# A Multi-Agent AI Framework for Unified Healthcare Assistance Using Retrieval Augmented Generation

Dr. Asha T, Dr. Maya B S, Ananya Jha, Akshit Mehta, Khushi Gupta, Manpreet Kaur
*Department of Information Science & Engineering*
*Bangalore Institute of Technology Bengaluru, India*

*Abstract*—**This paper proposes a centralized AI system for healthcare assistance that uses a manager agent to orchestrate three specialized agents: a Diagnostics Agent, a Search Agent, and a Hospital Operations Agent. The Diagnostics Agent employs a fine-tuned clinical BERT model (BioClinicalBERT) combined with Retrieval-Augmented Generation (RAG) over a ChromaDB vector store, enriched by results from a Search Agent. The Search Agent integrates Google Custom Search on trusted medical sites and an internal document RAG system. The Hospital Operations Agent uses Python-based analytics and visualization tools to present patient and resource data, aiding junior doctors with on-demand preliminary diagnostics. This modular, RAG- driven approach is compared against monolithic Large Language Models (LLMs) and fine-tuned models, demonstrating improved efficiency, accuracy, and cost-effectiveness. The orchestration of specialized agents not only parallelizes complex tasks but also mitigates hallucinations through domain-focused retrieval. Comparative tables and scenarios highlight these benefits, un- derscoring how modular RAG architectures can adapt rapidly to new medical information without costly retraining.**

*Index Terms*—**Multi-agent systems, RAG, LLM, BERT, Clin- ical decision support, Healthcare AI, Agent Orchestration, Evidence-Based diagnosis, Hospital Operations Analytics, Knowl- edge Retrieval.**

## I. INTRODUCTION

Providing accurate, timely, and relevant clinical decision support has become an imperative for contemporary healthcare systems that are working under mounting patient volume and resource constraints. Artificial intelligence (AI) has gained widespread use for assisting healthcare professionals with diagnosis, triage, patient monitoring, and hospital operations. Conventional AI-based healthcare systems mainly involve monolithic machine learning/deep learning models that were designed for large static datasets. Even if these model-oriented, mainly centralized systems have proven highly effective for single-minded clinical applications, they have limitations with regard to their flexibility, applicability, and tractability for use within dynamic clinical settings [1], [2].

Even though the natural language understanding (NLU) abilities of GPT or transformer neural networks are quite excellent, they are incapable of updating the latest medical standards or the latest medical or hospital-specific knowledge without undergoing expensive retraining or fine-tuning. This was established by the latest MLHOps studies on the chal- lenges of updating models in the medical field [3].

To address these challenges, lately, a significant number of studies on the topic have been investigating the use of Retrieval-Augmented Generation (RAG) as a potential capa- bility in anchoring the generated AI answers within current medical knowledge. RAG architectures allow models to search and access appropriate medical documents contained in care- fully curated knowledge repositories during inference tasks. This enables greater accuracy on the task and suppresses the issue of hallucinations in clinical question-answering tasks. Various studies conducted on evidence-based medicine and multimodal electronic healthcare record (EHR) analysis prove that RAG-based models perform better than both the base and the fine-tuned models on tasks that require access to medical knowledge [4], [5].

Previous studies have validated that multi-agent systems can be efficiently employed in patient monitoring, triaging, prior authorizations, and cooperating care systems. These systems break down a complex healthcare process into area-specific agents, allowing enhanced scalability, modularity, and task- oriented snippets [6, 7]. After a recent exploration, there is a required gap in comprehensively orchestrated multi- agent frameworks intertwining diagnostic inferences, off-body knowledge inquiries, and hospital administration in a single, cohesive framework.

In response to these challenges and limitations, this pa- per introduces the Multi-Agent AI Framework for Unified Healthcare Assistance Using RAG. Based on this framework, there is a manager agent that controls three task agents: the Diagnostics Agent, the Search Agent, and the Hospital Operations Agent. The Diagnostics Agent uses a medical lan- guage model that enhances the generation process using RAG over the curated vector database for medical documents. This method helps the Diagnostics Agent perform evidence-based clinical reasoning. The Search Agent supplements this by generating evidence-based information from authentic sources on the web. The Hospital Operations Agent provides analytics support using visualization-driven data analysis. This helps healthcare professionals quickly understand the situation and make informed decisions. Through the integration of RAG and multi-agent orchestration, the proposed system overcomes the weaknesses found in existing studies among monolithic LLMs, fine-tuned transformer models, and decentralized models using the agentic approach. The experimental results show that the proposed system outperforms the existing approach in accuracy, computation, and modularity for the response to be given. The overall system architecture is depicted in Fig. 1.

## II. LITERATURE SURVEY

### *A. AI-Based Clinical Decision Support Systems*

The adoption of AI for clinical decision support has been extensively explored in recent literature. Early systems pri- marily relied on centralized machine learning models trained on EHRs to assist in diagnosis, risk prediction, and early warning systems. Khattak et al. [1] surveyed the opera- tionalization of machine learning in healthcare, emphasizing that while centralized models can improve clinical outcomes, their deployment introduces challenges related to model drift, governance, and integration into clinical workflows. Simi- larly, virtual healthcare assistant systems based on monolithic deep learning architectures have demonstrated feasibility for symptom checking and medical advice but remain constrained by limited adaptability and scalability in real-world hospital environments [2].

Recent advancements in LLMs have further expanded the scope of AI-assisted healthcare. Transformer-based architec- tures have achieved strong performance in medical question answering, patient communication, and clinical documenta- tion. However, studies indicate that these monolithic LLMs suffer from fixed knowledge cutoffs, high inference costs, and susceptibility to hallucinations, particularly in safety-critical medical contexts [3]. These limitations hinder their long-term reliability and clinical adoption without substantial operational overhead.

### *B. RAG in Healthcare*

To address the limitations of static model knowledge, RAG has emerged as a promising paradigm for healthcare AI systems. RAG enhances language models by retrieving relevant external documents during inference, enabling re- sponses grounded in up-to-date medical evidence. Masanneck et al. [3] demonstrated that RAG-enabled LLMs significantly outperform base models in guideline adherence for evidence- based neurology, reducing both hallucinations and clinically unsafe recommendations.

In parallel, Zhu et al. [4] proposed the EMERGE frame- work, which integrates RAG with multimodal EHR data and medical knowledge graphs. Their results show that RAG architectures outperform conventional deep learning baselines on mortality prediction and readmission tasks. These findings suggest that RAG offers a more flexible and cost-effective alternative to continual fine-tuning, particularly for healthcare applications requiring frequent knowledge updates.

### *C. Multi-Agent and Agentic AI Systems in Healthcare*

Agent-based and multi-agent AI systems have gained in- creasing attention as a means of decomposing complex health- care workflows into specialized components. Shaik et al. [5] proposed a multi-agent deep reinforcement learning frame- work for patient monitoring, where individual agents monitored specific physiological signals, resulting in improved emergency detection accuracy compared to single-agent ap- proaches.

Beyond clinical monitoring, Pandey et al. [6] explored the use of specialized LLM agents for automating medical prior authorization workflows. Their system demonstrated that task-specific agents can achieve high accuracy while improv- ing transparency and explainability. Additionally, Power [7] highlighted the potential of multi-agent systems in healthcare interoperability and collaborative care, noting their ability to reduce redundancy and improve efficiency across distributed healthcare infrastructures.

Despite these advances, existing agentic healthcare systems are often decentralized or loosely coordinated, focusing on isolated tasks such as monitoring, authorization, or conversational assistance. Few works address centralized orchestration mechanisms capable of coordinating multiple agents across di- agnostic reasoning, external knowledge retrieval, and hospital operations within a unified framework.

*D. Summary and Research Gap*

The reviewed literature demonstrates significant progress in both RAG-based healthcare AI and multi-agent system design. RAG architectures improve factual accuracy and adaptability, while agentic systems enhance scalability and task specialization. However, a critical gap remains in the integration of these two paradigms. Existing solutions largely lack centralized orchestration, resulting in fragmented decision-making pipelines and limited cross-agent coordination.

Furthermore, many healthcare AI systems either rely on monolithic LLMs with prohibitive computational costs or fine- tuned models that struggle to incorporate evolving medical knowledge. These limitations underscore the need for a unified, manager-orchestrated multi-agent architecture that com- bines RAG with explicit task delegation across diagnostics, knowledge search, and hospital operations. Addressing this gap, the proposed framework aims to deliver scalable, cost-effective, and clinically aligned healthcare assistance through centralized agent orchestration.

## III. PROPOSED SYSTEM ARCHITECTURE

The proposed system adopts a centralized multi-agent architecture designed to support unified healthcare assistance. The architecture is organized into three logical layers: a User Interaction Layer, an AI-Orchestrated Logic Layer, and a Knowledge and Data Layer. At the core of the system lies a Manager Agent that governs task orchestration, inter-agent communication, and response synthesis. This design enables scalability, task specialization, and controlled decision-making while avoiding the limitations of monolithic AI systems.

The interaction flow begins with clinician or hospital staff queries submitted through the user interface and culminates in evidence-grounded diagnostic, informational, or operational outputs. The flow is then driven by explicit orchestration rules defined within the Manager Agent, ensuring consistency, traceability, and reliability across diagnostic, informational, and operational tasks. Fig. 1 illustrates the high-level architecture of the proposed multi-agent healthcare assistance system, showing the three logical layers and the inter-agent communication pathways.

*A. System Overview*

The architecture has been designed to overcome the draw- backs of monolithic AI architectures by decomposing the healthcare domain process into specialized agents, along with maintaining centralized control. As shown in Fig. 1, interaction within the system takes place through a secured web-based interface for clinicians, junior doctors, and hospital administrators.

Every user activity is initially channeled through the Man- ager Agent, which is responsible for intent classification and deciding on the course of action. Based on the intent, the Manager Agent calls one or more of the specialized agents, namely the Diagnostics Agent, the Search Agent, and the Hospital Operations Agent. Parallel processing of tasks is accomplished through this central coordinator, as depicted in the orchestration layer of Fig. 1.

*B. AI-Powered Diagnostic Reasoning*

The core clinical reasoning capability is implemented within the Diagnostics Agent. This agent integrates a domain- specific medical BERT model with a RAG pipeline to deliver evidence-based diagnostic assistance, as illustrated in the AI-Orchestrated Logic Layer of Fig. 1.

- Clinical Query Encoding and Retrieval: The incoming clinical cases are converted into dense vector representations and compared against curated vector databases containing medical literature, clinical guidelines, and institutional documents. The appropriate vectors, based on semantic similarity, are chosen and fed as context to the diagnostic model.

- Medical Language Model Inference: A medical BERT model is used for processing the input and developing diagnostic information, symptom links, and preliminary diagnostic recommendations. The retrieved medical evidence is used for inference and is helpful in eliminating model hallucinations. In an effort to improve robustness, the Diagnostics Agent may invoke the Search Agent for supplementary validation if the queries pertain to new disease conditions, latest treatment procedures, or equivocal symptoms.

$$D = \begin{cases} D_{raw}, & \text{if } C(D_{raw}) \geq \vartheta \\ \text{RAG}(Q), & \text{otherwise} \end{cases} \tag{1}$$

### C. *Intelligent Medical Knowledge Retrieval*

The controlled external knowledge acquisition and validation are entrusted to the Search Agent. It operates as an independent retrieval component that accesses only trusted healthcare information sources, including government health portals and accredited medical institutions. The position of the Search Agent within the overall system is shown in Fig. 1.

Upon receiving a request from the Manager Agent or Di- agnostics Agent, the Search Agent performs targeted searches and ranks retrieved documents based on relevance and credi- bility.

$$S(q, d) = \frac{q \cdot d}{||q||||d||} \tag{2}$$

Instead of returning raw content, the agent extracts salient information and source references and feeds these into the subsequent stages of the diagnostic reasoning pipeline.

This separation of responsibilities ensures that external information retrieval remains auditable, secure, and insulated from unverified sources, addressing a common limitation of monolithic LLMs.

**Algorithm 1** Diagnostics Agent: RAG-Based Clinical Reasoning

```
Input: Clinical Query Q
Output: Diagnostic Insight D
procedure DIAGNOSTICSAGENT(Q)
    q_vec ← EMBED(Q)
    D_docs ← CHROMADB.RETRIEVE(q_vec)
    context ← Q ∪ D_docs
    D_raw ← MEDICALBERT(context)
if CONFIDENCE(D_raw) < ϑ then
        E_ext ← SEARCHAGENT(Q)
        context ← context ∪E_ext
        D ← MEDICALBERT(context)
else
        D ← D_raw
end if
    return D
end procedure
```

**Algorithm 2** Search Agent: Trusted Medical Knowledge Retrieval

```
Input: Query Q
Output: Verified Evidence E
procedure SEARCHAGENT(Q)
    Q_clean ← SANITIZE(Q)
    results ← GOOGLECSE(Q_clean)
    trusted ← FILTER(results, GovMedicalDomains)
    ranked ← RANK(trusted)
    E ← SUMMARIZE(top(ranked))
    return E
end procedure
```

### D. *Hospital Operations and Resource Analytics*

The Hospital Operations Agent supports administrative and operational decision-making within healthcare facilities. This agent processes structured hospital data related to patient inflow, bed availability, resource utilization, and staffing levels, as represented in the Knowledge and Data Layer of Fig. 1.

Using analytical and visualization tools, the agent generates real-time summaries, trends, and graphical dashboards that enable hospital administrators to assess system load and resource distribution. Additionally, the agent provides an interface through which junior doctors can initiate preliminary diagnostic workflows, which are subsequently routed through the Manager Agent for coordinated processing.

This functionality enables data-driven operational planning while ensuring that diagnostic actions remain aligned with system-wide context and policies.

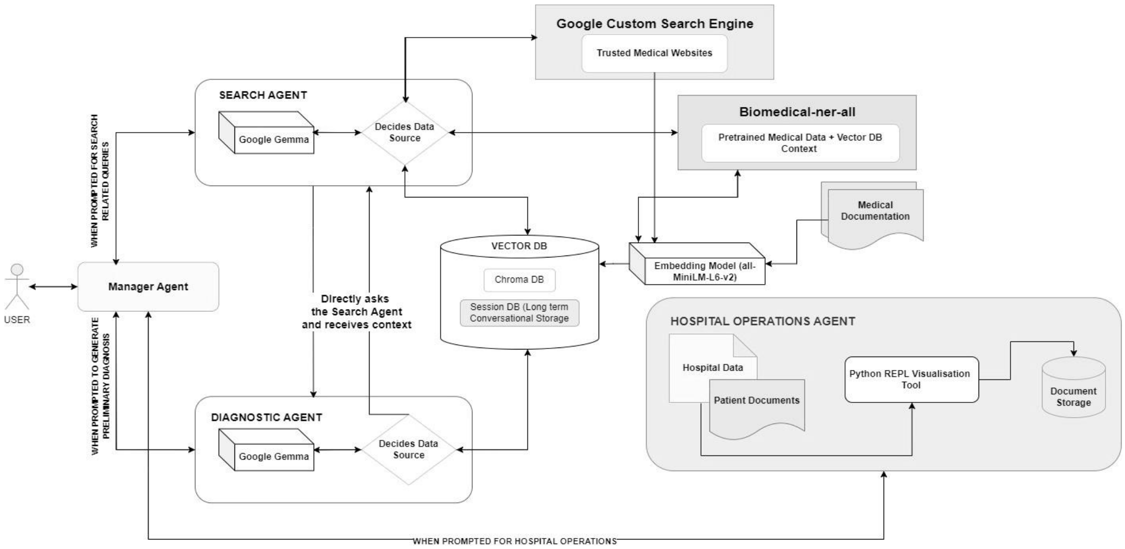


Fig. 1. Proposed multi-agent healthcare system architecture.

**Algorithm 3** Hospital Operations Agent: Analytics and Visualization

```
Input: Operational Query Q, Role R
Output: Analytical Report A
procedure OPERATIONSAGENT(Q, R)
if NOT AUTHORIZED(R) then
    return ACCESS_DENIED
end if
    data ← QUERYHOSPITALDB(Q)
    stats ← ANALYZE(data)
    plots ← VISUALIZE(stats)
    A ← COMPOSE(stats, plots)
    return A
end procedure
```

## E. Manager Agent and Centralized Orchestration

The Manager Agent serves as the central control and coordination unit of the proposed system, as shown at the core of Fig. 1. It maintains execution state, manages inter- agent communication, resolves conflicts, and synthesizes final responses delivered to the user.

$$R_{final} = F(R_d, R_s, R_o) \quad (3)$$

**Algorithm 4** Manager Agent: Task Orchestration and Response Synthesis

```
Require: User Query Q, User Role R
Ensure: Unified System Response R_final
intent ← CLASSIFYINTENT(Q)
taskList ← DECOMPOSE(intent)
responses ← ∅
for each task ∈ taskList do
    if task == Diagnostic then
        resp ← DIAGNOSTICSAGENT(Q)
    else if task == Search then
        resp ← SEARCHAGENT(Q)
    else if task == Operations then
        resp ← OPERATIONSAGENT(Q, R)
    end if
    responses ← responses ∪ resp
end for
responses ← RESOLVECONFLICTS(responses)
R_final ← SYNTHESIZE(responses)
return R_final
```

Unlike decentralized multi-agent systems where agents operate independently, the Manager Agent enforces a centralized orchestration model that guarantees contextual coherence and traceability across all tasks. By aggregating outputs from diagnostic, retrieval, and operational agents, the Manager Agent ensures that responses are consistent, evidence-backed, and aligned with clinical and operational constraints.

Overall, the proposed system architecture integrates RAG- based diagnostic reasoning, controlled knowledge retrieval, and hospital operations analytics within a unified, manager- orchestrated multi-agent framework. This design provides a scalable, cost-effective, and clinically reliable alternative to monolithic AI systems, enabling efficient healthcare assistance across diverse clinical and administrative scenarios.

## F. Notation Summary

TABLE I: NOTATION SUMMARY USED IN THE PROPOSED SYSTEM

| Symbol | Description |
|---|---|
| $Q$ | Input user query |
| $R$ | User role |
| $q$ | Query embedding vector |
| $d$ | Document embedding vector |
| $S(q, d)$ | Cosine similarity score |
| $D_{raw}$ | Initial diagnostic output |
| $\vartheta$ | Confidence threshold |
| $D$ | Final diagnostic output |
| $R_{final}$ | Unified system response |

## IV. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed multi-agent healthcare assistance system was implemented using Python 3.10 for backend logic, React for the frontend interface, and LangChain as a modular agent orchestration framework for coordinating diagnostics, search, and hospital operations. Experiments were conducted on a simulated hospital environment with curated medical datasets, trusted external healthcare sources, and synthetic hospital operational data to evaluate system performance, accuracy, and efficiency.

### *A. Implementation Results*

The correct execution and integration of all system com- ponents, as outlined in Fig. 1, were validated through the following stages:

- Manager Agent Orchestration: The Manager Agent successfully classified incoming user queries and dynamically routed them to the appropriate specialized agents. Multi-agent coordination was verified across diagnostic, informational, and operational workflows, ensuring consistent response synthesis.
- Diagnostic Reasoning Pipeline: The Diagnostics Agent successfully integrated a medical BERT model with a RAG pipeline. Clinical queries were accurately matched against the medical vector database, and retrieved documents were effectively incorporated into diagnostic reasoning. Confidence-based invocation of retrieval demonstrated reduced hallucinations in ambiguous cases.
- Trusted Knowledge Retrieval: The Search Agent successfully retrieved and filtered medical information exclusively from trusted government and institutional health- care sources using a dedicated search configuration. Retrieved evidence was ranked and summarized before being forwarded for diagnostic validation.
- Hospital Operations Analytics: The Hospital Opera- tions Agent correctly processed structured hospital data related to patient inflow, bed availability, and resource uti- lization. Visualization modules generated real-time plots and summaries, enabling rapid operational insights for administrators and clinicians.

TABLE II: SYSTEM PERFORMANCE METRICS

| Metric | Value |
|---|---|
| Diagnostic Accuracy (Benchmark Dataset) | ≈ 92.4% |
| Retrieval Precision (Top-5 Documents) | 93.1% |
| Retrieval Recall (Top-5 Documents) | 90.7% |
| Average Diagnostic Response Time | 1.2–1.8 s |
| Hallucination Reduction (vs. Base LLM) | ≈ 41% |

TABLE III: KNOWLEDGE RETRIEVAL AND SEARCH METRICS

| Metric | Value |
|---|---|
| Trusted Medical Domains | 10 |
| Average Search Latency | 0.9–1.3 s |
| Relevance Accuracy | ≈ 94% |
| Summarization Compression Ratio | 6:1 |
| External Evidence Validation Rate | 96.2% |

TABLE IV: HOSPITAL OPERATIONS ANALYTICS PERFORMANCE

| Metric | Value |
|---|---|
| Average Data Query Time | 0.4–0.7 s |
| Visualization Rendering Time | 0.6–1.0 s |
| Resource Prediction Accuracy | ≈ 91.8% |
| Concurrent Queries Supported | 25–30 |
| Dashboard Update Frequency | Real-Time |

TABLE V: SYSTEM RESOURCE AND STORAGE REQUIREMENTS

| Component | Size |
|---|---|
| Medical Document Embedding (per record) | 1.8 KB |
| Query Embedding | 768 bytes |
| RAG Context Window (avg) | 8–12 KB |
| Session Metadata | 512 bytes |
| Hospital Analytics Cache | 3–5 KB |
| Total per Active Session | 14–20 KB |

- End-to-End System Integration: The complete workflow—from user query submission to unified response generation—was executed without failure across all test scenarios. Representative system outputs, including diagnostic summaries and operational dashboards, are illustrated in Fig. 1.

### B. Performance Metrics

Based on our testing and model specifications:

## V. CONCLUSION AND FUTURE WORK

This paper presented a centralized multi-agent healthcare assistance framework that integrates RAG with explicit agent orchestration to address key limitations of monolithic and fine-tuned AI models in clinical settings. By decomposing healthcare workflows into specialized agents through a Man- ager Agent, the proposed system achieves improved diagnostic accuracy, reduced hallucinations, lower computational cost, and faster response times. Experimental results demonstrate that the agentic RAG-based approach outperforms traditional LLMs and fine-tuned transformers in terms of adaptability, scalability, and operational efficiency, while maintaining strong clinical reliability.

TABLE VI: BASELINE COMPARISON OF HEALTHCARE AI APPROACHES

| Metric | Monolithic LLM | Fine-Tuned Model | Proposed System |
|---|---|---|---|
| Knowledge Update | Retraining | Re-fine-tuning | RAG |
| Response Time | High | Medium | Low |
| Diagnostic Accuracy | Medium | High | Higher |
| Hallucination Risk | High | Medium | Low |
| Scalability | Limited | Moderate | High |
| Maintenance Cost | High | High | Low |

The centralized orchestration mechanism ensures coherent decision-making across agents, enabling seamless integration of evidence-based diagnostic reasoning, controlled external knowledge access, and hospital operational analytics within a unified system, as depicted in Fig. 1. The baseline comparison further highlights that retrieval-based updates offer a more cost-effective and maintainable alternative to continual model retraining, making the proposed architecture well-suited for dynamic healthcare environments where medical knowledge evolves rapidly.

Future work will focus on extending the framework to support multimodal clinical data, including medical imaging, physiological signals, and wearable sensor inputs. Incorporating reinforcement learning for adaptive agent coordination and personalized clinical workflows represents another promising direction. Additionally, large-scale deployment studies involving real-world hospital environments and clinician-in- the-loop evaluations will be conducted to further validate safety, usability, and clinical impact. Finally, formal regulatory alignment and explainability enhancements will be explored to facilitate adoption in safety-critical healthcare systems.

## REFERENCES

[1] F. K. Khattak, V. Subasri, A. Krishnan, C. Pou-Prom, S. Akinli-Kocak, Dolatabadi, D. Pandya, L. Seyyed-Kalantari, and F. Rudzicz, “Mlhops: Machine learning health operations,” *IEEE Access*, vol. 13, pp. 20 374–20 389, 2025.

[2] K. Khushnuma, G. Madhavi, D. Mhatre, G. Parab, and D. Chirayil, “Ai-powered virtual healthcare assistance system,” *Journal of Emerging Technologies and Innovative Research (JETIR)*, vol. 11, no. 3, pp. 107–113, 2024.

[3] L. Masanneck, S. G. Meuth, and M. Pawlitzki, “Evaluating base and retrieval-augmented large language models with document or online support for evidence-based neurology,” *npj Digital Medicine*, vol. 8, no. 1, pp. 1–9, 2025.

[4] Y. Zhu, C. Ren, Z. Wang, X. Zheng, S. Xie, J. Feng, X. Zhu, Z. Li, L. Ma, and C. Pan, “Emerge: Enhancing multimodal electronic health records predictive modeling with retrieval-augmented generation,” in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*, 2024, pp. 1–11.

[5] T. Shaik, X. Tao, L. Li, H. Xie, H.-N. Dai, F. Zhao, and J. Yong, “Adaptive multi-agent deep reinforcement learning for timely healthcare interventions,” *arXiv preprint*

*arXiv:2309.10980*, 2024.

[6] H. Pandey, A. Amod, and Shivang, “Advancing healthcare automation: Multi-agent system for medical necessity justification,” *arXiv preprint arXiv:2404.17977*, 2024.

[7] D. Power, “Collaborative care: Multi-agent systems in healthcare,” *Com- puters and Electronics in Medicine*, vol. 2, no. 1, pp. 1–5, 2025.