
# MedCompass - Smart Healthcare Chatbot

Final Project Report Submitted to
The Department of Computer Science
Faculty of Computer and Information Technology
Jordan University of Science and Technology
In Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Computer Science

Prepared by

Sara Alsanajleh [163418]
Ghada AbuShaqra [164188]
Maian Alabweh [161399]

Supervisor
Rasha Obaidat

January 2026

# نموذج حقوق الملكية الفكرية لمشاريع التخرج في قسم علوم الحاسوب

**يتم قراءة وتوقيع هذا النموذج من قبل الطلاب المسجلين لمشاريع التخرج في قسم علوم الحاسوب**

تعود حقوق الملكية الفكرية لمشاريع التخرج ونتائجها (مثل براءات الاختراع أو أي منتج قابل للتسويق) إلى جامعة العلوم والتكنولوجيا الأردنية. وتخضع هذه الحقوق إلى قوانين وأنظمة و تعليمات الجامعة المتعلقة بالملكية الفكرية وبراءات الاختراع.

بناءا على ما سبق أوافق على ما يلي:

1) أن أحفظ كافة حقوق الملكية الفكرية لجامعة العلوم والتكنولوجيا الأردنية في مشروع التخرج.
2) أن ألتزم بوضع اسم جامعة العلوم والتكنولوجيا الأردنية و أسماء جميع الباحثين المشاركين في المشروع على أي نشرة علمية للمشروع كاملا أو لنتائجه. و يشمل ذلك النشر في المجلات و المؤتمرات العلمية عامة أو النشر على المواقع الإلكترونية أو براءات الاختراع أو المسابقات العلمية.
3) أن ألتزم بأسس حقوق التأليف المعتمدة في جامعة العلوم والتكنولوجيا الأردنية.
4) أن أقوم بإعلام الجهة المختصة في الجامعة عن أي اختراع أو اكتشاف قد ينتج عن هذا المشروع و أن ألتزم السرية التامة في ذلك و أن أعمل من خلال الجامعة على الحصول على براءة الاختراع التي قد تنتج عن هذا المشروع.
5) أن تكون جامعة العلوم والتكنولوجيا الأردنية هي المالك لأي براءة اختراع قد تنتج عن هذا المشروع و تشمل هذه الملكية حق الجامعة في إعطاء التراخيص و التسويق و البيع كمؤسسة راعية و داعمة لكافة الأنشطة البحثية. و يكون حق للطالب شمول اسمه على براءة الاختراع كأحد المخترعين، و في حال تم إعطاء تراخيص أو تسويق و بيع لأي من منتجات المشروع يمنح المخترعون بما فيهم الطالب نسبة من الإيرادات حسب تعليمات البحث العلمي في جامعة العلوم والتكنولوجيا الأردنية.

إسم الطالب غادة ماهر أبو شقره التوقيع

إسم الطالب سارة جهاد السناجلة التوقيع

إسم الطالب ميان ايمن العبوة التوقيع

**إسم المشرف** رشا العبيدات **التوقيع**

تاريخ 20.1.2026

# Smart Medical Chatbot Using Multi-Agent Architecture and Retrieval-Augmented Generation

Ghada AbuShaqra, Sara Alsanajleh, Maian Alabweh and Rasha Obeidat[§]
Jordan University of Science and Technology Irbid, Jordan
{gmabushaqra22, sjalsanajleh22, maalabweh22}@cit.just.edu.jo
[§]rmobeidat@just.edu.jo

***Abstract*—Hospitals continue to face increasing pressure at the early stage of care delivery due to rising patient volumes, limited medical resources, long waiting times, and inefficient manual intake and routing processes. This paper presents the design and initial implementation of an AI-based patient intake and routing assistant that supports early-stage hospital workflows without providing clinical consultation or diagnosis. The system conducts structured multi-turn medical conversations in Arabic to collect patient information, identify urgent warning signs, recommend appropriate medical departments, and complete appointment booking through a connected database. The proposed solution integrates large language models for natural dialogue, retrieval-augmented generation to ground responses in verified medical knowledge, and a multi-agent architecture that separates registration, conversation management, medical retrieval, and scheduling into dedicated agents. A shared internal state ensures consistent workflow control and prevents conversational drift. A functional prototype is implemented using the Llama 3.1 large language model and is validated through scenario-based testing. The system demonstrates successful end-to-end execution from patient registration to confirmed appointment scheduling, produces reliable medical department recommendations grounded in retrieved knowledge, and correctly interrupts the workflow when emergency indicators are detected. These results confirm the technical feasibility of the proposed approach and highlight its potential to reduce administrative workload and improve patient routing within real hospital settings.**

***Index Terms*—medical chatbot, large language models, retrieval-augmented generation, multi-agent systems, Arabic medical dialogue, patient routing, appointment scheduling**

## I. Project Goals and Objectives

The primary goal of this project is to support healthcare systems during the early stages of service delivery by reducing operational pressure and improving overall efficiency. The system aims to assist hospitals in managing increasing patient volumes and service requests, while enabling more efficient use of available medical resources and improve the overall flow of patients from the first point of contact.

**The main objectives of the project include**

- reducing congestion in outpatient clinics and shortening waiting times during patient intake
- lowering administrative workload on medical staff by supporting basic intake and routing processes
- improving patient experience through clearer guidance and smoother early interaction with healthcare services
- enhancing patient routing by directing patients to appropriate departments from the beginning and reducing unnecessary referrals
- supporting patient safety by encouraging early recognition of urgent cases and prioritizing timely medical intervention
- demonstrating a practical and scalable solution that aligns with real hospital workflows and can be extended in the future

## II. Introduction

The rapid growth in healthcare demand, together with limited medical resources and rising patient expectations, has increased pressure on hospitals and medical centers. In many clinics, patients continue to experience crowded waiting areas, slow service delivery, and unclear routing between departments. As patient volumes grow, healthcare providers are required to allocate additional time, staff, and operational resources to manage waiting lists and prioritize patients. This situation increases the workload on medical teams and contributes to operational stress. Moreover, prolonged waiting times and unclear care processes lead to higher levels of patient frustration and reduced satisfaction, as highlighted in OECD reports [1]. These challenges are particularly evident at the early stage of care, where registration, symptom collection, and department assignment are often handled manually. Manual intake processes can introduce delays, miscommunication, and avoidable referrals, negatively affecting both patient experience and clinical efficiency.

In recent years, advances in artificial intelligence have made early-stage care processes easier to automate in a safe and controlled way, with the World Health Organization (WHO) [2] noting AI's potential to enhance health outcomes and support person-centred care when appropriately governed, thereby enabling a transformative shift in healthcare systems. Artificial Intelligence (AI), Large Language Models (LLMs), Multi-Agent Systems (MAS), and Retrieval-Augmented Generation (RAG) are now widely discussed as practical tools for building conversational healthcare support systems. LLMs can understand free-text patient descriptions and keep a natural conversation, but they may produce incorrect content if they work alone. RAG helps reduce this risk by grounding responses in retrieved documents instead of relying only on the model's internal memory, as introduced in the original RAG framework by Lewis et al. [3]. MAS design also supports reliability by splitting the workflow into smaller agents, where

each agent has one clear job, which improves control and testing compared to a single large chatbot module [4].

In many cases, systems either use an LLM without strong grounding, or use retrieval without a structured agent workflow. This creates a gap between research prototypes and real hospital workflows, where the system must do more than answer questions. It should collect symptoms gradually, detect urgency, recommend a department, and then complete a real booking step while keeping data consistent.

To address this gap, this paper presents the design and implementation of a Smart Medical Chatbot that combines LLMs, MAS, and RAG into one end-to-end system for early-stage healthcare services. The chatbot operates primarily in Arabic to support natural interaction for local users, while following strict safety boundaries by avoiding diagnosis and focusing on routing and workflow support. The system is built around a state-driven pipeline, where a shared internal state tracks patient demographics, extracted symptoms, urgency signals, recommended departments, and booking details. This design makes the workflow traceable and prevents random conversation drift.

The implementation achieves a working prototype that is validated through scenario-based testing. The system successfully executes a full workflow from patient registration to appointment booking, extracts structured medical information from unstructured Arabic text across multiple turns, produces consistent department recommendations grounded by retrieval, and correctly interrupts the workflow when emergency indicators are detected. These results show that the proposed approach can operate as a unified service workflow, not as separate disconnected features, and can serve as a practical front-line gateway that reduces administrative load and improves patient routing.

The main conclusion of this work is that combining MAS control with RAG grounding offers a more reliable direction for real clinical workflows than using a single-component chatbot. This project provides a functional foundation that can be extended with quantitative evaluation, larger knowledge coverage, and tighter integration with hospital information systems.

The remainder of this paper is organized as follows. Section III reviews the related work on medical chatbots, intelligent healthcare systems, and AI-based triage approaches, with a focus on identifying the limitations of existing solutions and the research gap addressed in this study. Section IV describes the proposed methodology and system architecture, including the database layer, the RAG pipeline, and the MAS workflow. Section V presents the experimental setup, baselines, and evaluation metrics. Section VI reports the functional results and scenario-based testing outcomes. Finally, Section VII concludes the paper and discusses limitations and possible future improvements.

## III. RELATED WORK

The healthcare sector is facing increasing pressure due to the growing number of patients, the shortage of medical staff, and the large amount of time spent on administrative tasks instead of direct patient care. This situation has created a clear need for systems that can reduce the burden on healthcare facilities. Such systems can handle a large number of patients and improve workflow by scheduling appointments, reminding patients, and reducing missed visits. This highlights the practical benefit of medical chatbots.

At a global level, the adoption of healthcare chatbots is supported by international collaboration and regulatory encouragement. As stated by Barreda et al. [5], the use of healthcare chatbots has been increasing due to international collaboration and knowledge sharing. The (WHO) has recognized the potential of artificial intelligence to improve public health and has encouraged the development of guidelines and standards for the implementation of these technologies. These initiatives aim to ensure global access to the benefits of healthcare chatbots and to promote equity and quality in healthcare services.

### *1) **Clinical and Administrative Chatbots**:*

Healthcare chatbots are designed to perform a wide range of functions that support both patients and healthcare systems. Laymouna et al. [6] highlighted that health care chatbots perform a variety of roles within the health care system, encompassing both clinical and non-clinical tasks. These roles include patient data collection, facilitation of health information flow, symptom assessment, support for health education, disease management, and research data gathering. Collectively, these functions enhance workflow efficiency and enable the provision of care to a larger patient population. Hindelang et al. [7] conducted a systematic review to evaluate the impact of chatbots on medical history-taking. They found that these systems are capable of efficiently collecting medical information, enhancing patient engagement, and facilitating data acquisition, thereby reducing the workload of healthcare staff and accelerating processes within the healthcare system. Nelson et al. [8] discussed how AI-powered intelligent assistants can reduce pressure on hospital triage and appointment scheduling processes by using natural language processing and machine learning. These systems allow the assessment of patient symptoms, analysis of patient data, and prioritization of cases based on urgency. In addition, they can automatically schedule appointments more efficiently than traditional manual methods. Li et al. [9] conducted a study to evaluate an artificial intelligence tool called Smart-doctor in a pediatric outpatient clinic. The results showed that the use of this AI system significantly reduced patient waiting time, which decreased from about 21.8 minutes in the conventional system to around 8.8 minutes with AI assistance. In addition, both consultation time and total visit time were reduced. Parental satisfaction increased by approximately 17.5%. The study also reported lower costs for medical tests and medications compared to the traditional approach.

### *2) **Large Language Models in Healthcare**:*

LLMS have introduced advanced conversational capabilities

into healthcare, enabling more natural interaction and clinical support. Yuan et al. [10] indicate that LLMS possess advanced capabilities in medical knowledge retrieval, textual response generation, and dialogue-based interaction within healthcare contexts, in addition to their role in supporting clinical decision-making processes. Furthermore, they highlight that the adoption of such models can contribute to reducing the workload of healthcare professionals by automating routine tasks, including medical documentation, report summarization, and clinical workflow support. As noted by Wang et al. [11], LLMs have made significant progress, becoming capable of generating natural text and performing complex tasks. This advancement is particularly evident in the healthcare sector, where LLMs support diagnosis, facilitate patient-provider communication, and answer medical questions. However, these models still face limitations, such as hallucinations and reliance on outdated information, highlighting the importance of employing RAG techniques to address these challenges. Iqbal et al. [12] explained in their review that large language models such as ChatGPT have a clear ability to support disease diagnosis and clinical decision-making, which can enhance the quality of care provided to patients. These findings highlight the importance of investing in such models to develop interactive systems that help guide patients and perform initial patient triage, which directly aligns with the aim of this project to design a medical assistant capable of improving the patient journey before reaching a specialist physician.

*3) **Prompt Engineering for Medical LLM Applications**:*

Advanced prompt engineering offers a practical alternative to fine-tuning for improving medical LLM performance. As noted by Maharjan et al. [13], prompt engineering can outperform fine-tuning approaches in medical question-answering tasks, even when using open-source models. The OpenMedLM platform showed that improving prompt design, including few-shot prompting, chain-of-thought reasoning, and self-consistency, is able to raise model performance to advanced levels without the need for additional training data or high computational cost. Their results confirm that relying on open-source models combined with strong prompt engineering represents a promising direction for developing safe and reliable medical applications, which directly supports our methodology of building an LLM-based system without applying fine-tuning.

*4) **AI-Based Medical Triage and Department Recommendation**:*

Medical triage is a critical application area where chatbot and LLM performance must be evaluated against clinical standards. Alwaked et al. [14] used an intelligent chatbot as a tool to support the triage process and assist nursing staff in the emergency department. The system was designed to help speed up the initial assessment stage and organize patient flow within the department. It also showed a supportive role for nurses, especially less experienced or non-clinical staff, by helping them classify case severity into different levels and guide patients according to their priority. A study by Masanneck et al. [15] showed that LLMS such as GPT-4-based ChatGPT can triage patients in emergency departments with a level of accuracy similar to untrained doctors, although they do not reach the level of professional experts, which highlights their future potential as a clinical decision support tool with a need for further development. Sarbay et al. [16] studied the performance of the ChatGPT model in predicting case severity in emergency medicine using 50 medical scenarios generated according to the ESI v4 triage guideline. Emergency medicine specialists' assessments were used as the reference standard, and the model's predictions were compared with them. The results showed a limited level of agreement between ChatGPT and physicians (Cohen's Kappa = 0.341). The overall sensitivity was 57.1% and the specificity was 34.5%, with an F1 score of 0.461. However, the model showed better performance when dealing with high-acuity cases (ESI-1 and ESI-2), achieving a sensitivity of 76.2%, a specificity of 93.1%, and an F1 score of 0.821. This indicates that the model is more capable of identifying high-risk emergency cases than lower-acuity conditions. Systems based on LLMs, such as ChatGPT 4.0, have the potential to improve medical triage, potentially enhancing both the efficiency and quality of care. As demonstrated by Gebrael et al. [17], ChatGPT showed high sensitivity in identifying patients with metastatic prostate cancer who require hospital admission. Shi et al. [18] introduced C-PATH, a medical chatbot system based on LLMS. The system aims to collect patient symptoms through multi-turn conversations and then analyze them to guide patients to the appropriate medical departments using an initial triage process. Their results show that C-PATH can generate clear and understandable medical conversations for users and provide accurate department recommendations based on clinical medical knowledge. This highlights the potential of such systems to improve patient experience and facilitate access to appropriate healthcare services.

*5) **Retrieval Augmented Generation in Medical Systems**:*

RAG is essential in medical applications because it reduces hallucinations and grounds model responses in verified medical knowledge. Yu He et al. [19] developed an LLM integrated with a RAG framework using 35 preoperative clinical guidelines. The system was evaluated through 1,260 assessed responses, including 336 human-generated answers, 336 LLM-only responses, and 588 LLM-RAG responses. The findings showed that GPT-4.0 achieved the highest accuracy among the baseline LLMs at 80.1%, which further improved to 91.4% when enhanced with RAG. This performance exceeded the accuracy of human-generated instructions, which reached 86.3%. In addition, the LLM-RAG system produced responses within 15–20 seconds, compared to approximately 10 minutes required by human evaluators, demonstrating non-inferior performance to humans with a low hallucination rate. Patil et al. [20] used an intelligent medical system called RAGMed. This system is a medical assistant based on an LLM supported by the RAG approach. The system aims to improve answers

to patient questions, make appointment scheduling easier, and summarize clinical notes. The results showed that combining a medical knowledge base with vector retrieval methods improved answer accuracy, reduced hallucinations, and helped lower the administrative workload on medical staff. Zhao et al. [21] proposed MedRAG, a framework based on RAG enhanced with a diagnostic knowledge graph for the medical domain. The system retrieves relevant electronic health records and integrates them with the diagnostic knowledge graph to process patient manifestations, enabling the generation of accurate diagnoses and tailored treatment recommendations. Experimental results demonstrate that MedRAG outperforms traditional RAG models in terms of diagnostic accuracy and specificity, while also providing proactive follow-up questions to support more informed medical decision-making.

*6) **Multi Agent Systems in Healthcare**:*

MAS provide a dynamic and adaptive framework for designing healthcare systems, contributing to improved resource management and enhanced coordination of patient care. As explained by Borkowski and Ben-Ari [22], multi-agent artificial intelligence systems have the ability to improve medical decision-making and coordinate the patient journey within hospitals by distributing tasks among specialized agents that work collaboratively. The study indicates that this type of system can enhance diagnostic accuracy, organize patient flow, and support clinical resources, which directly aligns with the objective of our proposed system. A review by Isern and Moreno [23] indicated that the use of intelligent agents and MAS in healthcare is increasing. These systems provide flexibility and adaptability to manage the complexity and diversity of healthcare delivery. The applications reviewed encompassed various healthcare objectives and involved multiple stakeholders, suggesting promising opportunities for future research in this field.

*7) **Research Gap**:*

Based on reviewing previous studies in the field of medical chatbots, it can be observed that most existing systems focus on only one or two artificial intelligence components. **None of these papers** present a fully integrated solution that combines conversational intelligence, medical knowledge grounding, and structured system-level task management within a single framework. Some studies mainly rely on LLMs to conduct medical conversations, but without grounding the responses in verified medical knowledge. Other studies use RAG techniques to improve answer accuracy, yet they often apply them within simple system architectures that do not clearly separate tasks or manage the dialogue flow in a structured way. As a result, the integration between conversational intelligence, medical knowledge retrieval, and system-level task management remains limited.

In addition, the majority of existing medical chatbot systems are designed and evaluated primarily in English, with **limited support for the Arabic language**, which creates a clear gap in providing effective medical conversational systems for Arabic-speaking users. This work aims to address this gap by **integrating three core components into a single unified system**: LLMs for natural medical dialogue, RAG to support responses with trusted medical knowledge, and a multi-agent architecture to distribute medical and administrative responsibilities in a clear and organized manner. By combining these components, the proposed system seeks to improve reliability, reduce unsupported or hallucinated responses, and ensure a logical interaction flow starting from patient registration, through symptom collection and department identification, and finally appointment booking.

The summary of the reviewed related work is presented in Table I.

## IV. Methodology

This section describes the methodology followed to design and develop the proposed smart medical chatbot system. The methodology is divided into several interconnected components, including the large language model layer, prompt engineering, retrieval-augmented generation, multi-agent architecture, database design, and additional AI-driven features.

### *A. Large Language Models and Prompt Engineering*

The system relies on large language models (LLMs) to conduct medical conversations with patients and guide them toward the appropriate medical department. At the current stage, the system is implemented using Llama 3.1 as the main conversational model. In future stages, additional models such as Qwen 2.5, Qwen 3, and Falcon 3 will be evaluated and compared to study their behavior and suitability for medical dialogue in Arabic.

Since no labeled training dataset is available, the behavior of the models is controlled using prompt engineering rather than fine-tuning. The prompts are carefully designed to define the role of the assistant, limit its responsibilities, and ensure safe medical interaction. The model is instructed to act as a medical assistant in a Jordanian hospital, communicate in Arabic while understanding local expressions, ask structured and sequential questions, and avoid providing any final medical diagnosis. Instead, its role is limited to collecting symptoms, identifying potential emergency cases, and routing the patient to the appropriate department.

Several types of prompts are used within the system. These include a system prompt that defines strict behavioral rules, a greeting prompt to initiate the conversation, prompts to generate the next medical question based on previously collected information, and a department extraction prompt that forces the model to output only the name of the most suitable department. In addition, a dedicated emergency detection prompt is used to identify high-risk symptoms that require immediate medical attention.

To clarify how prompt engineering is applied in practice, an example of the system prompt used in the proposed chatbot is illustrated in Figure 1.

TABLE I: Summary of Related Work

| Paper | Methodology | Results | Drawbacks |
|---|---|---|---|
| [5] | Scoping review conducted using PRISMA guidelines with systematic searches in PubMed, Web of Science, and Scopus. | Identified diverse healthcare chatbot applications, including mental health support, medical information provision, appointment management, health education, chronic condition management, and COVID-19-related services. | Limited evidence on cost-effectiveness, alongside challenges related to implementation, system interoperability, ethical considerations, and gaps in robust outcome evaluation. |
| [18] | C-PATH is built on a fine-tuned LLaMA3-based large language model using a multi-stage training pipeline that includes medical knowledge acquisition, conversational dialogue training, and clinical summarization. Structured medical data are transformed into patient-friendly multi-turn conversations, with GPT-3.5 Turbo used to rewrite the DDXPlus-derived conversations for improved consistency. The system incorporates multi-turn dialogue management to maintain coherence during extended interactions. | The system demonstrates high performance in conversational quality and department-level triage. GPTScore evaluation shows strong understandability and informativeness, with high accuracy in medical department recommendations. Classification results further confirm reliable department recommendation performance on rewritten dialogue datasets. | Despite strong performance, the system remains susceptible to common LLM limitations, including hallucinated or incorrect medical responses in patient-facing scenarios. |
| [19] | Built an LLM-RAG system using 35 preoperative guidelines, embedded the documents, retrieved relevant chunks per query, and evaluated multiple LLMs against human answers using 14 anonymized clinical scenarios. | GPT-4-RAG achieved 91.4% accuracy, compared to 80.1% for GPT-4 alone and 86.3% for human responses, with response times of 15–20 seconds versus about 10 minutes for humans and a low hallucination rate. | The study used simulated scenarios and a single guideline set, limiting generalizability. Text conversion may lose information from figures, and system performance depends on retrieval quality despite low hallucination rates. |
| [16] | Preliminary cross-sectional study evaluating ChatGPT for emergency triage prediction using 50 case scenarios based on the Emergency Severity Index (ESI) and comparing ChatGPT predictions with expert consensus. | Moderate overall performance with fair agreement (Cohen's Kappa = 0.341). Better performance for high-acuity cases (ESI-1 and ESI-2) with sensitivity 76.2%, specificity 93.1%, and AUC 0.846. | Scenario-based evaluation, small sample size, poor performance for lower acuity cases, and limited generalizability to real emergency department settings. |
| [17] | Retrospective study using ChatGPT 4.0 to assist ER triage for 56 patients with metastatic prostate cancer using structured and unstructured EHR data. Performance compared to ER physicians. | High sensitivity for admission (95.7%), high diagnosis agreement (87.5%), more comprehensive treatment recommendations, but low specificity for discharge. | Low discharge specificity, ESI not predictive, potential hallucinations, and retrospective design limiting generalizability. |
| [9] | Randomized clinical trial comparing AI-assisted outpatient service "Smart-doctor" with conventional service in a pediatric hospital. | Significantly reduced queuing time, consultation time, and total visit time, with higher patient satisfaction scores (p ¡ 0.01). | Single-hospital setting and lack of long-term clinical outcome evaluation. |
| [14] | Rule-based chatbot designed to support emergency triage by collecting patient information and assisting nurses during initial assessment. | Improved organization of triage process and support for nurses, especially less experienced staff. | Descriptive evaluation without numerical results, single-hospital testing, and need for frequent rule updates. |
| [15] | Compared ChatGPT and multiple LLMs with untrained doctors and expert raters using 124 emergency case vignettes based on the Manchester Triage System. | GPT-4-based ChatGPT performed close to untrained doctors but below expert raters, with a tendency toward overtriage. | Scenario-based evaluation only, no improvement as decision support, and no real-world clinical validation. |
| [11] | Enhanced RAG framework based on Qwen-14B-Chat, integrating improved preprocessing, prompt design, and fusion of sparse and dense retrievers. | Outperformed baseline models in medical knowledge understanding and semantic reasoning tasks, achieving a 3.86% overall improvement. | Evaluation limited to competition datasets and automatic metrics without real clinical deployment. |
| [21] | Proposed MedRAG integrating Retrieval-Augmented Generation with knowledge graph–elicited reasoning, evaluated on CPDD and DDXPlus datasets. | Outperformed all RAG baselines across difficulty levels, with significant gains from knowledge graph reasoning. | Evaluation limited to benchmark datasets without real-world clinical deployment. |
| [13] | Evaluated multiple open-source LLMs using the OpenMedLM platform with sequential prompt engineering strategies across four medical QA benchmarks. | Yi-34B achieved strong benchmark performance, outperforming prior open-source state-of-the-art models. | Limited to academic benchmarks without assessment of real clinical interaction or deployment risks. |

| Key Prompt Design Principles for the Medical Chatbot | |
|---|---|
| أنت مساعد طبي ذكي يعمل في مستشفى أردني. | |
| دورك الأساسي: | القواعد الصارمة: |
| التحدث بالعربية الفصحى مع فهم اللهجة الأردنية. | لا تقدم تشخيصاً نهائياً أبداً – فقط توجيه للقسم المناسب. |
| طرح أسئلة طبية متسلسلة ومنطقية لفهم حالة المريض. | اطرح سؤالاً واحداً في كل مرة. |
| جمع معلومات كافية عن الأعراض. | كن متعاطفاً ومطمئناً للمريض. |
| تحديد القسم الطبي المناسب بناءً على الأعراض. | إذا شككت بحالة طارئة، أخبر المريض بالذهاب للمستشفى فوراً. |
| اكتشاف الحالات الطارئة فوراً. | اجمع معلومات عن: نوع الألم، مدته، شدته، أعراض مصاحبة. |

Fig. 1: Example of the system prompt defining the role and behavioral rules of the medical chatbot

The prompt defines the assistant's role, core responsibilities, and strict behavioral constraints, ensuring safe and structured medical interaction while preventing diagnosis or unsupported medical advice. The same prompt structure will be reused when testing other language models. This allows a fair comparison between models and ensures that any performance difference is related to the model itself rather than changes in system logic.

### *B. Retrieval Augmented Generation and Medical Knowledge Base*

To improve the accuracy and reliability of responses, the system integrates a retrieval-augmented generation (RAG) component. The medical knowledge used by the RAG module is stored in a structured JSON file that contains information about medical departments, common symptoms, and guidance rules. This dataset was created manually by a general practitioner with sufficient clinical experience, ensuring that the content reflects real medical practice.

The proposed RAG process is implemented as an integral part of the system architecture. First, the medical knowledge is preprocessed and indexed. When the patient describes symptoms, the system retrieves the most relevant entries from the knowledge base. These retrieved texts are then injected into the prompt provided to the language model, allowing it to generate responses grounded in verified medical content rather than relying only on its internal knowledge.

Although the current implementation uses a single curated JSON file, the system is designed to support future expansion. Additional data sources such as clinical guidelines, hospital protocols, or publicly available medical references can be integrated later without changing the core architecture. This makes the system scalable and adaptable to different clinical settings.

### *C. Multi-Agent System Architecture*

The chatbot is implemented using a multi-agent architecture to separate responsibilities and improve system clarity. Four main agents are defined, each responsible for a specific task within the interaction flow.

1) **Registration Agent** handles the initial interaction with the patient. It is responsible for collecting basic patient information required to start the process, such as identification details and consent to proceed.
2) **Conversation Agent** manages the medical dialogue. It interacts directly with the patient, asks follow-up questions, collects symptoms gradually, and ensures that the conversation follows the predefined medical rules. This agent focuses on maintaining a natural and understandable interaction.
3) **RAG Agent** responsible for retrieving relevant medical knowledge based on the collected symptoms. It communicates with the knowledge base, selects the most relevant medical content, and provides it to the language model to support informed decision-making.
4) **Booking Agent** handles appointment scheduling. After the appropriate department is identified, this agent retrieves available appointment slots from the database and presents them to the patient. The patient is asked to choose the most suitable date and time, and only after confirmation does the agent finalize the booking. This design ensures that appointments are booked based on patient preference rather than automatically.

This separation of roles makes the system easier to maintain, test, and extend, as each agent can be improved independently. Figure 2 illustrates the workflow of the proposed multi-agent system.

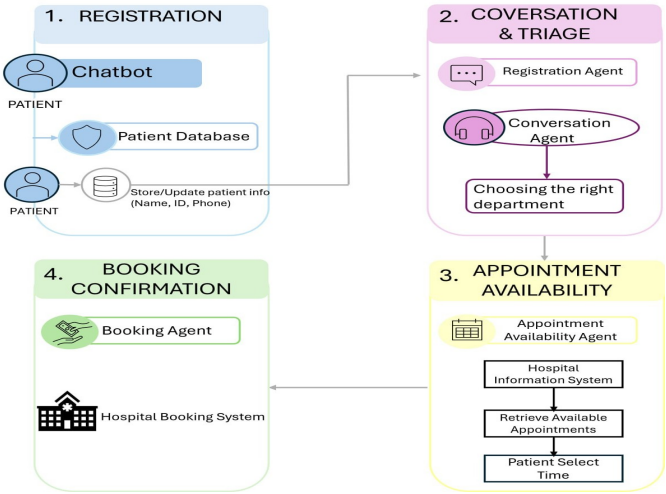


Fig. 2: End-to-end workflow of the proposed multi-agent smart medical chatbot system.

### *D. Database Design*

The system uses a relational database to store structured information related to doctors, departments, and appointment

schedules. The database includes predefined doctors, their associated departments, available dates, and time slots. Appointment status is updated dynamically based on patient bookings.

An example of the appointments table structure is shown in Table II. This table stores the appointment date, time, availability status, and links to both doctor and patient records. The database design supports future extensions such as appointment history tracking and statistical analysis of patient flow.

TABLE II: Sample appointment table structure

| ID | Doctor ID | Department ID | Appointment Date | Time | Status |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 2025-11-27 | 09:00 | BOOKED |
| 423 | 2 | 1 | 2025-12-07 | 11:00 | AVAILABLE |
| 432 | 2 | 1 | 2025-12-08 | 16:00 | AVAILABLE |
| 426 | 2 | 1 | 2025-12-07 | 16:00 | AVAILABLE |
| 960 | 6 | 2 | 2025-12-23 | 14:00 | AVAILABLE |

### *E. AI-Based Features*

Several AI-driven features are planned to enhance the user experience and system effectiveness:

- **Automatic appointment reminders:** The system sends an automated reminder to the patient one day before the scheduled visit, helping reduce missed appointments and improve hospital workflow.
- **Selectable operating modes:** The system allows patients to choose between two operating modes:
  - **Medical consultation and guidance mode:** This mode focuses on providing general medical guidance using prompt engineering, without performing appointment booking.
  - **Intake and booking mode:** This mode focuses on symptom collection, department identification, and appointment booking, which aligns with the primary objective of the system.

## V. Experiments

To assess the effectiveness and safety of the proposed system, a set of evaluation criteria is defined focusing on correct patient routing, emergency handling, and workflow completion.

### *A. Experimental Setup And Baselines*

The evaluation is conducted using scenario-based testing to validate the proposed smart medical chatbot. Since this work represents an initial implementation phase, the experiments focus on functional validation and system behavior rather than large-scale quantitative evaluation.

A set of medical scenarios written in Arabic was developed in collaboration with a general practitioner. Each scenario represents a realistic patient complaint and includes an expected medical department and, when applicable, emergency warning signs. The system is tested through multi-turn conversations, starting from patient registration, followed by symptom collection, department recommendation, and either appointment booking or emergency interruption.

To assess the contribution of the proposed system architecture, the evaluation includes comparison with simplified internal baselines. The first baseline is an LLM-only version, where the chatbot relies solely on the language model and prompt rules without retrieval-augmented generation. The second baseline is a single-agent version that uses retrieval-based medical knowledge but does not separate tasks into multiple agents. These baselines enable evaluating the impact of retrieval grounding and multi-agent workflow design within the same experimental setup.

### *B. Evaluation Metrics*

To prepare for full system evaluation in later stages, a set of approximately **200 medical scenarios** has been developed in collaboration with a general practitioner. These scenarios cover different symptom patterns and medical cases and will be used for comprehensive system testing after completing the full implementation.

The system is evaluated using the following metrics, which focus on the correctness of medical routing, safety handling, and overall workflow reliability:

- **Department Accuracy:** measures how often the system recommends the correct medical department based on the patient's symptoms, compared to the expected outcome of each scenario.
- **Emergency Detection Recall:** measures the system's ability to correctly identify urgent cases that require immediate medical attention.
- **Emergency Detection Precision:** measures how many of the cases flagged as emergencies by the system are truly emergency cases.
- **Workflow Success Rate:** measures the percentage of scenarios in which the system successfully completes the full process, from patient intake to appointment booking or emergency interruption.

## VI. Results

At current stage, the focus was on implementing the core system components and verifying that the proposed design is functional and feasible in practice, rather than conducting a full quantitative evaluation. The primary outcome of this phase is the successful implementation of an end-to-end smart medical chatbot prototype. The system integrates patient registration, medical conversation, department recommendation, and appointment booking into a single workflow. The implemented prototype demonstrates that the proposed architecture can support a complete interaction starting from patient data entry and ending with confirmed appointment scheduling.

The conversational interface allows the system to interact with patients in Arabic, collect symptoms gradually, and ask follow-up questions in a structured manner. Based on the provided information, the system is able to suggest an appropriate medical department and proceed to appointment booking. The interaction flow reflects the different stages of system operation, including patient registration, medical dialogue, symptom clarification, and final appointment confirmation.

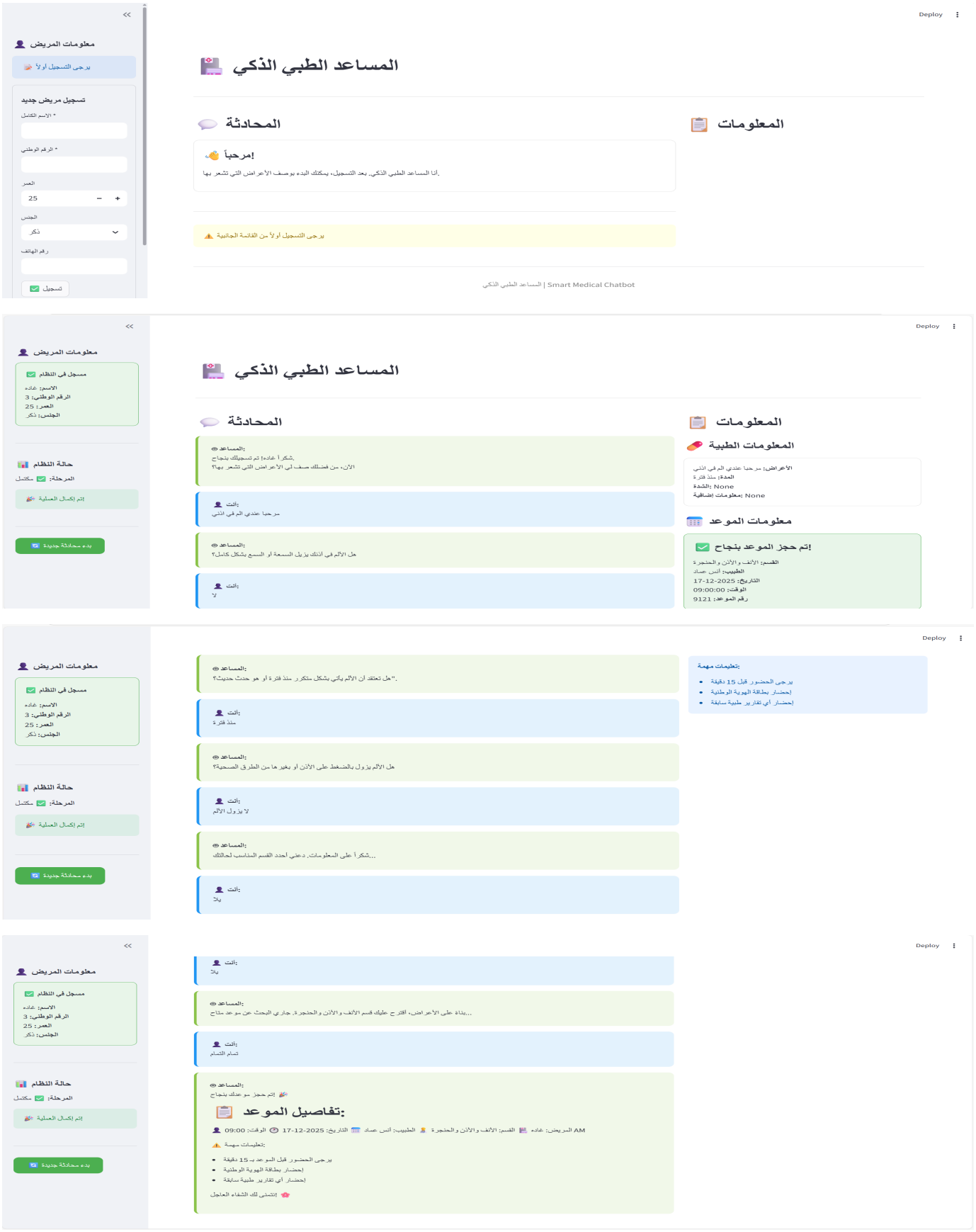


Fig. 3: User interface of the proposed system showing the complete interaction flow, including patient registration, multi-turn Arabic medical conversation, appointment booking confirmation, and final system output summary.

Figure 3 illustrates the complete user interaction flow of the proposed smart medical chatbot. The first interface shows the initial patient registration stage, where basic patient information such as name, age, gender, and identification details are collected before starting the medical conversation. This step ensures that all interactions are linked to a registered patient profile.

After successful registration, the system initiates a structured Arabic medical dialogue. As shown in the subsequent interfaces, the chatbot asks follow-up questions to gradually collect symptoms and clarify the patient's condition. The conversation is conducted in a multi-turn manner, allowing the system to refine its understanding based on user responses while maintaining a clear and organized dialogue flow.

Based on the collected information, the system identifies the most appropriate medical department and transitions to the appointment booking stage. The booking interface illustrates the intended workflow of the appointment scheduling process, which is designed to be fully activated in later stages of the project.

Finally, the last interface presents a summary of the interaction, including patient details and confirmed appointment information. This summary ensures transparency for the patient and verifies that the system has correctly completed the full workflow. Overall, these interfaces demonstrate that the proposed system successfully integrates patient registration, medical conversation, department recommendation, and appointment booking into a single coherent process.

In addition to the conversational flow, a database structure was implemented to manage doctors, departments, and appointment schedules.Overall, the results obtained in this phase confirm that the core idea of the project is technically achievable. The successful implementation of the main components provides a solid foundation for future work, which will focus on full-scale testing, performance evaluation, and comparison between different language models.

## VII. Conclusion

This project presented the design and initial implementation of a smart medical chatbot aimed at supporting the early stages of healthcare service delivery. The system focuses on automated patient interaction, symptom-based guidance, and appointment booking, with the goal of reducing pressure on hospitals and improving patient flow during the first point of contact. The developed chatbot combines conversational interaction in Arabic with knowledge-supported decision making and structured database operations within a single workflow. Through guided medical dialogue, the system is able to collect patient symptoms step by step and suggest the most suitable medical department without providing any final diagnosis. This approach helps organize patient requests while keeping the interaction simple and understandable.

An important aspect of this work is the integration of retrieval-based medical knowledge and a multi-agent system design. By separating system responsibilities into dedicated agents and supporting the language model with curated medical data, the system achieves better control, clearer task management, and more reliable behavior. This design also improves maintainability and allows future extensions without major changes. Overall, the successful implementation of the prototype confirms that the proposed approach is technically feasible and suitable for further development. While the current phase focuses on functional validation rather than full clinical evaluation, the results show strong potential for using intelligent chatbots as supportive tools in healthcare environments. Future work will focus on system evaluation, expanding medical data, and improving integration with real hospital systems.

## References


[1] OECD, *Waiting Times for Health Services: Next in Line*. Paris: OECD Publishing, 2020. [Online]. Available: https://doi.org/10.1787/242e3c8c-en

[2] World Health Organization, "Who outlines considerations for regulation of artificial intelligence for health," Oct. 2023. [Online]. Available: https://www.who.int/news/item/19-10-2023-who-outlines-considerations-for-regulation-of-artificial-intelligence-for-health

[3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2021. [Online]. Available: https://arxiv.org/abs/2005.11401

[4] M. Wooldridge, *An Introduction to Multi-Agent Systems*. John Wiley & Sons, 2002. [Online]. Available: https://uranos.ch/research/references/Wooldridge_2001/TLTK.pdf

[5] M. Barreda, D. Cantarero-Prieto, D. Coca, A. Delgado, P. Lanza-León, J. Lera, R. Montalbán, and F. Pérez, "Transforming healthcare with chatbots: Uses and applications–a scoping review," *Digital Health*, vol. 11, p. 20552076251319174, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11915287/

[6] M. Laymouna, Y. Ma, D. Lessard, T. Schuster, K. Engler, and B. Lebouché, "Roles, users, benefits, and limitations of chatbots in health care: Rapid review," *J Med Internet Res*, vol. 26, p. e56930, Jul. 2024, https://doi.org/10.2196/56930. [Online]. Available: https://www.jmir.org/2024/1/e56930

[7] M. Hindelang, S. Sitaru, and A. Zink, "Transforming health care through chatbots for medical history-taking and future directions: Comprehensive systematic review," *JMIR Medical Informatics*, vol. 12, p. e56628, Aug. 2024.

[8] J. Nelson, A. Wills, and J. Owen, "Streamlining patient triage and appointment scheduling with ai assistants," May 2025.

[9] X. Li, D. Tian, W. Li, Y. Hu, B. Dong, H. Wang, J. Yuan, B. Li, H. Mei, S. Tong, L. Zhao, and S. Liu, "Using artificial intelligence to reduce queuing time and improve satisfaction in pediatric outpatient service: A randomized clinical trial," *Front. Pediatr.*, vol. 10, p. 929834, 2022.

[10] M. Yuan, P. Bao, J. Yuan, Y. Shen, Z. Chen, Y. Xie, J. Zhao, Y. Chen, L. Zhang, L. Shen, and B. Dong, "Large language models illuminate a progressive pathway to artificial healthcare assistant: A review," 2023. [Online]. Available: https://arxiv.org/abs/2311.01918

[11] Y. Wang, Y. Wan, X. Lei, Q. Chen, and H. Hu, "A retrieval augmented generation based optimization approach for medical knowledge understanding and reasoning in large language models," *Array*, vol. 28, p. 100504, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590005625001316

[12] U. Iqbal, A. Tanweer, A. R. Rahmanti, D. Greenfield, L. T.-J. Lee, and Y.-C. J. Li, "Impact of large language model (chatgpt) in healthcare: an umbrella review and evidence synthesis," *Journal of Biomedical Science*, vol. 32, no. 1, p. 45, 2025, epub 2025 May 7. [Online]. Available: https://jbiomedsci.biomedcentral.com/

[13] J. Maharjan, A. Garikipati, N. P. Singh, L. Cyrus, M. Sharma, M. Ciobanu, G. Barnes, R. Thapa, Q. Mao, and R. Das, "Openmedlm: Prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models," 2024. [Online]. Available: https://arxiv.org/abs/2402.19371

[14] M. H. Alwaked, F. A. Alammar, S. M. Algfari, A. S. Alghamdi, A. M. Almuhaylib, M. A. Alzaher, A. T. Alharbi, A. A. Alsaif, G. M. Alamri, H. S. Alsaif, A. S. Al-Otaibi, Z. H. Alanazi, M. H. Alwaked, F. A. Al-Sharari, F. A. H. Alsarimi, and A. Z. Alumtairi, "The use of chatbots for triage and emergency nursing support: Review article," *International Journal of Health Sciences*, vol. 5, no. S1, pp. 1207–1218, 2021.

[15] L. Masanneck, L. Schmidt, A. Seifert, T. Kölsche, N. Huntemann, R. Jansen, M. Mehsin, M. Bernhard, S. G. Meuth, L. Böhm, and M. Pawlitzki, "Triage performance across large language models, chatgpt, and untrained doctors in emergency medicine: Comparative study," *Journal of Medical Internet Research*, vol. 26, p. e53297, 2024, epub 2024 Jun 14.

[16] Sarbay, G. B. Berikol, and U. Özturan, "Performance of emergency triage prediction of an open access natural language processing based chatbot application (chatgpt): A preliminary, scenario-based cross-sectional study," *Turkish Journal of Emergency Medicine*, vol. 23, no. 3, pp. 156–161, Jul. 2023.

[17] G. Gebrael, K. K. Sahu, B. Chigarira, N. Tripathi, V. Mathew Thomas, N. Sayegh, B. L. Maughan, N. Agarwal, U. Swami, and H. Li, "Enhancing triage efficiency and accuracy in emergency rooms for patients with metastatic prostate cancer: A retrospective analysis of artificial intelligence-assisted triage using chatgpt 4.0," *Cancers*, vol. 15, no. 14, p. 3717, 2023. [Online]. Available: https://www.mdpi.com/2072-6694/15/14/3717

[18] Q. Shi, Q. Han, and C. Soares, "C-path: Conversational patient assistance and triage in healthcare system," 2025. [Online]. Available: https://arxiv.org/abs/2506.06737

[19] Y. Ke, L. Jin, K. Elangovan, H. R. Abdullah, N. Liu, A. T. H. Sia, C. R. Soh, J. Y. M. Tung, J. C. L. Ong, and D. S. W. Ting, "Development and testing of retrieval augmented generation in large language models – a case study report," 2024. [Online]. Available: https://arxiv.org/abs/2402.01733

[20] R. Patil, M. Abbidi, and S. Fannon, "Ragmed: A rag-based medical ai assistant for improving healthcare delivery," *AI*, vol. 6, no. 10, p. 240, 2025. [Online]. Available: https://www.mdpi.com/2673-2688/6/10/240

[21] X. Zhao, S. Liu, S.-Y. Yang, and et al., "Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot," 2025. [Online]. Available: https://arxiv.org/abs/2502.04413

[22] A. A. Borkowski and A. Ben-Ari, "Multiagent ai systems in health care: Envisioning next-generation intelligence," *Federal Practitioner*, vol. 42, no. 5, pp. 188–194, May 2025, epub 2025 May 14.

[23] D. Isern and A. Moreno, "A systematic literature review of agents applied in healthcare," *Journal of Medical Systems*, vol. 40, no. 2, p. 43, 2016, epub 2015 Nov 21. [Online]. Available: https://link.springer.com/journal/10916