

ClinIQ: An AI-Based Clinical Decision Support System using Multi-Agent Reasoning, Retrieval-Augmented Generation, and Knowledge Graphs

Tanishka Gaikwad, Rudraksh Nawale, Om Badade, Vishal Nagare, and Alka Khade
Department of Computer Engineering
Terna Engineering College, Navi Mumbai, India

Abstract—Clinical decision-making requires clinicians to consider patient history, symptoms, laboratory findings, medications, and supporting medical documents. Reviewing these sources can be demanding, particularly when information is distributed across multiple records. Large language models (LLMs) offer opportunities to assist with clinical information synthesis, but their outputs may be incomplete, insufficiently grounded, or difficult to verify. This paper presents ClinIQ, an AI-based clinical decision support system that integrates retrieval-augmented generation (RAG), knowledge graph reasoning, and a multi-agent analysis workflow. The system combines structured patient information and user-provided clinical findings with relevant retrieved context to generate draft differential diagnoses, recommendations, and risk observations. Meta Llama 3 70B Instruct is accessed through Amazon Bedrock for language-based suggestion generation. The system also incorporates clinician review and decision recording to retain professional oversight. This paper describes the system architecture, implementation, data flow, and functional evaluation approach. ClinIQ is intended to support clinical information review rather than replace professional judgment; its clinical safety and effectiveness require further formal validation.

Index Terms—Clinical decision support, large language models, retrieval-augmented generation, knowledge graphs, multi-agent systems, explainable AI.

I. INTRODUCTION

Clinical decision support systems (CDSSs) help healthcare professionals interpret patient information and consider possible diagnoses or management options. However, clinical information may be distributed across patient histories, laboratory reports, medication records, examination findings, and uploaded documents. Synthesizing these sources into a coherent view can require considerable effort.

Large language models have demonstrated the ability to summarize text and generate context-sensitive responses. Their use in healthcare nevertheless presents challenges, including unsupported statements, incomplete use of patient context, and difficulty distinguishing generated suggestions from established clinical evidence. A useful decision-support system

must therefore combine language generation with relevant information retrieval, structured relationships, and appropriate human oversight.

Retrieval-augmented generation provides a way to condition model responses on retrieved information rather than relying exclusively on the model's internal parameters [1], [3]. Knowledge graphs represent entities and their relationships in a structured form, potentially supporting the exploration of connections among clinical concepts [6], [7]. Multi-agent approaches divide analysis into interacting tasks and have been investigated for clinical reasoning and information synthesis [2], [8].

This paper presents ClinIQ, a clinical decision support prototype that integrates RAG, a knowledge graph, and a multi-agent analysis workflow. Its purpose is to organize patient context and produce draft suggestions for review by a qualified healthcare professional.

The main contributions of this work are:

- An integrated clinical decision support workflow combining patient context, retrieval-augmented generation, knowledge graph integration, and multi-agent analysis.
- An implementation that connects a web interface and Python-based service layer with patient data storage, document storage, and a hosted LLM.
- A human-in-the-loop workflow in which AI-generated suggestions are presented for clinician review and decisions can be recorded.

II. RELATED WORK

A. Retrieval-Augmented Clinical Systems

RAG combines information retrieval with language generation to provide additional context for model responses. ClinicalRAG investigates heterogeneous knowledge retrieval for clinical decision support, while SurgeryLLM explores

retrieval-augmented language modelling in surgical workflows [1], [3].

The effectiveness of a RAG-based system depends on the relevance and quality of the retrieved material, the way context is constructed, and the model's ability to use that context appropriately. Retrieval alone does not guarantee that generated recommendations are clinically correct.

B. Knowledge Graphs and Multi-Agent Reasoning

Knowledge graphs represent entities and relationships explicitly. Their integration with RAG has been studied as a way to combine structured relationships with text-based knowledge [6], [7]. Related work also explores agent-based medical question resolution and clinical reasoning [?], [2], [9].

Multi-agent approaches can divide a complex task into smaller analytical responsibilities. However, the reliability of such systems depends on the coordination of agents, the evidence available to them, and the validation of their outputs [?], [8].

C. Research Motivation

Existing research explores RAG, knowledge graphs, and agent-based reasoning in different clinical settings. ClinIQ brings these components together in a working decision-support prototype and connects the AI analysis workflow to patient records, clinical document handling, and clinician decision recording. The system is intended as an implementation-oriented contribution; it does not claim that integration alone establishes clinical accuracy or improved patient outcomes.

III. SYSTEM ARCHITECTURE

ClinIQ is organized around a clinical analysis workflow that receives patient context, incorporates retrieved information and knowledge graph relationships, and uses multi-agent reasoning to support language-based suggestion generation. The resulting suggestions are presented for professional review.

Fig. 1 illustrates the AI workflow rather than the deployment infrastructure. The RAG, knowledge graph, and multi-agent components are shown as connected components of the analysis process. The diagram does not imply that every request follows a single rigid sequence through these components.

A. Clinical Input and Context Preparation

The analysis begins with information associated with a selected patient and the clinical details supplied for the current request. These may include symptoms, observations, medical history, medication information, laboratory findings, and references to uploaded documents. The backend prepares the available information for the analysis request.

The usefulness of the resulting suggestions depends on the completeness, accuracy, and relevance of the supplied patient context. Missing, outdated, or contradictory information can affect the generated output and must be considered during professional review.

B. Retrieval-Augmented Generation

The RAG component supports analysis by incorporating relevant context retrieved from available clinical information. This approach is intended to ground generation in the information available to the system rather than relying solely on a model's pre-trained knowledge.

The exact relevance of retrieved content depends on the underlying data, retrieval process, and context supplied to the model. Retrieved information should not automatically be treated as authoritative evidence; its relevance and quality must be assessed.

C. Knowledge Graph Integration

The knowledge graph is integrated into the AI analysis workflow to make represented relationships among clinical concepts available during analysis. Graph-derived context can complement textual patient information by exposing relationships that may not be apparent from an isolated record or document.

The graph can support reasoning over represented relationships, but its usefulness is bounded by the coverage, correctness, and currency of the information it contains. The graph should not be interpreted as a complete representation of medical knowledge.

D. Multi-Agent Analysis and Suggestion Generation

The multi-agent component supports the analysis workflow by coordinating processing of available clinical context and information provided by the retrieval and knowledge graph components. The resulting context is used in language-based suggestion generation through Meta Llama 3 70B Instruct accessed via Amazon Bedrock.

The system can generate draft differential diagnoses, recommendations, and risk observations. These outputs are intended to support review and must not be treated as confirmed diagnoses, prescriptions, or autonomous treatment decisions.

E. Clinician Oversight

ClinIQ incorporates a human-in-the-loop workflow. A clinician can review the generated suggestions and, where supported by the application, accept, modify, or reject them and record a decision. This preserves the distinction between model-generated content and the decision made by the health-care professional.

IV. DATA SOURCES AND IMPLEMENTATION

A. Data Sources

ClinIQ uses information available through the application's patient records, clinical forms, analysis requests, and uploaded documents. The principal categories are summarized in Table I.

The prototype also supports a local JSON-based data fallback for development and demonstration. The availability and completeness of records depend on the selected environment and the data entered or uploaded into the application.

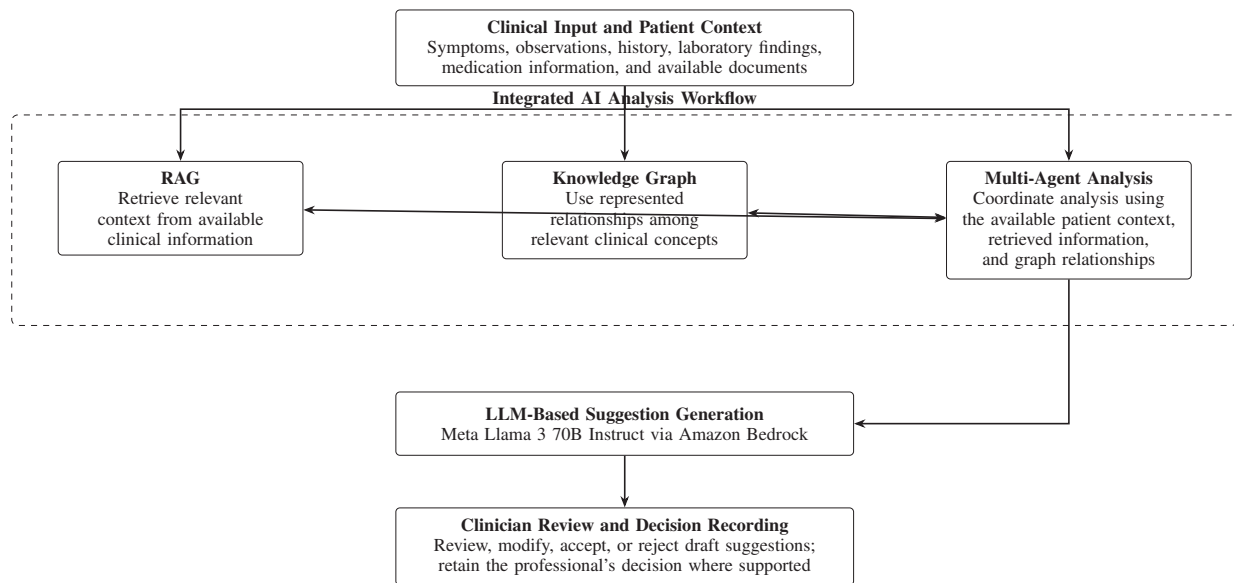


Fig. 1. Conceptual AI workflow of ClinIQ. Retrieval, knowledge graph integration, and multi-agent analysis interact within the analysis workflow. The LLM generates draft suggestions for clinician review.

TABLE I
PRINCIPAL DATA CATEGORIES IN CLINIQ

Category	Examples
Patient records	Patient details, history, appointments, and medication records.
Clinical inputs	Symptoms, vital signs, examination observations, and user-entered findings.
Uploaded documents	Laboratory reports, imaging, prescriptions, and clinical notes in supported file formats.
Analysis context	Clinical information assembled for an AI analysis request.
AI outputs	Draft differential diagnoses, recommendations, and risk observations.
Decision records	Clinician decisions, comments, and associated audit information where implemented.

TABLE II
IMPLEMENTATION TECHNOLOGIES

Component	Technology
Frontend	React 19, TypeScript, Vite, Tailwind CSS v4.
Backend API	Python, FastAPI, Uvicorn, Pydantic.
Authentication	JWT-based authentication using <code>python-jose</code> ; <code>Passlib</code> and <code>bcrypt</code> .
Primary database	Supabase, with a local JSON-based fallback for development.
AI inference	Amazon Bedrock with Meta Llama 3 70B Instruct.
Document storage	Amazon S3 and time-limited pre-signed URLs.
Configuration	Environment variables and <code>python-dotenv</code> .
Testing	cURL requests and Python test scripts.

B. Implementation Stack

The system uses a web-based interface and a Python service layer. The major implementation technologies are summarized in Table II.

The frontend communicates with the backend API, which manages application requests and coordinates access to configured data services and the AI model. Supabase provides the primary database integration, while the local JSON fallback supports local demonstration. Amazon S3 is used for document storage, with server-side encryption and time-limited pre-signed URLs when configured in the deployment.

These technologies describe the software implementation and are separate from the AI workflow illustrated in Fig. 1.

C. Analysis Workflow

The system's analysis process can be summarized as follows:

- 1) An authorized user selects a patient and supplies the clinical information required for an analysis request.
- 2) Relevant patient context and available document references are assembled by the application.
- 3) The integrated analysis workflow uses available retrieval results, knowledge graph relationships, and multi-agent processing to prepare context for generation.
- 4) The prepared context is sent to Meta Llama 3 70B Instruct through Amazon Bedrock to generate draft suggestions.
- 5) The application presents the suggestions for clinician review.
- 6) The clinician's decision can be recorded through the application's supported decision workflow.

The information available to each stage depends on the configured data sources, the contents of the request, and the functionality enabled in the running environment.

V. FUNCTIONAL EVALUATION APPROACH

The prototype includes test scripts intended to exercise key application workflows, including authentication, document uploads, clinical analysis requests, clinician decision handling, and cleanup operations. These tests address application functionality and integration behavior rather than establishing clinical effectiveness.

A complete evaluation of a clinical decision support system should distinguish software correctness from the quality and safety of its clinical suggestions. Functional tests can establish whether application operations behave as expected under the tested conditions, but they do not independently demonstrate diagnostic accuracy, reduced clinician workload, or improved patient outcomes.

Further evaluation should include a documented test dataset, predefined evaluation criteria, reproducible test procedures, and quantitative results. Clinical output assessment should involve qualified reviewers and an appropriate reference standard. Any reported measurements should be calculated from recorded test outputs and should include the evaluation conditions under which they were obtained.

VI. DISCUSSION

ClinIQ combines structured patient information, document-based context, knowledge graph relationships, multi-agent analysis, and LLM-based suggestion generation in one application workflow. This design provides a foundation for exploring how different information sources and reasoning components can assist clinical review.

The integration has several potential benefits. Retrieval can make relevant available information accessible during generation; the knowledge graph can expose represented relationships; and multi-agent processing can organize analysis across components. Clinician review provides a point at which generated content can be checked against professional knowledge and the patient's circumstances.

These potential benefits should not be confused with demonstrated clinical improvements. A language model can produce incorrect or unsupported statements, retrieved information can be irrelevant, and a knowledge graph can contain incomplete or inaccurate relationships. The presence of multiple agents does not automatically eliminate these risks.

VII. LIMITATIONS AND FUTURE WORK

The current system is a decision-support prototype and should not be considered a clinically validated diagnostic system. Its output quality depends on the available patient information, retrieval quality, graph coverage, model behavior, and the way the components interact.

Future work should include a reproducible quantitative evaluation of retrieval relevance, output completeness, factual consistency, and the quality of generated suggestions. A clinician-led evaluation using an appropriately governed dataset would be necessary to assess clinical usefulness and identify failure modes.

Additional work should document the knowledge graph schema, retrieval configuration, agent responsibilities, and evaluation protocol in sufficient detail for reproducibility. Privacy, access control, data retention, and auditability should also be assessed under the intended deployment conditions. Clinical use would require appropriate validation, governance, and oversight.

VIII. CONCLUSION

This paper presented ClinIQ, an AI-based clinical decision support prototype integrating retrieval-augmented generation, knowledge graph integration, and multi-agent analysis with a hosted large language model. The system uses patient context and available clinical information to generate draft suggestions that can be reviewed by a healthcare professional. Its implementation connects a web interface and backend service with configured data and document storage services.

ClinIQ provides a basis for further investigation of integrated AI assistance in clinical information review. Its clinical accuracy, safety, and practical effectiveness remain subjects for systematic evaluation. Professional oversight remains essential, and the system is not intended to replace clinical judgment.

REFERENCES

- [1] Y. Lu, X. Zhao, and J. Wang, "ClinicalRAG: Enhancing clinical decision support through heterogeneous knowledge retrieval," in *Proc. KnowLLM*, 2024.
- [2] C. U. Ögdu, K. Arslanoglu, and M. Karakose, "An adaptive multi-agent LLM-based clinical decision support system integrating biomedical RAG and web intelligence," *IEEE Access*, 2025.
- [3] C. S. Ong, N. T. Obey, Y. Zheng, A. Cohan, and E. B. Schneider, "SurgeryLLM: A retrieval-augmented generation large language model framework for surgical decision support and workflow enhancement," *npj Digital Medicine*, 2024.
- [4] F. Gaber *et al.*, "Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis," *npj Digital Medicine*, 2025.
- [5] M. A. Cypko *et al.*, "Large language models with retrieval-augmented generation enhance expert modelling of Bayesian network for clinical decision support," *International Journal of Computer Assisted Radiology and Surgery*, 2025.
- [6] S. Wang, H. Yang, and G. Bai, "Construction of intelligent decision support systems through integration of retrieval-augmented generation and knowledge graphs," *Scientific Reports*, 2025.
- [7] V. Yadav, Gaurav, A. Rana, and S. Sharma, "MedRAG-Agent: Medical query resolution by employing a multi-agent, knowledge graph-enhanced RAG-based AI framework," in *Proc. 2025 IEEE 6th Global Conference on Advancement Technology (GCAT)*, 2025.
- [8] P. Shi, Y. Shui, and Q. Pan, "ER-MedRAG: A multi-agent reinforcement learning framework for reliable clinical retrieval-augmented reasoning," 2026.
- [9] Y. Lu, G. Fu, W. Wu, X. Zhao, S. Y. Goi, and J. Wang, "DoctorRAG: Medical RAG fusing knowledge with patient analogy through textual gradients," *arXiv*, 2025.
- [10] J. Moll, J. Lueberstedt, *et al.*, "Agentic clinical reasoning over longitudinal myeloma records: A retrospective evaluation against expert consensus," *arXiv*, 2026.
- [11] M. H. Doshi, D. Haldar, *et al.*, "Generative AI for medical digital twin using RAG: A clinical decision support approach," in *Proc. 2025 2nd International Conference on Artificial Intelligence Innovation in Healthcare Industry (ICAIIHI)*, 2025.
- [12] K. G. Maid and S. Gaikwad, "AI medical health assistant: RAG approach," in *Proc. 2025 International Conference on Information, Implementation, and Innovation in Technology (I2ITCON)*, 2025.