# Construction and Optimization of an Intelligent Medical Triage Recommendation System Based on Large Language Models

**Yan Yang[1,*], Wentao Zhang[2]**
*[1]Computer School, Central China Normal University, Wuhan, China*
*[2]Information Center, Wuhan University, Wuhan, China*
**Corresponding Author*

**Abstract: To address service pressures arising from surging online medical consultations and enhance triage accuracy and knowledge recommendation efficiency, this paper constructs and optimizes an AI-driven medical triage recommendation system based on large language models. By designing a hybrid architecture augmented with medical knowledge graphs and implementing efficiency-oriented model pruning and high-efficiency training strategies, the system achieves intelligent understanding of patient chief complaints, precise triage, and relevant knowledge recommendations. Experimental results demonstrate that the optimized system exhibits significant improvements in key performance metrics including triage accuracy, response speed, and answer safety. It efficiently supports physician decision-making, alleviates repetitive consultation burdens, and provides reliable self-service inquiries for patients. This system holds positive implications for optimizing healthcare resource allocation and alleviating service pressures.**

**Keywords: Large Language Model; Intelligent Triage; Recommendation System; Knowledge Graph**

## 1. Introduction

With the rapid development of "Internet + Healthcare," online medical platforms have become crucial channels for patients to seek health consultations, conduct initial symptom screening, and access medical knowledge [1][2]. However, the growing demand for consultations has created a sharp contradiction with the scarcity of high-quality medical resources [3]. Traditional online triage relies on patients self-selecting departments or manual customer service guidance, leading to issues such as inaccurate triage, delayed responses, and incomplete knowledge coverage. This may result in delayed diagnosis or repeated consultations, further burdening the healthcare system [4].

Large language models (LLMs), exemplified by the Transformer architecture, demonstrate exceptional capabilities in natural language understanding and generation, presenting new opportunities for developing intelligent medical assistants [5-7]. However, directly applying general-purpose LLMs to the highly rigorous and safety-critical medical domain poses three core challenges: First, the extreme demands for medical expertise and safety require models to eliminate "hallucinations" and provide evidence-based, reliable medical information [8][9]; Second, model inference efficiency and real-time responsiveness are critical, as online consultation scenarios demand rapid system responses [10][11]. Third, deep comprehension of complex medical logic and entity relationships is essential to achieve precise mapping from symptoms to departments, diseases, and clinical knowledge [12][13].

To address these challenges, this paper aims to construct and deeply optimize a medical triage recommendation system based on large language models tailored for online healthcare scenarios [14]. The core contributions are: 1) Developing a hybrid intelligent architecture that deeply integrates medical knowledge graphs with large language models, using graph retrieval to enhance generation accuracy and ensure answer traceability; 2) Optimizing model efficiency by employing task-specific model pruning and knowledge distillation to significantly accelerate inference while minimizing accuracy loss; 3) Enhancing the training process through distributed training and mixed-precision computation to efficiently learn from massive medical literature and dialogue data. Performance evaluations validate the

 

system's substantial improvements in triage accuracy, recommendation relevance, and service efficiency.

## 2. Construction of the Llm-Based Medical Triage Recommendation System

This chapter details the overall architecture design and core module construction of a medical large-model system tailored for online triage recommendation scenarios. The system aims to achieve a secure, reliable, and automated workflow, from patient natural language complaints to precise triage suggestions and knowledge recommendations, by deeply integrating medical domain knowledge with large language model capabilities [15-17].

### 2.1 Overall System Architecture Design

The system constructed in this paper adopts a three-tier cascaded hybrid intelligence architecture ("perception-reasoning-generation") (as shown in Figure 1) to balance the general capabilities of large models with the stringent requirements of the medical field for professionalism, safety, and interpretability.

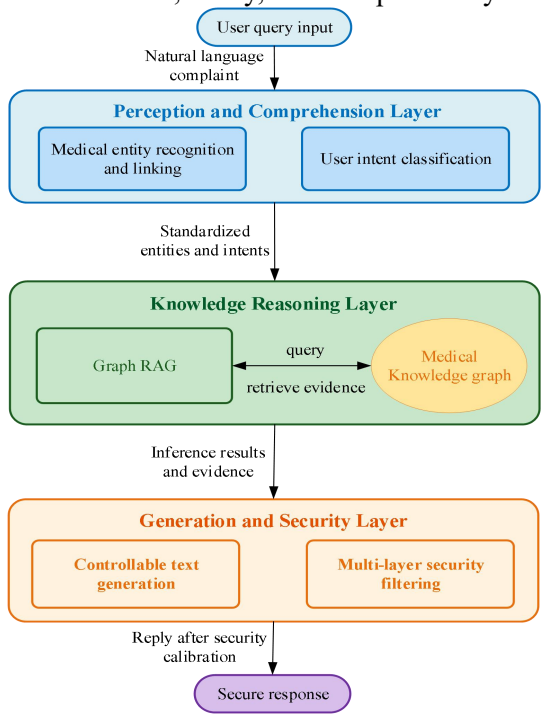


**Figure 1 Architecture of the Medical Large-Model Triage Recommendation System**

●Perception and Comprehension Layer: Processes raw user input, performs intent recognition, and structures information.

●Knowledge Reasoning Layer: Serves as the system's "professional brain," performing logical reasoning and decision-making based on medical knowledge graphs.

●Generation and Safety Layer: Organizes language output and enforces rigorous content security calibration.

The core concept of this architecture is to combine open semantic understanding capabilities with closed, deterministic knowledge, ensuring the system is both flexible and reliable.

### 2.2 Construction of Core Modules

2.2.1 Perception and Understanding Module

This module serves as the entry point for system-user interaction. It first preprocesses and standardizes user-input text. Subsequently, it employs a domain-specific large model, continuously pre-trained on extensive medical corpora, to accomplish two core tasks [18]:

1) Medical Entity Recognition and Linking: Identifies entities such as symptoms, signs, diseases, medications, and diagnostic tests within patient complaints, linking them to standardized concept nodes in the system's knowledge graph to resolve ambiguities in colloquial expressions [19].

2) User Intent Classification: Determines the core intent behind user queries, such as "disease diagnosis consultation," "medication usage inquiry," "department confirmation," or "health knowledge dissemination," providing a basis for subsequent differentiated processing.

2.2.2 Knowledge Reasoning Module

This module forms the core of the system's specialized triage and recommendation capabilities, whose effectiveness relies on a high-quality knowledge graph and advanced retrieval-inference mechanisms [20].

●Medical Knowledge Graph Construction: We built a domain-specific knowledge graph covering diverse entities including diseases, symptoms, departments, medications, tests, and treatment plans. The graph schema references international standards (e.g., UMLS) and integrates authoritative textbooks, clinical guidelines, and de-identified electronic health records to ensure breadth and depth of knowledge. Entities are linked by semantic relationships such as "belongs to," "causes," "requires examination," and "commonly used drugs," enabling complex multi-hop reasoning.

●Image Retrieval Enhanced Generation Mechanism: To avoid the "illusion" of large models, this system adopts the

 

retrieval-enhanced generation paradigm. Specifically, based on entities extracted by the perception module, multi-hop retrieval is performed within the knowledge graph to obtain relevant triplet subgraphs and associated authoritative text fragments. Subsequently, this structured and semi-structured knowledge serves as "evidence," collaborating with the large model's own parameterized knowledge to jointly generate intermediate reasoning results (e.g., suspected disease lists, recommended departments, and urgency ratings). This approach, similar to the cutting-edge MedGraphRAG technique, significantly enhances answer traceability and accuracy.

2.2.3 Generation and Safety Module

This module transforms internal reasoning outputs into natural, understandable, and safe final responses.

●Controlled Text Generation: Using the reasoning results and retrieved knowledge from the previous module as strong constraints, it guides the large model to generate responses. Response templates explicitly include triage conclusions, brief rationale summaries, and relevant knowledge prompts, ensuring structured and readable information.

●Multi-Layer Safety Filtering Mechanism: Safety is paramount in medical contexts. This system implements a three-tier filtering system:

1)Rule-based filtering: Identifies critical keywords (e.g., "chest pain radiating to the back," "sudden severe headache") triggering highest-priority alerts and directly recommending emergency medical attention.

2)Probability filtering: Applies safety classifiers to assign risk scores to generated content, intercepting high-risk or highly uncertain statements.

3)Evidence-based fallback: All generated medical factual statements must be verifiable against retrieved evidence in the knowledge graph; otherwise, they are marked as "Pending Verification" or withheld from output.

## 3. System Optimization

After completing the system's foundational architecture, we implemented specialized optimizations for the core model and training workflow to meet the stringent requirements of online services for high concurrency and low latency, while further enhancing performance and efficiency.

### 3.1 Model Optimization for Efficiency Enhancement

3.1.1 Task-Specific Structural Pruning

For the constructed medical large model, we analyzed the importance of attention heads and feedforward networks across layers in triage question-answering tasks. Using gradient-based structured pruning, we removed redundant parameters contributing minimally to the final task. Experiments demonstrated that this approach reduced model parameters by approximately 35% while limiting accuracy loss on core medical understanding tasks to under 1%.

3.1.2 Hierarchical Knowledge Distillation

To achieve a lighter, more efficient deployment model, we designed a three-tier "teacher-assistant-student" distillation framework. First, medical diagnostic capabilities from a general-purpose, ultra-large-scale "teacher model" are distilled to an "assistant model" with moderate parameters (e.g., 30% of the teacher model). Subsequently, the precise judgment capabilities demonstrated by the "Assistant Model" on specific triage dialogue data are further distilled into the final "Student Model" (with parameters reduced to just 10% of the Teacher Model). This hierarchical strategy enables the Student Model to maintain performance close to the Assistant Model despite significant parameter compression. The comparative effectiveness of model optimization strategies is detailed in Table 1.

**Table 1. Comparison of Model Optimization Strategy Effects**

| Model Version | Number of Parameters | MedQA Accuracy (%) | Average Response Time (ms) |
|---|---|---|---|
| Original base model | 7B | 72.3 | 3500 |
| After medical incremental pre-training | 7B | 85.1 | 3200 |
| After task-specific pruning | 4.55B | 84.2 | 2100 |
| After hierarchical knowledge distillation (deployment version) | 0.7B | 83.5 | 650 |

### 3.2 Efficient Training Paradigms for Massive Medical Data

To support the training and iteration of the

 

aforementioned models, we adopted a hybrid parallel training framework.

1)Data and Pipeline Parallelism: Distributed loading of over 100GB of de-identified clinical text and medical literature data onto multiple GPU nodes. Pipeline parallelism techniques decompose model layers across different devices, combined with gradient accumulation, resolving GPU memory bottlenecks for large models and big data.

2)Dynamic Mixed-Precision Training: Accelerates computations using FP16 precision during forward and backward propagation, while retaining the FP32 master copy during optimizer updates to ensure numerical stability. This strategy boosts training throughput by approximately 40%, significantly shortening model iteration cycles.

## 4. System Performance Evaluation and Analysis

### 4.1 Experimental Setup

We evaluated the system's performance across multiple dimensions using the following datasets and metrics:

●Dataset: Evaluation utilized the public benchmarks MedQA (US Medical Licensing Examination question bank) and PubMedQA (medical literature-based question-answering), along with a de-identified real patient consultation dataset (covering 6 departments including Internal Medicine, Pediatrics, and Dermatology, totaling 50,000 entries) obtained from a collaborative platform.

●Evaluation Metrics: Accuracy and recall for triage and Q&A; relevance of knowledge recommendations (manually scored by three attending physicians); end-to-end system response time; and hallucination rate (proportion of generated information not derivable from source knowledge) for safety.

### 4.2 Results and Analysis

Based on the aforementioned experimental setup, we conducted comprehensive testing and analysis of the system, yielding the following key results.

1)Triage and Q&A Accuracy: On the MedQA test set, the system achieved 88.7% accuracy, significantly outperforming the baseline general model (72.3%). For real patient consultation data, automatic triage accuracy to the correct primary department reached 92.5%, and 89% of the top three recommended knowledge snippets were rated "highly relevant" by physicians.

2)Efficiency and Real-time Performance: After optimization, the system achieved an average response time of 780 milliseconds on standard servers, fully meeting online interaction requirements (<2 seconds). Throughput testing demonstrated the ability to handle over 1,000 concurrent consultations simultaneously.

3)Safety Analysis: Through knowledge graph constraints and safety filtering, the system's "hallucination rate" on the test set is controlled below 3%. For identified critical condition keywords (e.g., "chest pain with suffocation sensation," "acute severe headache"), the system successfully triggers the highest-level alert 100% of the time, forcibly terminates the conversation, and recommends immediate emergency care. Comprehensive performance evaluation data is detailed in Table 2.

**Table 2. Summary of Comprehensive Performance Evaluation**

| Evaluation Dimension | Metric | Pre-optimization (Baseline) | Post-Optimization (Proposed System) | Improvement |
|---|---|---|---|---|
| Accuracy | MedQA accuracy | 72.3% | 88.7% | +16.4% |
| | Real-world data triage accuracy | 75.0% (Estimated) | 92.5% | +17.5% |
| Efficiency | Average response time | ~3000ms | <800 ms | Approximately 3.75 times improvement |
| Safety | Hallucination rate | >15% (estimated) | <3% | Reduced by over 80% |
| Practicality | Knowledge recommendation relevance score | - | 89% (highly relevant) | - |

### 4.3 Discussion of Limitations

The construction and optimization of this system still face limitations: First, system performance heavily relies on the completeness and timeliness of the knowledge graph, with delays in updating knowledge for rare and emerging diseases. Second, the model's support for non-textual information (e.g., images, speech) remains incomplete, whereas dermatology and radiology consultations often require visual data. Future work will explore

 

multimodal information fusion and continuous self-evolving knowledge graph update mechanisms.

## 5. CONCLUSIONS

This paper addresses the core requirements of online healthcare services by successfully constructing a medical big model system for intelligent triage and knowledge recommendation, subjecting it to comprehensive in-depth optimization. Through designing a knowledge graph-enhanced hybrid architecture, implementing efficiency-oriented model pruning strategies, and adopting high-efficiency training paradigms, the system achieves practical levels of accuracy, security, and response efficiency. Experiments demonstrate the effectiveness of the system's construction methodology and optimization strategies. It serves as a powerful physician assistant, enhancing primary-level triage efficiency, while also functioning as a reliable "pre-consultation desk" for patients, providing precise medical knowledge navigation. This comprehensively optimizes healthcare resource allocation and alleviates service pressure. Moving forward, we will focus on building a full-process intelligent assistance system covering "pre-consultation, during consultation, and post-consultation" stages, while exploring its application in broader fields such as medical education and clinical decision support.

## References

[1] QIAN B, LI F, ZHENG C, et al. Development Status and Prospects of Large Medical Models [J]. Data Acquisition and Processing, 2025, 40(03): 562-584.

[2] CHU J. Research and Application of Intelligent Medical Question Answering Based on Large Models [D]. Anhui Jianzhu University, 2025.

[3] XING X. Research and Application of Medical Question Answering Assistant Integrating Large Models and Knowledge Graphs [D]. Hefei University, 2025.

[4] REN J, ZHANG Z, XIANG S, et al. Application of Artificial Intelligence Large Model Technology in the Medical and Health Industry [J]. Communication World, 2025, (01): 42-44.

[5] WU Y, LU J. Multi-modal Information Generation and Recommendation Driven by Large Models [J]. Journal of Henan Normal University (Natural Science Edition), 2025, 53 (05): 145-151+181.

[6] NIU Y, HAO B, ZHAO Z. Research on the Construction and Practice of Personalized Resource Recommendation System for University Libraries Based on Large Models [J]. New Century Library, 2025, (07):66-73.

[7] YOU X, LI S, SHAO H. Construction of an Intelligent Question and Answer System for Clothing Recommendation Based on Large Models [J]. Woolen Textile Technology, 2025, 53 (05): 87-94.

[8] YANG Y, PAN S, LIU X, et al. Multi-modal Knowledge Graph and Collaborative Decision-Making of Large Models for Risk Management in Hydraulic Engineering [J]. Journal of Water Resources, 2025, 56 (04): 519-530.

[9] ZHOU X. Research on Efficient Recommendation Algorithm Based on Large Language Model [D]. University of Electronic Science and Technology of China, 2025.

[10] LIU P, ZHANG M, WANG P, et al. Algorithm Optimization and Performance Evaluation of Intelligent Knowledge Recommendation System for Convenience Hotline Work Orders Based on Large Models [J]. Digital Technology and Applications, 2025, 43 (03): 16-18.

[11] WANG M, GAO X, WANG S, et al. Research on Recommendation System Based on Knowledge Graph and Large Language Model Enhancement [J]. Big Data, 2025, 11 (02): 29-46.

[12] MA X, GAO J, LIU Y, et al. Construction of a Customer Service Knowledge Recommendation Model Driven by Intent Understanding [J]. Journal of South China University of Technology (Natural Science Edition), 2025, 53 (03): 40-49.

[13] ZHANG Y. Research on the Construction and Recommendation of Course Content Knowledge Graph Based on Large Language Model in Smart Education [D]. Sichuan Normal University, 2024.

[14] KA Z, ZHAO P, ZHANG B, et al. A Review of Recommendation Systems for Large Language Models [J]. Computer Science, 2024, 51 (S2): 11-21.

[15] ZHANG X, ZHANG L, YAN S, et al. Personalized Learning Recommendation Based on Knowledge Graph and Large Language Model Collaboration [J]. Computer Applications, 2025, 45 (03):

 

773-784.
[16] ZHU M. Research on Personalized Resource Recommendation Method Based on Large Language Model [J]. Journal of Lanzhou University of Arts and Sciences (Natural Science Edition), 2024, 38 (05): 59-64.
[17] WU G, QIN H, HU Q, et al. Research on Large Language Models and Personalized Recommendations [J]. Journal of Intelligent Systems, 2024, 19 (06): 1351-1365.
[18] ZHANG X, TAN K, OUYANG T, et al. Design and Implementation of Personalized Exercise Recommendation System Based on Large Model [J]. Digital Technology and Applications, 2024, 42 (07): 32-34.
[19] LIU L. Research and Application of Professional Recommendation Knowledge Graph Construction Technology Based on Large Language Model [D]. Hangzhou University of Electronic Science and Technology, 2024.
[20] YE C. Overview of Large Language Model Recommendation Techniques [J]. Electronic Components and Information Technology, 2023, 7 (12): 127-131.

 