08
TextRNN[32]	81.25	81.03	81.14
BERT-TextCNN[33]	84.01	82.80	83.81
RoBERTa-WWM - TextCNN	84.33	83.95	84.14
RBTRC (Ours)	87.03	86.22	86.62

The experimental results demonstrate that the RBTRC model outperforms other models in terms of precision, recall, and F1-score, highlighting its effectiveness in intent recognition classification.

Performance Evaluation

The professional effectiveness of the system in the medical question-answering domain was evaluated from two perspectives: subjective evaluation and objective performance evaluation.

Subjective evaluation

a) Comparison of Responses from Different Models: The comparison of intelligence, specialization, and timeliness is illustrated in Tables 10-12, showcasing the responses of three models to three medical questions. Compared to ChatGLM2 and ChatGPT, the proposed question-answering system demonstrates superior professionalism and refinement. For diagnostic tests and medication recommendations, the proposed system not only lists specific tests and drug names but also provides detailed explanations of the purposes of these tests, mechanisms of action, indications, and precautions for the medications. For instance, in the case of the query regarding "medication recommendations for neonatal respiratory distress syndrome", ChatGLM2 exhibited hallucinations typical of large language models, recommending drugs (dexamethasone and methylprednisolone) aimed at relieving respiratory distress based solely on the keyword "respiratory difficulty." It failed to consider the neonate's age and the safety of the recommended medications, revealing deficiencies in contextual understanding and domain-specific refinement. In contrast, the proposed system leverages relevant professional knowledge from the dynamic medical knowledge graph, incorporating age restrictions, indications, and safety considerations to ensure more targeted and secure medication recommendations. Lastly, for question 3, both ChatGLM2 and ChatGPT were unable to provide up-to-date information as their training data is limited to pre-2024. In contrast, the proposed system dynamically responds to queries, demonstrating excellent timeliness.

Table 10. Comparison of question-answering for diagnostic procedures for a disease

Model	What tests should I undergo for emphysema?
	Response
ChatGLM2	If diagnosed with emphysema, the following tests may be necessary: 1. Pulmonary function tests: This is the most common diagnostic method for emphysema. Pulmonary function tests include lung capacity, vital capacity, and pulmonary function indices, which measure lung function and the degree of airway obstruction. 2. Respiratory system examination: Includes auscultation, oral examination, and chest X-rays to assess airway conditions and check for other abnormalities. 3. Blood tests: Blood tests can evaluate the presence of chronic obstructive pulmonary disease (COPD) and other related health issues, such as cardiovascular diseases and diabetes. 4. Lung tissue biopsy: If the cause of emphysema is clear, such as allergic rhinitis or bronchial asthma, a lung tissue biopsy may be required for further diagnosis and treatment. 5. Imaging tests: CT or MRI scans can measure the density and structure of lung tissue, helping determine the severity and cause of emphysema.
ChatGPT	The diagnosis and evaluation of emphysema typically require a series of tests to determine the severity and impact of the condition. Common tests include: 1. Chest X-ray: Helps identify structural changes in the lungs. 2. Computed tomography (CT): Provides more detailed images of the lungs, helping determine the extent and severity of emphysema. 3. Pulmonary function tests: Measures the airflow capacity of the lungs, including forced expiratory volume (FEV1) and forced vital capacity (FVC), to assess lung function damage. 4. Blood gas analysis: Measures oxygen and carbon dioxide levels in the blood to evaluate respiratory efficiency. 5. Respiratory secretion examination: Analyzes sputum to check for infections or other abnormalities. 6. Alpha-1 antitrypsin level test: A hereditary disease test, sometimes associated with emphysema. 7. Physical fitness and endurance tests: Evaluate the patient's overall health and ability to perform daily activities.
PLMD-QA (Ours)	Emphysema is a common type of chronic obstructive pulmonary disease (COPD), with main symptoms including chronic cough, sputum production, shortness of breath, and breathing difficulties. Regular physical examinations are crucial for patients with emphysema to monitor disease progression and take appropriate treatment measures. Recommended tests include: 1. Pulmonary Function Evaluation (1) Pulmonary Function Tests: a. Vital Capacity (VC): Evaluate lung capacity. b. Maximum Expiratory Flow-Volume Curve (MEFV): Analyze the relationship between airflow and volume. c. Residual Volume/Total Lung Capacity Ratio (RV/TLC): Assess airway obstruction. (2) Respiratory Function Tests: a. Forced Expiratory Volume/Forced Vital Capacity Ratio (FEV1/FVC): Evaluate airway patency. b. Alveolar-Arterial Oxygen Gradient: Assess gas exchange efficiency. (3) Diffusion Function: a. Diffusion Coefficient (D): Assess oxygen transfer through alveolar walls. b. Pulmonary Diffusion Area: Reflect the effective gas exchange surface area. (4) Lung Compliance: a. Lung Elasticity Modulus (CME): Evaluate lung tissue elasticity. b. Lung Volume Index (VFI): Assess lung expansion capacity. 2. Blood Gas Analysis (1) Arterial Blood Gas Analysis: a. Arterial Oxygen Pressure (PaO2): Assess oxygen content in the blood. b. Arterial Carbon Dioxide Pressure (PaCO2): Monitor CO2 elimination. c. pH Value: Evaluate acid-base balance. 3. Imaging Tests a. Chest X-ray: Assess the extent and type of lung lesions, including nodules, pneumonia, and emphysema. 4. Cardiac Function Evaluation a. Echocardiography: Evaluate the relationship between cardiac and pulmonary functions, including heart structure and blood flow dynamics. 5. Tissue and Genetic Tests a. Lung Tissue Biopsy: Obtain lung tissue samples for pathological examination to confirm diagnosis. b. Genetic Testing: Evaluate hereditary factors associated with emphysema through genetic assessment...

In summary, the proposed system not only integrates professional knowledge effectively but also accurately addresses time-sensitive questions. Its responses exhibit greater coherence and logical structure, further validating the significant impact of P-Tuning in improving the quality of generated answers from large models.

Table 11. Comparison of question-answering for medications for a disease

Model	What medications are recommended for neonatal respiratory distress syndrome (NRDS)?
	Response
ChatGLM2	Neonatal respiratory distress syndrome (RDS) is a common neonatal condition caused by a deficiency or insufficiency of pulmonary surfactant, leading to neonatal respiratory difficulty and hypoxemia. The medications commonly used for treating RDS include: 1. Glucocorticoids: Glucocorticoids are commonly used for treating neonatal respiratory distress syndrome as they can reduce pulmonary edema and alleviate respiratory distress. Common glucocorticoids include prednisone, methylprednisolone, and dexamethasone. 2. Diuretics: Diuretics can help reduce fluid accumulation in the body, thereby alleviating pulmonary edema and improving respiratory distress. Commonly used diuretics include furosemide and bumetanide. 3. Pulmonary Surfactant Substitutes: Pulmonary surfactant is critical for regulating alveolar surface tension. Deficiency in pulmonary surfactant leads to RDS. Certain drugs can act as substitutes, such as pulmonary surfactant receptor antagonists and autologous plasma replacement therapy. 4. Oxygen Therapy: Oxygen therapy is required to treat hypoxemia caused by RDS and improve oxygen supply...
ChatGPT	The treatment of neonatal respiratory distress syndrome (NRDS) mainly includes the following medications and supportive measures: 1. Surfactants: This is the key medication for treating NRDS. Surfactants help reduce alveolar surface tension and improve lung function. Common surfactants include aminoglycoside-based surfactants (e.g., Poractant alfa and Beractant) and phospholipid-based surfactants (e.g., Lucinactant). 2. Oxygen Therapy: Provides high-concentration oxygen to improve blood oxygen saturation. 3. Respiratory Support: Non-invasive ventilation (e.g., CPAP) or mechanical ventilation may be needed to maintain respiratory function. 4. Glucocorticoids: In certain cases, especially before delivery, mothers may be given glucocorticoids to promote fetal lung development (e.g., dexamethasone). 5. Other Supportive Treatments...
PLMD-QA (Ours)	Neonatal respiratory distress syndrome is a common neonatal condition that often results in hypoxemia, respiratory difficulty, and increased heart rate in newborns. Correct treatment is crucial for neonatal health. The following are commonly used medications for treating NRDS: 1. Surfactant Class = a. Medication: Porcine lung phospholipid injection. b. Mechanism: Surfactants reduce alveolar surface tension, prevent alveolar collapse, and improve pulmonary gas exchange. c. Indications: Mainly used for preterm infants with NRDS caused by insufficient pulmonary surfactant synthesis. d. Precautions: Monitor the infant's oxygen saturation and heart rate during use to avoid overdose-related lung

	<p>damage.</p> <p>2. Bronchodilators a. Medication: Ambroxol hydrochloride injection, ambroxol hydrochloride glucose injection, injectable ambroxol hydrochloride. b. Mechanism: These drugs dilate the bronchi and reduce airway resistance, thereby improving respiratory function. c. Indications: Used in cases of bronchospasm or airway obstruction. d. Precautions: Monitor heart rate changes to avoid excessive excitement or tachycardia.</p> <p>3. Mucolytics a. Medication: Bromhexine hydrochloride glucose injection. b. Mechanism: Helps dilute and clear respiratory secretions, improving ventilation. c. Indications: Suitable for cases with excessive viscous sputum to help clear the airway. d. Precautions: Evaluate the nature of secretions before use to ensure no allergic reactions...</p>
--	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Table 12. Comparison of question-answering for overview of a disease

Model	Do you know about the dengue fever outbreak in 2024?
	Response
ChatGLM2	Due to the limited data on the dengue fever outbreak in 2024, I am unable to provide detailed information. (Training data ends in 2023, so I cannot provide detailed information on diseases emerging in 2024.)
ChatGPT	<p>Dengue fever is an acute infectious disease caused by the dengue virus, primarily transmitted through mosquitoes. Below are some key points about dengue fever in 2024:</p> <p>1. Cause a. Virus: Dengue fever is caused by the dengue virus (DENV), which has four serotypes (DEN-1, DEN-2, DEN-3, DEN-4). Infection with one serotype does not confer lifelong immunity, and individuals remain susceptible to other serotypes.</p> <p>2. Transmission Pathway ... 3. Symptoms ... 4. Diagnosis ... 5. Treatment ... 6. Prevention ...</p> <p>7. Situation in 2024 a. Epidemic Trend: The prevalence of dengue fever in 2024 may vary by region, with some areas experiencing peak outbreaks. It is essential to stay informed through local public health announcements. If you or anyone around you develops similar symptoms, especially in areas where dengue fever is prevalent, seek medical attention promptly and take appropriate protective measures.</p>
PLMD-QA (Ours)	<p>1. What is Dengue Fever? Dengue fever is an acute infectious disease caused by the dengue virus, primarily transmitted by mosquitoes, especially <i>Aedes aegypti</i>. The virus has four serotypes (DEN-1, DEN-2, DEN-3, DEN-4). Infection with one serotype does not confer lifelong immunity, leaving individuals susceptible to other serotypes in the future.</p> <p>2. Dengue Fever Situation in 2024 According to data from Brazil's Ministry of Health, since the beginning of 2024, cases of dengue fever have surged, with suspected and confirmed cases reaching over 6.48 million and fatalities rising to 5,139. An additional 2,048 deaths are under investigation. These figures highlight the severity of the epidemic and emphasize the need for public attention to preventive measures.</p> <p>3. Transmission of Dengue Fever Dengue virus is primarily transmitted through mosquito bites. When an infected mosquito bites a healthy individual, the virus enters the body through the mosquito's saliva.</p> <p>4. Symptoms of Dengue Fever a. Common symptoms include high fever, severe headaches, pain behind the eyes, joint and muscle pain, rashes, nausea, and vomiting. b. Severe dengue fever can lead to dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS), which require urgent medical attention.</p> <p>5. Treatment Currently, no specific antiviral treatment exists for dengue fever. Treatment focuses on symptom management, including: a. Fluid replacement. b. Use of fever-reducing medications (avoid aspirin and other NSAIDs). c. Maintaining fluid balance.</p> <p>6. Prevention Measures a. To prevent dengue fever, consider the following: Use mosquito repellents, wear long-sleeved clothing, use mosquito nets, and eliminate standing water to reduce mosquito breeding. b. While dengue vaccines are available in some regions, their indications and efficacy vary. Follow local health authorities' guidance for vaccination...</p>

b) **Experimental Setup:** The system also invited three medical teams (China Pharmaceutical University [CPU], Capital Medical University [CMU], and Gansu Medical College [GSMC]) to evaluate the responses generated by different models to verify the system's effectiveness. A total of 100 questions were randomly selected from the real dialogue dataset and input into the three models to generate answers. When comparing responses to the same question, medical team members selected the better answer based on content quality, without being informed of the source model.

As shown in Figure 3, the horizontal axis represents the different medical teams, and the vertical axis indicates the number of satisfactory answers for each model. Overall, the proposed system demonstrated superior satisfaction levels for most questions, validating its effectiveness. Medical team members noted that for simpler questions, Model 1 (the proposed system) performed significantly better than the other two models. For more complex questions, Model 3 (ChatGPT) showed a slight edge over Model 1, far surpassing Model 2 (ChatGLM2). This could be attributed to ChatGPT's training data, which includes some professional content, giving it an advantage in handling certain complex questions. Notably, medical team members highlighted that Model 1 (the proposed system) provided more detailed and organized responses enriched with professional knowledge, demonstrating strong potential for broader practical applications.

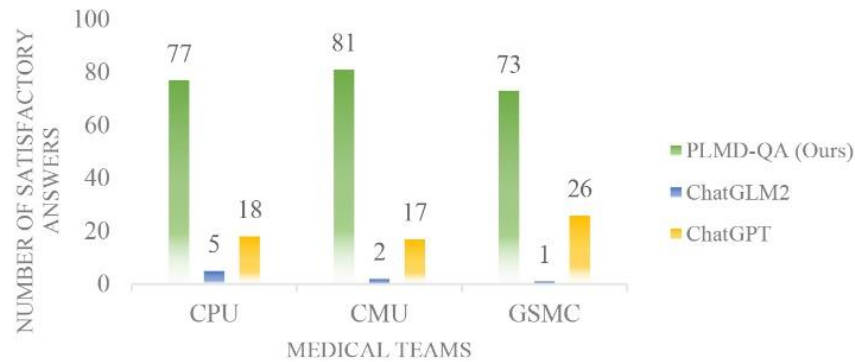


Figure 3. Evaluation of satisfactory answers from various models by three medical team

Objective performance evaluation

This experiment designed and organized 200 objective medical questions of varying difficulty levels, including multiple-choice, true/false, fill-in-the-blank, multi-select, ranking, matching, and clinical decision analysis questions. These questions were answered by different models, and the accuracy of each model on different types of questions was calculated to evaluate the system's objective performance. The questions were categorized by difficulty into three levels: Simple Questions (SQ), Moderate Questions (MQ), and Difficult Questions (DQ). SQ typically retrieve professional information directly from the knowledge graph without requiring complex reasoning. MQ require moderate reasoning and involve the association of multiple knowledge points. DQ demand complex reasoning and a deep understanding of medical knowledge.

According to the objective performance evaluation results in Table 13, the accuracy of responses decreases as question difficulty increases. For average accuracy, the proposed system significantly outperforms ChatGLM2 and slightly exceeds ChatGPT, indicating that the proposed question-answering system can significantly enhance the capabilities of large language models in the medical domain. From the perspective of the number of correctly answered questions, the proposed system answers more questions correctly, objectively validating its performance.

Table 13. Objective performance evaluation

Model	Accuracy (%)				
	SQ(120)	MQ(60)	DQ(20)	All	Avg
ChatGLM2	68.3%(82)	55.0%(33)	20.0%(4)	59.5%(119)	47.8%
ChatGPT	85.0%(102)	71.7%(43)	45.0%(9)	77.0%(154)	67.2%
PLMD-QA (Ours)	90.8%(109)	80.0%(48)	35.0%(7)	82.0%(164)	68.6%

CONCLUSION

This study proposed an intelligent question-answering system that integrates large language models with a dynamic medical knowledge graph, aiming to provide convenient and accurate question-answering services for healthcare practitioners and patients, assist in medical decision-making and diagnosis, and enhance the quality and efficiency of healthcare services. Based on multiple experiments, the following conclusions were drawn:(1) A large-scale dynamic medical knowledge graph integrating multi-source data was constructed, encompassing 86,211 medical entities and 703,281 relationships. The system consolidated fragmented medical knowledge and implemented structured and standardized knowledge storage and representation on the Neo4j platform, providing a comprehensive information foundation for medical research. (2)High-accuracy models were developed for named entity extraction (RBBSC) and intent recognition (RBTRC). The RBBSC model, based on the improved RoBERTa-WWM and combining Self-Attention mechanisms with BiLSTM, effectively captured contextual information and achieved significant performance improvements in medical named entity extraction, with an F1-score of 0.782. The RBTRC model, which combines the strengths of TextCNN and TextRNN, demonstrated the ability to capture both global and local features, achieving an accuracy of 0.8703, outperforming other baseline methods. (3)A novel question-answering paradigm based on P-Tuning, integrating

large language models with a medical knowledge graph, was proposed. This paradigm achieved professional and interpretable question-answering enhancements and multi-turn intelligent interactions. In question-answering tests, the system outperformed other models in response quality and medical team satisfaction. (4)The Med-BERT information filtering module was designed to effectively restrict the system from answering questions outside the medical domain, improving efficiency. Additionally, an entity alignment unit was introduced after entity extraction to better identify and understand the core aspects of user queries.

In summary, this study demonstrates the system's broad application potential in medical auxiliary consultations, patient health education, and medical research. It not only enhances the convenience of healthcare services but also promotes the development of medical informatization and intelligence.

REFERENCES

- [1] R. Wang and X. Zhang, "Medical knowledge graph question-answering system based on hybrid dynamic masking and multi-strategy fusion," *Journal of Frontiers of Computer Science & Technology*, vol. 18, no. 10, pp. 2770–2786, 2024.
- [2] X. Chen, S. Jia, and X. Yang, "A review: Knowledge reasoning over knowledge graph," *Expert Systems with Applications*, vol. 141, p. 112948, 2020.
- [3] X. Hu, J. Duan, and D. Dang, "Natural language question answering over knowledge graph: the marriage of SPARQL query and keyword search," *Knowledge and Information Systems*, vol. 63, no. 4, pp. 819–844, 2021.
- [4] Y. Lan, G. He, J. Jiang, J. Jiang, W. X. Zhao, and J.-R. Wen, "Complex knowledge base question answering: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 11, pp. 11,196–11,215, 2022.
- [5] Y. Chen, Y. Jia, C. Tan, W. Chen, and M. Zhang, "Method for complex question answering based on global and local features of knowledge graph," *Journal of Software*, vol. 34, no. 12, pp. 5614–5628, 2023.
- [6] G. Tan, Y. Chen, and Y. Peng, "Hybrid domain feature knowledge graph smart question answering system," *Computer Engineering and Applications*, vol. 56, no. 3, pp. 232–239, 2020.
- [7] D. Lukovnikov, A. Fischer, J. Lehmann, and S. Auer, "Neural network-based question answering over knowledge graphs on word and character level," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 1211–1220.
- [8] Y. Hao, Y. Zhang, K. Liu, S. He, Z. Liu, H. Wu, and J. Zhao, "An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 221–231.
- [9] P. Han and L. Gu, "Research on extraction of Chinese medical entities based on hybrid deep learning," *Library and Information Service*, vol. 66, no. 14, pp. 119–127, 2022.
- [10] M. Bakhshi, M. Nematbakhsh, M. Mohsenzadeh, and A. M. Rahmani, "Data-driven construction of SPARQL queries by approximate question graph alignment in question answering over knowledge graphs," *Expert Systems with Applications*, vol. 146, p. 113205, 2020.
- [11] Z. Ma, S. Wang, H. Zhang, and R. Li, "Joint model of intelligent QA intent recognition based on knowledge graph," *Journal of Computer Engineering & Applications*, vol. 59, no. 6, pp. 171–178, 2023.
- [12] Z. Yang, T. Xu, Z. Jin, M. Kong, G. Li, and X. Du, "Construction and application of knowledge graph of sheep goat disease," *Journal of Huazhong Agricultural University*, vol. 42, no. 3, pp. 63–70, 2023.
- [13] B. Ji, R. Liu, S. Li, J. Yu, Q. Wu, Y. Tan, and J. Wu, "A hybrid approach for named entity recognition in Chinese electronic medical record," *BMC Medical Informatics and Decision Making*, vol. 19, pp. 149–158, 2019.
- [14] Q. Xu, Y. Zhou, B. Liao, Z. Xin, W. Xie, C. Hu, and A. Luo, "Named entity recognition of diabetes online health community data using multiple machine learning models," *Bioengineering*, vol. 10, no. 6, p. 659, 2023.
- [15] V. Sciannameo, D. J. Pagliari, S. Urru, et al., "Information extraction from medical case reports using OpenAI InstructGPT," *Computer Methods and Programs in Biomedicine*, vol. 255, p. 108326, 2024.
- [16] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

- [17] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *International Conference on Machine Learning*, PMLR, 2020, pp. 3929–3938.
- [18] S. Lu, I. Bigoulaeva, R. Sachdeva, H. T. Madabushi, and I. Gurevych, "Are emergent abilities in large language models just in-context learning?" *arXiv preprint arXiv:2309.01809*, 2023.
- [19] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al., "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.
- [20] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," *arXiv preprint arXiv:2005.00661*, 2020.
- [21] X. Liu, K. Ji, Y. Fu, W. L. Tam, Z. Du, Z. Yang, and J. Tang, "P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks," *arXiv preprint arXiv:2110.07602*, 2021.
- [22] Y. Wang, B. Yu, Y. Zhang, T. Liu, H. Zhu, and L. Sun, "Tplinker: Single-stage joint extraction of entities and relations through token pair linking," *arXiv preprint arXiv:2010.13415*, 2020.
- [23] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for Chinese BERT," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3504–3514, 2021.
- [24] N. Fanourakis, V. Efthymiou, D. Kotzinos, and V. Christophides, "Knowledge graph embedding methods for entity alignment: experimental review," *Data Mining and Knowledge Discovery*, vol. 37, no. 5, pp. 2070–2130, 2023.
- [25] Q. Wang and X. Li, "Chinese news title classification model based on ERNIE-TextRCNN," in *Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing*, 2022, pp. 147–151.
- [26] P. C. Sukhwai, V. Rajan, and A. Kankanhalli, "A Joint LLM-KG System for Disease Q&A," *IEEE Journal of Biomedical and Health Informatics*, 2024.
- [27] D. Gavrilov and N. Balagansky, "Ahead-of-Time P-Tuning," *arXiv preprint arXiv:2305.10835*, 2023.
- [28] J. L. Jimmy, K. Nongmeikappam, and S. K. Naskar, "BiLSTM-CRF Manipuri NER with character-level word representation," *Arabian Journal for Science and Engineering*, vol. 48, no. 2, pp. 1715–1734, 2023.
- [29] D. Contractor, B. Patra, and P. Singla, "Constrained BERT BiLSTM CRF for understanding multi-sentence entity-seeking questions," *Natural Language Engineering*, vol. 27, no. 1, pp. 65–87, 2021.
- [30] X. Yin, S. Zheng, and Q. Wang, "Fine-grained Chinese named entity recognition based on RoBERTa-WWM BiLSTM-CRF model," in *2021 6th International Conference on Image, Vision and Computing (ICIVC)*, IEEE, 2021, pp. 408–413.
- [31] I. Alshubaily, "TextCNN with attention for text classification," *arXiv preprint arXiv:2108.01921*, 2021.
- [32] C. Eang and S. Lee, "Improving the accuracy and effectiveness of text classification based on the integration of the BERT model and a recurrent neural network (RNN_Bert_Based)," *Applied Sciences*, vol. 14, no. 18, p. 8388, 2024.
- [33] F. Guan and T. Tezuka, "A medical Q&A system with entity linking and intent recognition," in *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, 2022, pp. 820–829.