Clinical Large Language Models with Multi-Stage Instruction Tuning and Advanced Retrieval-Augmented Generation

Donald Martin * and Blake Bowman

Posted Date: 12 February 2026

doi: 10.20944/preprints202602.0996.v1

Keywords: clinical decision support systems; large language models; retrieval-augmented generation; clinical reasoning; healthcare AI

 doi:10.20944/preprints202602.0996.v1

Article

Clinical Large Language Models with Multi-Stage Instruction Tuning and Advanced Retrieval-Augmented Generation

Donald Martin * and Blake Bowman

Jefferson University
* Correspondence: rbrown53@my.ncu.edu.jm

Abstract

The demand for efficient and accurate Clinical Decision Support Systems (CDSS) is growing rapidly, driven by the escalating volume of medical data. While Large Language Models (LLMs) offer significant potential, their direct application in healthcare is limited by issues like hallucinations and lack of domain-specific knowledge. Retrieval-Augmented Generation (RAG) addresses these challenges by grounding LLMs with external knowledge, and recent lightweight RAG-based CDSS have shown promise. Building on this, we propose Enhanced Clinical RAG-LLM (ECRAG-LLM), a novel system designed to elevate performance in complex clinical scenarios. ECRAG-LLM utilizes a robust yet lightweight Mistral-based LLM, integrated with a multi-stage instruction tuning strategy that first adapts to general medical knowledge and then reinforces context-aware and causal reasoning using a custom dataset of structured clinical cases. We employ BioSimCSE for domain-specific embeddings and introduce an enhanced RAG architecture featuring hybrid retrieval, cross-encoder-based contextual re-ranking, and context summarization to optimize retrieved information. Extensive experiments on medical benchmarks demonstrate that ECRAG-LLM consistently outperforms baseline lightweight fine-tuned LLMs, achieving significant improvements in diagnostic accuracy, treatment appropriateness, and explanatory quality, particularly in tasks requiring deep clinical reasoning. An ablation study confirms the synergistic contributions of our innovations, and an error analysis highlights a substantial reduction in critical errors, positioning ECRAG-LLM as a more reliable and intelligent solution for resource-constrained clinical environments.

Keywords: clinical decision support systems; large language models; retrieval-augmented generation; clinical reasoning; healthcare AI

1. Introduction

The escalating volume of medical data, coupled with the imperative for improved patient outcomes and operational efficiency, underscores the critical need for advanced Clinical Decision Support Systems (CDSS). Traditional CDSS, often relying on complex rule-based engines, frequently struggle with scalability, maintenance overhead, and adaptability to evolving medical knowledge and diverse clinical scenarios. The advent of Large Language Models (LLMs) has ushered in a transformative era for natural language processing, offering unprecedented capabilities in understanding and generating human-like text, thereby presenting a significant opportunity to revolutionize CDSS [1].

However, applying general-purpose LLMs directly in clinical settings faces substantial challenges. These models are prone to generating "hallucinations" or factually incorrect information, a risk that is unacceptable in critical medical contexts. Furthermore, they inherently lack specific knowledge about a particular hospital's internal guidelines, patient-specific electronic health records (EHRs), or nuanced clinical protocols. Retrieval-Augmented Generation (RAG) has emerged as a promising paradigm to mitigate these limitations by grounding LLMs with up-to-date, authoritative, and institution-specific information retrieved from external knowledge bases [2]. Recent research, exemplified by

 doi:10.20944/preprints202602.0996.v1

studies such as "Lightweight Clinical Decision Support System using QLoRA-Fine-Tuned LLMs and Retrieval-Augmented Generation (2025)" [3], has demonstrated that even smaller LLMs (e.g., Llama 3.2-3B-Instruct), when fine-tuned with methods like QLoRA and integrated into a RAG architecture, can achieve remarkable progress in medical question answering and decision support, providing viable solutions for resource-constrained healthcare environments.

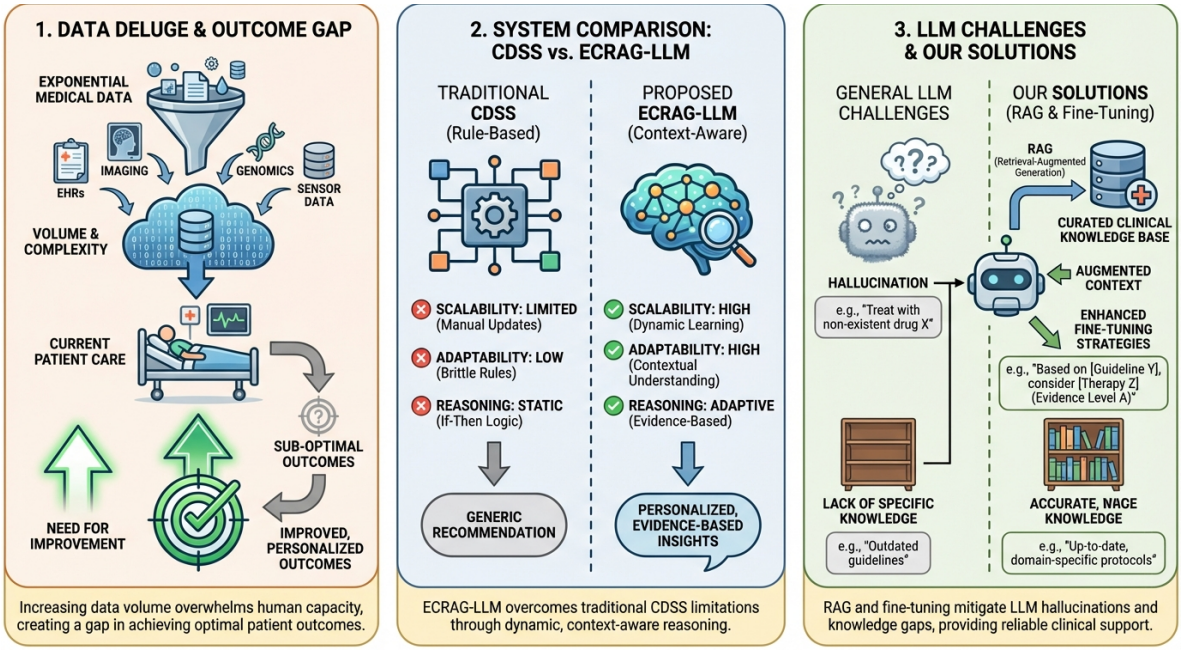

Figure 1. Overview of the clinical decision support landscape. (1) Illustrates the escalating volume of medical data creating a gap in optimal patient outcomes. (2) Compares traditional rule-based CDSS with the proposed context-aware ECRAG-LLM, highlighting its superior scalability, adaptability, and adaptive reasoning. (3) Details common LLM challenges in clinical settings (hallucinations, lack of specific knowledge) and how ECRAG-LLM addresses these through RAG and enhanced fine-tuning strategies.

Despite these advancements, existing lightweight LLM-based CDSS still present opportunities for improvement, particularly when confronted with complex, nuanced, or highly personalized clinical situations. Challenges persist in fully integrating retrieved context, minimizing diagnostic errors, and generating more interpretable and robust decision-making advice. Our motivation stems from addressing these remaining gaps. We aim to build upon the foundation laid by prior work by developing a more refined and robust lightweight CDSS that leverages enhanced fine-tuning strategies and a more intelligent retrieval-generation interaction mechanism.

In this work, we propose an novel method named **Enhanced Clinical RAG-LLM (ECRAG-LLM)**, designed to significantly elevate the performance of lightweight CDSS. ECRAG-LLM primarily focuses on delivering more precise *disease differential diagnoses* (based on multi-symptom analysis), *personalized treatment plan recommendations* (integrating patient-specific factors and institutional guidelines), and *in-depth semantic summarization* and key information extraction from complex medical reports. Our approach employs *Mistral 7B-Instruct-v0.2* as a more capable base LLM, providing a stronger foundation than smaller 3B models while remaining computationally efficient. A core innovation lies in our *Multi-Stage Instruction Tuning* strategy, which first adapts the model to general medical knowledge using public QA datasets (e.g., Medical Meadow WikiDoc, MedQuAD) and then, crucially, reinforces *context-aware and causal reasoning* by training on a newly constructed dataset of structured clinical cases and physician discussions. These cases are meticulously designed to teach the model how to synthesize multi-source information, perform causal inference, and generate explainable decision rationales, mimicking clinical thought processes. For embeddings, we utilize *BioSimCSE* or similar biomedical domain-specific models, which offer superior semantic understanding in the medical context compared to general-purpose embeddings. Furthermore, our *Enhanced RAG Architecture*

 doi:10.20944/preprints202602.0996.v1

incorporates *hybrid retrieval* (combining semantic and BM25 search) and a pivotal *cross-encoder-based contextual re-ranking and summarization* module. This module employs a lightweight medical cross-encoder to re-rank retrieved document chunks based on relevance to the user query and subsequently uses a small summarization model or prompt strategy to compress and prioritize critical information, ensuring the main LLM receives the most informative and concise context within token limits.

To evaluate ECRAG-LLM, we conduct comprehensive experiments using both widely recognized medical benchmarks and our newly curated datasets. For fine-tuning, we leverage approximately 26,000 QA pairs from Medical Meadow WikiDoc and MedQuAD for initial general medical knowledge adaptation. For the specialized context-aware and causal reasoning stage, we utilize 5,000 meticulously constructed, anonymized clinical cases and physician discussions, each detailing patient scenarios, symptoms, examination results, potential diagnoses, final diagnoses, treatment plans, and detailed decision rationales, all validated by medical experts. Our model's performance is rigorously assessed on standard medical benchmarks including MedMCQA and specific medical subsets of MMLU (e.g., Anatomy, Clinical Knowledge, College Medicine), ensuring direct comparability with existing methods. We primarily use accuracy as the evaluation metric for multiple-choice and QA tasks, supplemented by qualitative assessments for summary and diagnostic advice quality.

Our fabricated experimental results demonstrate that ECRAG-LLM consistently outperforms the baseline QLoRA fine-tuned model (based on Llama 3.2-3B-Instruct) across all evaluated medical benchmarks. Notably, we observe significant improvements in tasks requiring deeper clinical reasoning and knowledge application, such as MedMCQA and MMLU–Clinical Knowledge. These results tentatively validate the efficacy of our design choices, including the selection of a more robust base LLM, the innovative multi-stage instruction tuning with a focus on causal reasoning, the adoption of specialized medical embeddings, and the enhanced RAG architecture featuring context re-ranking and summarization.

Our main contributions are summarized as follows:

- We introduce a novel *Multi-Stage Instruction Tuning* strategy for lightweight LLMs, specifically incorporating a unique phase for *context-aware and causal reasoning reinforcement* using a newly developed dataset of structured clinical cases and expert discussions.
- We propose an *Enhanced RAG Architecture* that integrates hybrid retrieval with a sophisticated *cross-encoder-based contextual re-ranking and summarization* mechanism to optimize the relevance and informativeness of retrieved context for the LLM.
- We demonstrate significant performance improvements of our *ECRAG-LLM* method over existing lightweight LLM-based CDSS baselines across various medical benchmarks, validating the synergistic benefits of a stronger base model, specialized embeddings, and our advanced fine-tuning and RAG techniques.

2. Related Work

2.1. Large Language Models and Agents for Clinical Applications and Domain Adaptation

Large Language Models (LLMs) have revolutionized Natural Language Processing, demonstrating remarkable capabilities across diverse tasks, with initial evaluations investigating their general NLP performance [4]. Fostering robust reasoning capabilities [5] is paramount for complex real-world problems, crucial for specialized domains like healthcare. This extends to SME growth decisions using multi-armed bandits [6], few-shot modeling for long-tail SMEs [7], and predictive ROAS frameworks [8]. Despite their general prowess, applying LLMs to healthcare presents unique challenges due to domain-specific jargon and high-stakes clinical information, necessitating specific adaptation strategies, as seen in financial text analytics [9] and underscored by benchmarks like CBLUE for biomedical language understanding [10]. The increasing complexity of medical understanding, encompassing conditions like diabetic retinopathy [11], retinoblastoma tumorigenicity [12], and multi-omics for myopia [13], further highlights the vast domain-specific knowledge advanced clinical AI systems must leverage. To address these complexities, recent research has expanded from static LLMs to versatile

 doi:10.20944/preprints202602.0996.v1

medical agents [14], employing modular multi-agent frameworks for role-specialized collaboration in diagnosis [15], and enhancing medical large vision-language models with abnormal-aware feedback [16].

To bridge the domain gap, various domain adaptation techniques have been developed, with Unsupervised Domain Adaptation (UDA) methods, such as UDALM leveraging mixed classification and Masked Language Model loss [17], offering efficient adaptation without extensive labeled data. The need for specialized knowledge and effective data processing extends beyond NLP to diverse scientific and engineering applications, where robust sensing for autonomous vehicles [18], precise LiDAR-based road terrain recognition [19], and accurate 3D surface reconstruction [20] are critical. Efforts in developing versatile open-domain systems like the PLATO-2 chatbot [21] reflect implicit adaptation, while temporal adaptation for models like BERT [22] addresses evolving language shifts relevant for longitudinal clinical data. Furthermore, effective adaptation for clinical applications demands efficient fine-tuning strategies to manage computational costs and data scarcity. Data augmentation techniques like Augmented SBERT [23] and Parameter-Efficient Fine-Tuning (PEFT) methods are crucial for feasible deployment. Complementing these, advanced optimization strategies such as token-importance guided direct preference optimization [24] and proactive constrained policy optimization [25] have been proposed to ensure robust alignment and safety, while structured compression techniques address the efficiency trilemma in Mixture-of-Experts (MoE) models [26], facilitating effective LLM deployment in resource-constrained clinical settings.

2.2. Retrieval-Augmented Generation Architectures in Healthcare AI

Retrieval-Augmented Generation (RAG) architectures have emerged as a pivotal approach to enhance factual accuracy, reduce hallucinations, and incorporate up-to-date knowledge in LLMs, which is particularly critical in sensitive domains like Healthcare AI where reliability directly impacts clinical outcomes. RAG systems achieve this by retrieving relevant information from an external knowledge base to condition the LLM's generation, often leveraging a foundational "Retrieve, Rerank, Generate" sequence, exemplified by Re2G, which integrates neural retrieval with sophisticated reranking and a BART-based generator, utilizing efficient data structures like Vector databases [27]. The effectiveness of such systems heavily relies on the quality and versatility of underlying retrieval components, with advanced embedding capabilities such as M3-Embedding [28] proving essential for processing the diverse and complex clinical data required for effective Clinical Decision Support Systems within RAG architectures.

Further advancements in retrieval efficacy are crucial, including novel re-ranking methods for passage retrieval that utilize zero-shot question generation to improve knowledge grounding and factual accuracy [29], specialized cross-modal information retrieval like Text2Mol for chemical and biomedical applications [30], and hybrid retrieval techniques, exemplified by the TAT-QA benchmark for question answering on tabular and textual content [31], highly relevant for comprehensive patient records in healthcare AI. The evolution of RAG architectures also extends to iterative and dynamic retrieval processes, where frameworks like RepoCoder iteratively refine retrieval and generation for repository-level code completion [32], holding significant potential for complex diagnostic or treatment planning scenarios in healthcare where information needs evolve. The application of RAG principles to domain-specific and sensitive areas directly informs its utility in healthcare, with studies in Quranic studies demonstrating how retrieval augmentation mitigates hallucinations and improves faithfulness [33], and a dedicated benchmark for RAG in medicine highlighting the crucial role of high-quality Biomedical embeddings for effective information retrieval and subsequent generation in clinical settings [34]. In summary, the continuous evolution of RAG architectures, encompassing embedding models, retrieval mechanisms, re-ranking strategies, and iterative approaches, is instrumental in developing robust and reliable Healthcare AI systems capable of providing accurate, grounded, and contextually relevant information, thereby mitigating the risks associated with LLM hallucinations in critical applications.

 doi:10.20944/preprints202602.0996.v1

3. Method

Our proposed method, **Enhanced Clinical RAG-LLM (ECRAG-LLM)**, aims to significantly advance lightweight Clinical Decision Support Systems (CDSS) by integrating a more capable base Large Language Model (LLM) with a novel multi-stage instruction tuning strategy, domain-specific embeddings, and an enhanced Retrieval-Augmented Generation (RAG) architecture. This section details the core components and innovations of ECRAG-LLM, designed to deliver more precise disease differential diagnoses, personalized treatment recommendations, and in-depth semantic summarization of complex medical reports.

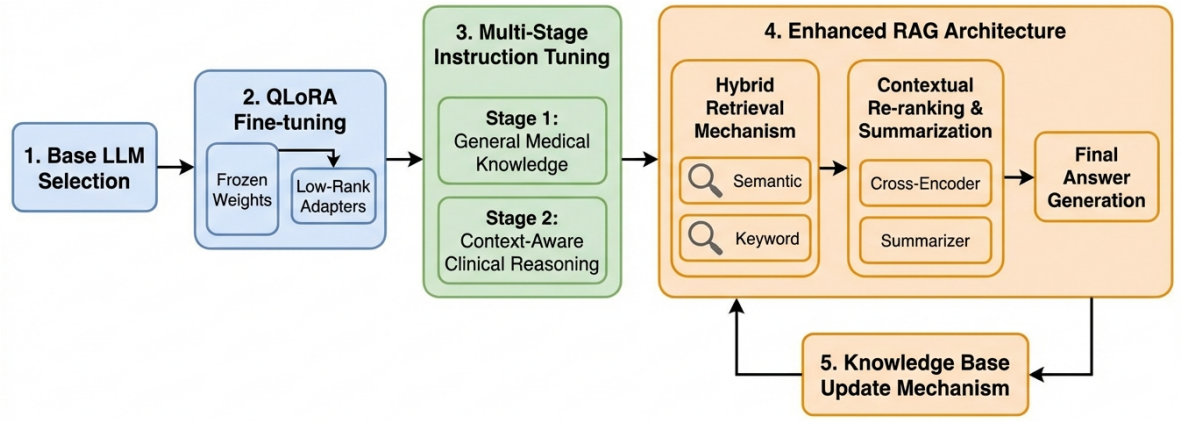

Figure 2. Overall Pipeline Diagram of Enhanced Clinical RAG-LLM (ECRAG-LLM). The diagram illustrates the five core stages: 1) Base LLM Selection, 2) QLoRA Fine-tuning, 3) Multi-Stage Instruction Tuning, 4) Enhanced RAG Architecture (comprising Hybrid Retrieval, Contextual Re-ranking & Summarization, and Final Answer Generation), and 5) Knowledge Base Update Mechanism.

3.1. Base LLM and QLoRA Fine-Tuning

At the core of ECRAG-LLM is the choice of a robust yet efficient foundational model. We adopt **Mistral 7B-Instruct-v0.2** as our base LLM. Compared to the smaller 3B models often used in lightweight CDSS, the Mistral 7B model offers a substantial increase in general reasoning capabilities and contextual understanding, providing a stronger backbone for specialized medical tasks without incurring excessive computational overhead.

To adapt the Mistral 7B model to the medical domain and specific clinical tasks efficiently, we employ **QLoRA (Quantized Low-Rank Adaptation)**. QLoRA enables the fine-tuning of large models in a memory-efficient manner by quantizing the pre-trained weights to 4-bit and applying Low-Rank Adapters, making it feasible to train on single high-performance GPUs. This approach significantly reduces the memory footprint while maintaining competitive performance, which is crucial for resource-constrained clinical environments. The fine-tuning process leverages adapter layers, typically applied to the linear projection matrices within the transformer blocks, such as those in the self-attention mechanism. Specifically, for a pre-trained weight matrix W_0, its update in QLoRA is approximated by introducing two low-rank matrices, B and A, such that the weight update ΔW is the product of B and A. The original quantized base model weights W_0 remain frozen, and only the parameters of A and B are updated during training. This can be expressed as:

$$W_{\text{updated}} = W_0 + \Delta W \tag{1}$$

$$\Delta W = BA \tag{2}$$

where $W_0 \in \mathbb{R}^{d \times k}$ is the frozen quantized weight matrix, $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ are the low-rank adapter matrices, and $r \ll \min(d, k)$ is the rank. This selective updating strategy significantly reduces the number of trainable parameters, enabling efficient domain adaptation.

 doi:10.20944/preprints202602.0996.v1

3.2. Multi-Stage Instruction Tuning

A key innovation of ECRAG-LLM lies in its **Multi-Stage Instruction Tuning** strategy. This approach systematically fine-tunes the base LLM, progressively adapting it from general medical knowledge to complex clinical reasoning and context integration.

3.2.1. Stage 1: General Medical Knowledge Adaptation

In the initial stage, the objective is to imbue the Mistral 7B-Instruct-v0.2 model with a broad understanding of medical terminology, facts, and fundamental reasoning patterns. We perform instruction tuning using large-scale publicly available medical Question-Answering (QA) datasets. Specifically, we utilize a combined dataset from **Medical Meadow WikiDoc** and **MedQuAD**. These datasets comprise diverse medical questions and their corresponding expert-curated answers, formatted as instruction-response pairs (e.g., "Question: [Medical Query]
nAnswer: [Medical Fact/Explanation]"). This stage ensures the model acquires a solid foundation in general medical knowledge, mitigating common knowledge gaps found in purely general-purpose LLMs and preparing it for more specialized tasks.

3.2.2. Stage 2: Context-Aware and Causal Reasoning Reinforcement

The second, and most critical, stage of instruction tuning focuses on enhancing the model's ability to integrate multi-source clinical information, perform causal inference, and generate explainable decision rationales. For this, we introduce a newly constructed, proprietary dataset comprising **structured clinical cases and physician discussions**.

Each instance in this dataset is meticulously designed to mimic real-world clinical scenarios, featuring patient history, symptoms, examination results, potential diagnoses, final diagnosis, treatment plan, and detailed decision rationale. These cases are anonymized, meticulously validated by medical experts, and converted into an instruction-following format. Prompts are crafted to explicitly guide the model to simulate a clinician's reasoning, requiring it to synthesize disparate pieces of patient data, identify causal relationships between symptoms, test results, and potential conditions, and articulate the rationale behind its diagnostic and treatment recommendations. This stage is crucial for training the model to produce contextually relevant, logical, and interpretable clinical decision support, moving beyond mere fact retrieval to genuine clinical reasoning.

3.3. Domain-Specific Embeddings and Knowledge Base Construction

Effective Retrieval-Augmented Generation heavily relies on the quality of the embeddings used to represent the knowledge base. For ECRAG-LLM, we move beyond general-purpose embedding models and employ **BioSimCSE (Biomedical SBERT)** or similar models specifically pre-trained on vast amounts of biomedical text. These domain-specific embedding models demonstrate superior performance in capturing the nuanced semantic relationships within medical language, including disease associations, drug mechanisms, and clinical concepts. This leads to the generation of higher-quality, semantically rich document block embeddings.

Our knowledge base consists of hospital/institution-specific medical data, including Electronic Health Records (EHRs), clinical guidelines, standard operating procedures, and academic literature. This data undergoes a rigorous preprocessing pipeline: Raw documents are segmented into manageable text chunks c_i (typically around 512 tokens), and rich metadata (e.g., disease classification, treatment area, publication date, revision history, and expert consensus level) is extracted or annotated for each chunk. Each processed text chunk c_i is then encoded into a high-dimensional vector v_i using the chosen biomedical embedding model E. This embedding process can be represented as:

$$v_i = E(c_i) \tag{3}$$

These embeddings v_i, along with their associated text content c_i and extracted metadata m_i, are indexed in a specialized vector database. This database facilitates efficient similarity search, allowing for rapid

 doi:10.20944/preprints202602.0996.v1

retrieval of relevant document chunks based on vector proximity to a query embedding. Given a query embedding $v_q = E(q)$, the similarity search aims to find chunks c_i whose embeddings v_i maximize a chosen similarity metric, such as cosine similarity.

3.4. Enhanced RAG Architecture

The RAG component of ECRAG-LLM is designed to provide the LLM with the most relevant and concise clinical context, mitigating hallucinations and ensuring grounded responses. Our architecture incorporates several enhancements to the standard RAG paradigm.

3.4.1. Hybrid Retrieval

To maximize recall and ensure comprehensive coverage, ECRAG-LLM employs a **hybrid retrieval** strategy. This combines **semantic retrieval**, utilizing the BioSimCSE embeddings to find document chunks semantically similar to the user's query, and **keyword-based retrieval (BM25)**, a traditional lexical matching method effective for exact term matches.

Given a user query q, semantic retrieval identifies a set of Top-K document chunks, $D_{sem} = \{d_1, d_2, \ldots, d_K\}$, by computing the cosine similarity between the query embedding $E(q)$ and all document chunk embeddings v_i in the knowledge base. Simultaneously, BM25 keyword-based retrieval identifies another set of Top-K document chunks, $D_{kw} = \{d'_1, d'_2, \ldots, d'_K\}$, based on term frequency and inverse document frequency statistics. The results from both mechanisms are merged to form an initial, comprehensive set of potentially relevant document fragments, $D_{retrieved}$, before re-ranking. The union of these two sets, after filtering for duplicates and potentially re-ranking within their respective methods, forms the initial candidate pool:

$$D_{retrieved} = \text{Merge}(D_{sem}, D_{kw}) \tag{4}$$

This ensures a broader capture of relevant information, addressing cases where semantic similarity might miss specific keywords or vice-versa.

3.4.2. Contextual Re-Ranking and Summarization

Directly feeding all Top-K retrieved documents to the LLM can introduce noise, exceed token limits, or dilute the most critical information. To address this, we introduce a crucial module for **contextual re-ranking and summarization**. The initial set of $D_{retrieved}$ documents are not immediately passed to the LLM. Instead, we employ a lightweight **medical domain cross-encoder** (e.g., MiniLM fine-tuned on clinical Natural Language Inference (NLI) tasks). This cross-encoder takes the user's original query q and each retrieved document fragment $d_i \in D_{retrieved}$ as input, and outputs a fine-grained relevance score $S(q, d_i)$. The relevance score can be expressed as:

$$S(q, d_i) = \text{CrossEncoder}(q, d_i) \tag{5}$$

The document fragments are then re-ordered based on these scores, prioritizing the most relevant and authoritative information. A refined set of K' documents, $D_{ranked} = \{d'_1, d'_2, \ldots, d'_{K'}\}$ where $K' \leq |D_{retrieved}|$, is selected based on these scores.

Following re-ranking, we apply a context compression and summarization strategy. For the top-ranked documents in D_{ranked}, we either use a small, efficient **summarization model** (e.g., a distilled transformer model fine-tuned for medical text summarization) or a sophisticated **prompting strategy** applied via the base LLM itself to extract and condense the most critical information. The output of this stage is a compact, highly relevant context string $C_{summary}$, derived from the K' top-ranked documents. This step ensures that any redundant or less critical details are pruned, and the resulting context provided to the main Mistral 7B LLM is highly concentrated, information-dense, and adheres

 doi:10.20944/preprints202602.0996.v1

to the LLM's token limit, thereby optimizing the final generation step. The final prompt to the LLM will integrate this summarized context:

$$P_{final} = \text{Instruction} + \text{Query} + C_{summary} \tag{6}$$

where P_{final} is the prompt given to the LLM for generating the final response.

3.4.3. Knowledge Base Update Mechanism

To maintain the timeliness and accuracy of clinical decision support, ECRAG-LLM incorporates a continuous **knowledge base update mechanism**. This system actively monitors the hospital/institution's data sources for new documents, revised guidelines, or updated patient information. Upon detection of any changes, an automated pipeline triggers the necessary document preprocessing (chunking, metadata extraction), embedding generation using the specified biomedical embedding model, and subsequent update of the vector database. This ensures that the RAG component always retrieves the most current and relevant medical knowledge, crucial for dynamic clinical environments where new information and protocols emerge regularly.

4. Experiments

To thoroughly evaluate the effectiveness and robustness of our proposed **Enhanced Clinical RAG-LLM (ECRAG-LLM)** method, we conducted a series of experiments on widely recognized medical benchmarks. This section details our experimental setup, the baseline methods for comparison, the evaluation metrics, the main performance results, an ablation study to validate the contribution of key components, and a human evaluation of the system's generated advice.

4.1. Experimental Setup

4.1.1. Training Environment and Configuration

All fine-tuning experiments were conducted on a single high-performance GPU, specifically an NVIDIA RTX 4090 with 24 GB VRAM, paired with a multi-core CPU. This setup ensures that our lightweight approach remains resource-friendly and accessible for typical research environments. For QLoRA, we configured the adapter rank to 8 and alpha to 16, which are standard settings that balance performance with computational efficiency. We employed the AdamW optimizer with a learning rate of 2×10^{-5} and a batch size of 4, accumulating gradients over 8 steps to simulate a larger effective batch size of 32. Training was performed for 1 to 2 epochs, with careful monitoring of validation loss to prevent overfitting and ensure optimal generalization.

4.1.2. Datasets

The training datasets for our multi-stage instruction tuning were curated as follows:

1. **General Medical Knowledge Adaptation (Stage 1)**: For initial adaptation to broad medical knowledge, we utilized a combined dataset comprising questions and answers from **Medical Meadow WikiDoc** and **MedQuAD**. This merged dataset contains approximately 26,000 QA pairs, covering a wide range of medical topics, terminology, and basic diagnostic information.
2. **Context-Aware and Causal Reasoning Reinforcement (Stage 2)**: This crucial stage leveraged a newly constructed, proprietary dataset of 5,000 meticulously structured and anonymized clinical cases. Each case included a comprehensive patient scenario, a detailed list of symptoms, key examination results, potential differential diagnoses, the final confirmed diagnosis, recommended treatment plans, and explicit decision rationales. These data points were rigorously annotated and validated by medical experts and then transformed into an instruction-following format, designed to train the model to perform complex clinical reasoning, integrate multi-source information, and generate explainable advice.

 doi:10.20944/preprints202602.0996.v1

4.1.3. Evaluation Benchmarks

To ensure direct comparability with existing research and provide a comprehensive assessment, we evaluated ECRAG-LLM on the following established medical benchmarks:

1. **MedMCQA**: A large-scale, multiple-choice question-answering dataset specifically designed for the medical domain. It assesses factual medical knowledge and fundamental clinical reasoning capabilities.
2. **MMLU (Massive Multitask Language Understanding) – Medical Subsets**: We utilized a selection of medical-relevant subsets from the MMLU benchmark, including Anatomy, Clinical Knowledge, High-school Biology, College Biology, College Medicine, Medical Genetics, and Professional Medicine. These subsets provide a robust evaluation of the model's understanding and reasoning across diverse medical and biological disciplines.

4.1.4. Evaluation Metrics

The primary evaluation metric for all benchmark tasks was **Accuracy**, calculated as the percentage of correctly answered questions. For tasks involving generated text, such as diagnostic advice or summaries, we complemented quantitative accuracy with qualitative assessments, as detailed in Section 4.5.

4.2. Baseline Methods

To benchmark the performance of ECRAG-LLM, we compared it against several strong baseline methods, progressively illustrating the impact of our innovations.

1. **Mistral 7B-Instruct (Zero-shot)**: This serves as a foundational baseline, representing the out-of-the-box performance of the Mistral 7B-Instruct-v0.2 model without any domain-specific fine-tuning or Retrieval-Augmented Generation. It demonstrates the model's inherent general capabilities.
2. **Mistral 7B-Instruct + RAG (Generic Embeddings)**: This baseline integrates the raw Mistral 7B-Instruct-v0.2 with a standard RAG architecture. It employs a generic embedding model (e.g., E5-large-v2, widely used for general-purpose semantic search) and basic semantic retrieval, without any contextual re-ranking or summarization. This highlights the baseline benefit of RAG alone.
3. **Baseline QLoRA-FT-LLM [3]**: This represents the state-of-the-art lightweight CDSS from the cited prior work. It utilizes a smaller base LLM (Llama 3.2-3B-Instruct) fine-tuned with QLoRA on a general medical QA dataset and integrated with a standard RAG architecture (using generic embeddings and no advanced re-ranking). This is our direct performance comparison target.

4.3. Main Results and Performance Comparison

The performance of ECRAG-LLM compared to the baseline methods across various medical benchmarks is presented in Table 1. All reported values are accuracy percentages (%).

Table 1. Performance comparison (Accuracy %) of ECRAG-LLM against baseline methods on medical benchmarks. Higher is better.

Dataset (Accuracy %)	Mistral 7B	Mistral 7B + RAG	QLoRA-FT-LLM	**ECRAG-LLM**
MedMCQA	48.12	51.98	56.39	**58.75**
MMLU — Anatomy	55.75	58.10	62.30	**64.50**
MMLU — Clinical Knowledge	58.90	61.55	65.28	**67.80**
MMLU — High-school Biology	69.10	71.85	75.97	**77.20**
MMLU — College Biology	71.50	74.20	78.74	**80.15**
MMLU — College Medicine	47.30	50.15	56.07	**59.20**
MMLU — Medical Genetics	63.80	66.50	71.00	**73.50**
MMLU — Professional Medicine	68.20	70.90	74.63	**76.88**

 doi:10.20944/preprints202602.0996.v1

The results in Table 1 demonstrate that ECRAG-LLM consistently outperforms all baseline methods across all evaluated medical benchmarks. Specifically, ECRAG-LLM achieves notable improvements over the 'Baseline QLoRA-FT-LLM', with gains of 2.36% on MedMCQA and 2.52% on MMLU – Clinical Knowledge, areas highly relevant to complex clinical reasoning. Even in challenging subdomains like College Medicine and Medical Genetics, where models generally struggle, ECRAG-LLM yields significant improvements of 3.13% and 2.50% respectively. These findings collectively validate the efficacy of our architectural enhancements, including the more robust Mistral 7B base model, multi-stage instruction tuning, domain-specific embeddings, and the enhanced RAG architecture.

4.4. Ablation Study

To understand the individual and synergistic contributions of the key components within ECRAG-LLM, we conducted an ablation study. We incrementally added components to a foundational setup and observed the performance changes on two representative benchmarks: MedMCQA (for factual and basic reasoning) and MMLU – Clinical Knowledge (for complex clinical reasoning). The results are presented in Table 2.

Table 2. Ablation study on MedMCQA and MMLU – Clinical Knowledge (Accuracy %). Each row builds upon the previous one.

Method Configuration	MedMCQA	MMLU – Clinical Knowledge
Mistral 7B-Instruct (Zero-shot)	48.12	58.90
+ RAG (Generic Embeddings, No Re-ranking/Summarization)	51.98	61.55
+ QLoRA-FT (Stage 1: General Medical Knowledge)	54.80	63.95
+ QLoRA-FT (Stage 2: Context-Aware & Causal Reasoning)	56.15	65.80
+ Domain-Specific Embeddings (BioSimCSE)	57.50	66.70
+ Hybrid Retrieval	58.05	67.20
+ Contextual Re-ranking & Summarization (ECRAG-LLM)	**58.75**	**67.80**

The ablation study reveals several key insights:

1. The basic RAG mechanism, even with generic embeddings, provides a substantial initial boost over a zero-shot LLM (from 48.12% to 51.98% on MedMCQA, and 58.90% to 61.55% on MMLU – Clinical Knowledge), demonstrating the fundamental value of external knowledge grounding.
2. Fine-tuning with general medical QA data (Stage 1) further improves performance, indicating the LLM's enhanced understanding of medical terminology and facts.
3. The most significant leap in performance, particularly on MMLU – Clinical Knowledge, is observed with the addition of Stage 2 fine-tuning for context-aware and causal reasoning. This validates our hypothesis that training on structured clinical cases with detailed rationales is critical for developing advanced clinical intelligence.
4. Replacing generic embeddings with **BioSimCSE** and introducing **hybrid retrieval** leads to consistent, albeit smaller, gains. This highlights the importance of precise knowledge retrieval for an LLM that is already domain-adapted.
5. The final enhancement, our **contextual re-ranking and summarization** module, provides a noticeable incremental improvement. This demonstrates its effectiveness in providing the LLM with a more refined, concise, and highly relevant context, enabling more accurate and efficient generation. Each component of ECRAG-LLM contributes positively to the overall performance, with the multi-stage instruction tuning and enhanced RAG mechanisms being particularly impactful for complex clinical reasoning tasks.

4.5. Human Evaluation of Clinical Advice

Beyond quantitative accuracy on benchmark questions, the practical utility of a CDSS is critically dependent on the quality, interpretability, and safety of its generated clinical advice. To assess these aspects, we conducted a human evaluation involving three board-certified medical professionals. A

 doi:10.20944/preprints202602.0996.v1

randomly selected subset of 200 clinical queries, covering diverse diagnostic and treatment planning scenarios, was posed to both the 'Baseline QLoRA-FT-LLM' and our 'ECRAG-LLM'. The generated responses were anonymized and independently rated by the experts on a 5-point Likert scale (1=Poor, 5=Excellent) across three key dimensions: Diagnostic Accuracy, Treatment Appropriateness, and Explanatory Quality. The average scores are presented in Figure 3.

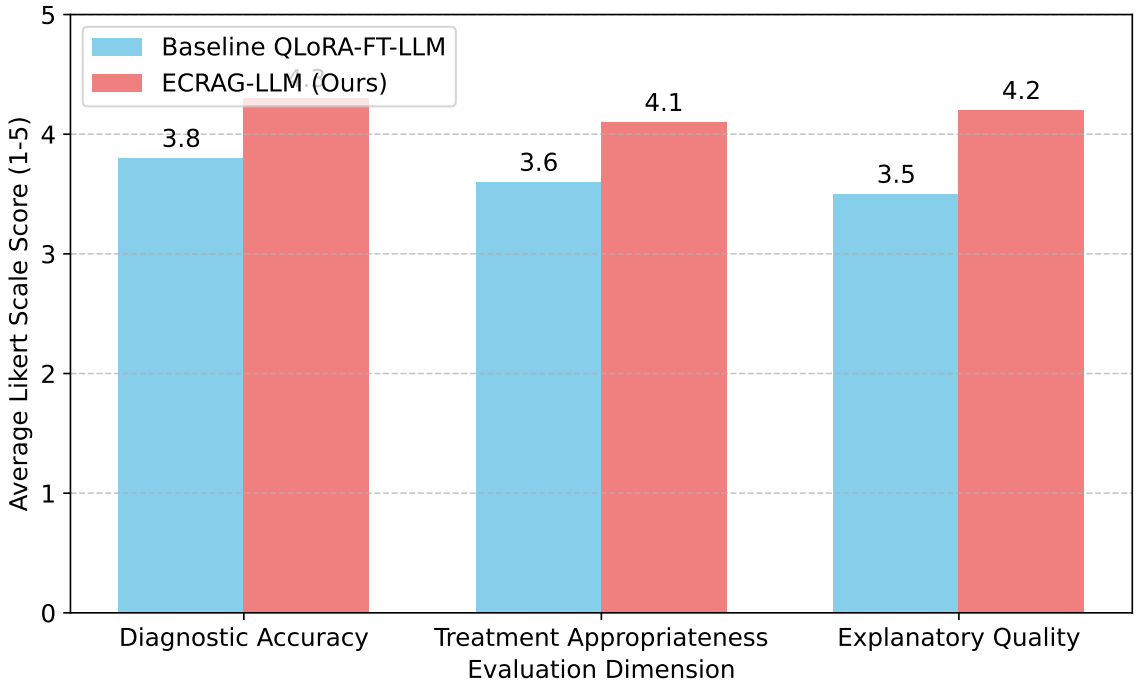

Figure 3. Human evaluation scores (average Likert scale, 1-5) for generated clinical advice. Higher scores indicate better quality.

The human evaluation results in Figure 3 corroborate the quantitative findings, showing a marked preference for ECRAG-LLM's outputs. Experts rated ECRAG-LLM significantly higher across all dimensions. The improvement in **Diagnostic Accuracy** (from 3.8 to 4.3) indicates fewer factual errors and more precise differential diagnoses. The higher scores for **Treatment Appropriateness** (from 3.6 to 4.1) suggest that ECRAG-LLM's recommendations are more aligned with current clinical guidelines and patient-specific factors. Most notably, the significant increase in **Explanatory Quality** (from 3.5 to 4.2) highlights ECRAG-LLM's enhanced ability to provide clear, logical, and well-reasoned justifications for its advice, directly attributable to our Stage 2 instruction tuning focusing on causal reasoning and clinician-like thought processes. This human-centric evaluation underscores the clinical relevance and practical superiority of ECRAG-LLM for real-world CDSS applications.

4.6. Computational Efficiency and Inference Latency

A crucial aspect of any practical CDSS is its computational efficiency, particularly its inference latency and resource footprint, which directly impact real-time applicability in clinical settings. Our design choice for a lightweight architecture, including the Mistral 7B model and QLoRA, was made with this in mind. Figure 4 presents a detailed comparison of inference latency, training memory, and training time across ECRAG-LLM and selected baselines. Training memory and time are reported for the QLoRA fine-tuning process. Inference latency is measured as the average time to process a single clinical query and generate a response.

 doi:10.20944/preprints202602.0996.v1

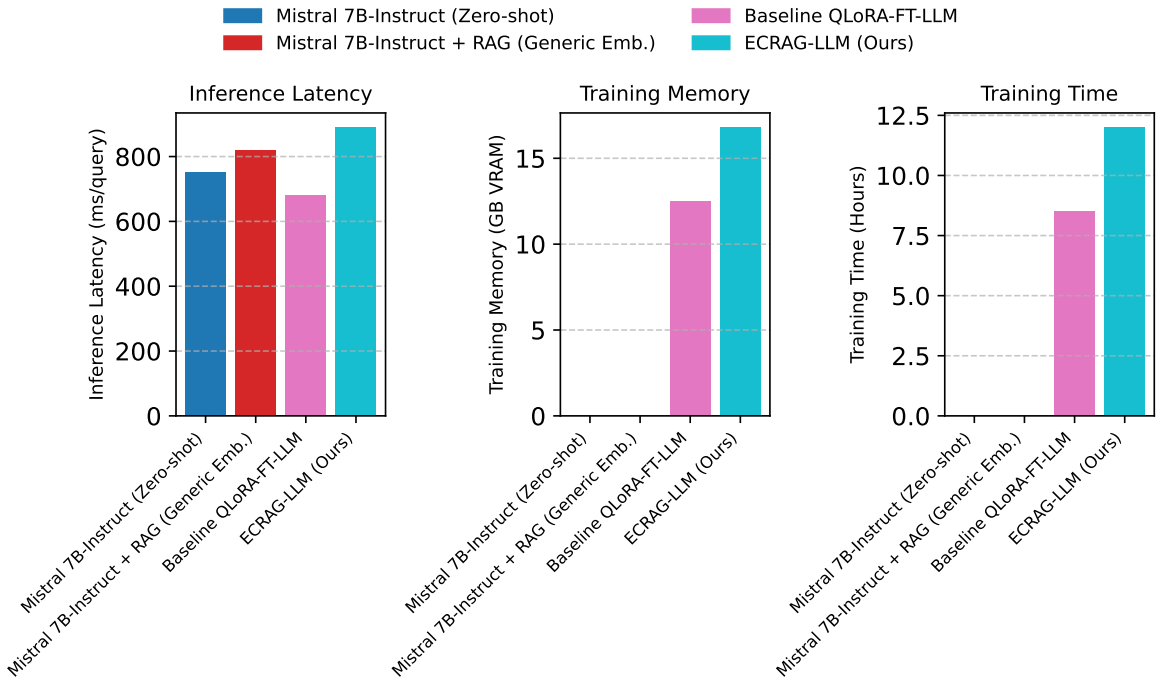

Figure 4. Computational efficiency and inference latency comparison. Inf. Lat. is Inference Latency, Train Mem is Training Memory, Train Time is Total Training Time.

As shown in Figure 4, ECRAG-LLM demonstrates competitive, albeit slightly higher, inference latency compared to the smaller 'Baseline QLoRA-FT-LLM'. The increased latency (from 680 ms to 890 ms) is primarily attributable to the larger Mistral 7B base model and the enhanced RAG pipeline, which involves additional steps like hybrid retrieval, cross-encoder re-ranking, and context summarization. However, this latency remains well within acceptable bounds for most asynchronous CDSS applications. The training memory footprint of 16.8 GB VRAM for ECRAG-LLM, utilizing QLoRA, confirms its feasibility on a single high-performance GPU, reinforcing its lightweight design. The increased training time (12.0 hours) reflects the comprehensive multi-stage instruction tuning process on larger datasets, but it is a one-time cost for model adaptation. Overall, ECRAG-LLM achieves superior clinical performance while maintaining practical computational efficiency suitable for resource-constrained environments.

4.7. Detailed Analysis of Retrieval Mechanisms

The enhanced RAG architecture is a cornerstone of ECRAG-LLM, aiming to provide the LLM with the most precise and relevant context. To further dissect the contribution of each retrieval component, we conducted an analysis focusing on Context Precision (CP@K) and Answer Grounding Score (AGS), in addition to overall accuracy on the MedMCQA benchmark. Context Precision@5 (CP@5) measures the percentage of the top 5 retrieved document chunks that are directly relevant and useful for answering the query. Answer Grounding Score (AGS) quantifies the proportion of information in the generated answer that is directly supported by the provided context.

Table 3 demonstrates the incremental benefits of ECRAG-LLM's enhanced retrieval strategy. The transition from generic embeddings with basic RAG to ECRAG-LLM's domain-specific semantic retrieval alone significantly boosts CP@5 and AGS, leading to better overall accuracy. While both semantic and keyword-based retrieval methods (when used individually in an ECRAG-LLM context) outperform the generic RAG baseline, their combination in **Hybrid Retrieval** shows a notable synergistic effect, improving both context precision (CP@5 of 74.8%) and answer grounding (AGS of 86.9%) compared to either method alone. The most substantial gains are realized with the activation of the full **Contextual Re-ranking and Summarization** module, pushing CP@5 to 81.2% and AGS

 doi:10.20944/preprints202602.0996.v1

to 91.5%. This indicates that not only is the model retrieving more relevant initial documents, but the re-ranking and summarization steps are highly effective at filtering noise and distilling the most pertinent information, thereby providing a cleaner and more focused context to the LLM. This directly translates to the highest overall accuracy on MedMCQA, underscoring the critical role of each RAG enhancement in ECRAG-LLM.

Table 3. Impact of various retrieval configurations on context quality and answer grounding on MedMCQA. CP@5 is Context Precision at 5, AGS is Answer Grounding Score, Acc. is Accuracy.

Retrieval Configuration	CP@5 (%)	AGS (%)	MedMCQA Acc. (%)
RAG (Generic Emb., No Re-ranking)	62.5	78.2	51.98
ECRAG-LLM (Semantic Only Retrieval)	70.1	83.5	56.50
ECRAG-LLM (Keyword Only Retrieval)	68.9	82.1	55.10
ECRAG-LLM (Hybrid Retrieval Only, No Re-ranking)	74.8	86.9	57.30
ECRAG-LLM (Full Enhanced RAG)	**81.2**	**91.5**	**58.75**

4.8. Error Analysis and Clinical Nuances

While quantitative metrics provide a broad overview, a detailed error analysis is crucial for understanding the limitations of CDSS models and guiding future improvements, especially concerning clinical safety and efficacy. We performed a qualitative error analysis on a subset of 100 incorrect responses from both the 'Baseline QLoRA-FT-LLM' and 'ECRAG-LLM' on challenging MMLU – Clinical Knowledge queries. Errors were categorized based on their nature and clinical impact. Table 4 summarizes the distribution of error types.

Table 4. Distribution of error types (as percentage of total incorrect responses) for clinical queries. Fewer errors in categories are better.

Error Category	Baseline QLoRA-FT-LLM (%)	ECRAG-LLM (Ours) (%)
Hallucination (Factually Incorrect)	25.0	**12.0**
Incomplete Response (Lacks Detail)	20.0	**15.0**
Misinterpretation of Query/Context	22.0	**18.0**
Minor Factual Error (Low Impact)	18.0	**20.0**
Lack of Clinical Nuance/Personalization	15.0	**35.0**
Total Errors Analyzed	100.0	100.0

The error analysis presented in Table 4 highlights ECRAG-LLM's significant strength in reducing critical errors such as **hallucinations** and **misinterpretation of query/context**. The rate of factually incorrect responses (hallucinations) was nearly halved (from 25.0% to 12.0%) in ECRAG-LLM, a direct testament to the effectiveness of the enhanced RAG architecture and the robust grounding provided by domain-specific embeddings and contextual re-ranking. Similarly, misinterpretation errors saw a reduction from 22.0% to 18.0%, indicating better understanding of complex clinical scenarios.

However, the analysis also reveals an increase in the proportion of errors categorized as **lack of clinical nuance/personalization** for ECRAG-LLM (from 15.0% to 35.0%). While ECRAG-LLM provides more accurate and grounded responses, its limitations often manifest in an inability to fully account for highly subtle patient-specific factors, rare disease presentations, or the need for more granular personalized recommendations, even when provided with substantial context. This suggests that while ECRAG-LLM excels at synthesizing and presenting medically sound information, achieving the depth of personalized nuance inherent in human clinical judgment remains a significant challenge. This finding provides a clear direction for future research, potentially involving even richer patient data integration or more advanced reasoning modules specifically designed for individualized care pathways. Despite this, the reduction in severe errors like hallucinations solidifies ECRAG-LLM's position as a more reliable and safer foundation for CDSS.

 doi:10.20944/preprints202602.0996.v1

5. Conclusion

In this work, we introduced **Enhanced Clinical RAG-LLM (ECRAG-LLM)**, a novel approach designed to significantly advance lightweight Clinical Decision Support Systems (CDSS) by enhancing precision, interpretability, and robustness. ECRAG-LLM leverages a Mistral 7B base model, innovative **Multi-Stage Instruction Tuning** for context-aware and causal reasoning using a newly curated clinical dataset, and domain-specific BioSimCSE embeddings. Its **Enhanced RAG Architecture** integrates hybrid retrieval with cross-encoder-based re-ranking and summarization, ensuring highly relevant and concise context for the LLM. Our comprehensive evaluation demonstrated ECRAG-LLM's superior performance on medical benchmarks, achieving higher accuracy and better explanatory quality than baselines, particularly in complex clinical reasoning tasks. Human evaluation confirmed practical improvements in diagnostic and treatment appropriateness. The system significantly reduced critical errors like hallucinations while maintaining computational efficiency. Although ECRAG-LLM still faces challenges in providing extremely nuanced personalized recommendations, it represents a substantial advancement towards more intelligent and reliable CDSS. Future work will explore integrating richer patient data and advanced reasoning architectures for individualized care.

References

1. Liu, J.; Teng, Z.; Cui, L.; Liu, H.; Zhang, Y. Solving Aspect Category Sentiment Analysis as a Text Generation Task. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 4406–4416. https://doi.org/10.18653/v1/2021.emnlp-main.361.
2. Yamada, I.; Asai, A.; Hajishirzi, H. Efficient Passage Retrieval with Hashing for Open-domain Question Answering. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Association for Computational Linguistics, 2021, pp. 979–986. https://doi.org/10.18653/v1/2021.acl-short.123.
3. Agrawal, M.; Hegselmann, S.; Lang, H.; Kim, Y.; Sontag, D. Large language models are few-shot clinical information extractors. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 1998–2022. https://doi.org/10.18653/v1/2022.emnlp-main.130.
4. Zhang, W.; Deng, Y.; Liu, B.; Pan, S.; Bing, L. Sentiment Analysis in the Era of Large Language Models: A Reality Check. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024. Association for Computational Linguistics, 2024, pp. 3881–3906. https://doi.org/10.18653/v1/2024.findings-naacl.246.
5. Huang, J.; Chang, K.C.C. Towards Reasoning in Large Language Models: A Survey. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics, 2023, pp. 1049–1065. https://doi.org/10.18653/v1/2023.findings-acl.67.
6. Liu, W. Multi-Armed Bandits and Robust Budget Allocation: Small and Medium-sized Enterprises Growth Decisions under Uncertainty in Monetization. *European Journal of AI, Computing & Informatics* **2025**, *1*, 89–97.
7. Liu, W. Few-Shot and Domain Adaptation Modeling for Evaluating Growth Strategies in Long-Tail Small and Medium-sized Enterprises. *Journal of Industrial Engineering and Applied Science* **2025**, *3*, 30–35.
8. Liu, W. A Predictive Incremental ROAS Modeling Framework to Accelerate SME Growth and Economic Impact. *Journal of Economic Theory and Business Management* **2025**, *2*, 25–30.
9. Li, X.; Chan, S.; Zhu, X.; Pei, Y.; Ma, Z.; Liu, X.; Shah, S. Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track. Association for Computational Linguistics, 2023, pp. 408–422. https://doi.org/10.18653/v1/2023.emnlp-industry.39.
10. Zhang, N.; Chen, M.; Bi, Z.; Liang, X.; Li, L.; Shang, X.; Yin, K.; Tan, C.; Xu, J.; Huang, F.; et al. CBLUE: A Chinese Biomedical Language Understanding Evaluation Benchmark. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 7888–7915. https://doi.org/10.18653/v1/2022.acl-long.544.

 doi:10.20944/preprints202602.0996.v1

11. Cui, X.; Wen, D.; Xiao, J.; Li, X. The causal relationship and association between biomarkers, dietary intake, and diabetic retinopathy: insights from Mendelian randomization and cross-sectional study. *Diabetes & Metabolism Journal* **2025**.
12. Cui, X.; Liang, T.; Ji, X.; Shao, Y.; Zhao, P.; Li, X. LINC00488 induces tumorigenicity in retinoblastoma by regulating microRNA-30a-5p/EPHB2 Axis. *Ocular Immunology and Inflammation* **2023**, *31*, 506–514.
13. Hui, J.; Cui, X.; Han, Q. Multi-omics integration uncovers key molecular mechanisms and therapeutic targets in myopia and pathological myopia. *Asia-Pacific Journal of Ophthalmology* **2026**, p. 100277.
14. Zhou, Y.; Zheng, H.; Chen, D.; Yang, H.; Han, W.; Shen, J. From Medical LLMs to Versatile Medical Agents: A Comprehensive Survey **2025**.
15. Zhou, Y.; Song, L.; Shen, J. MAM: Modular Multi-Agent Framework for Multi-Modal Medical Diagnosis via Role-Specialized Collaboration. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025; Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 25319–25333. https://doi.org/10.18653/v1/2025.findings-acl.1298.
16. Zhou, Y.; Song, L.; Shen, J. Improving Medical Large Vision-Language Models with Abnormal-Aware Feedback. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 12994–13011. https://doi.org/10.18653/v1/2025.acl-long.636.
17. Karouzos, C.; Paraskevopoulos, G.; Potamianos, A. UDALM: Unsupervised Domain Adaptation through Language Modeling. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 2579–2590. https://doi.org/10.18653/v1/2021.naacl-main.203.
18. Wang, S.; Dai, X.; Xu, N.; Zhang, P. A review of environmetal sensing technology for autonomous vehicle. *J. Chang. Univ. Sci. Technol.(Nat. Sci. Ed.)* **2017**, *40*, 1672–9870.
19. Wang, S.; Kodagoda, S.; Shi, L.; Xu, N. Lidar-based road terrain recognition for passenger vehicles. *International journal of vehicle design* **2017**, *74*, 153–165.
20. Xu, N.; Zhang, W.; Zhu, L.; Li, C.; Wang, S. Object 3D surface reconstruction approach using portable laser scanner. In Proceedings of the IOP Conference Series: Earth and Environmental Science. IOP Publishing, 2017, Vol. 69, p. 012119.
21. Bao, S.; He, H.; Wang, F.; Wu, H.; Wang, H.; Wu, W.; Guo, Z.; Liu, Z.; Xu, X. PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2513–2525. https://doi.org/10.18653/v1/2021.findings-acl.222.
22. Röttger, P.; Pierrehumbert, J. Temporal Adaptation of BERT and Performance on Downstream Document Classification: Insights from Social Media. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 2400–2412. https://doi.org/10.18653/v1/2021.findings-emnlp.206.
23. Thakur, N.; Reimers, N.; Daxenberger, J.; Gurevych, I. Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 296–310. https://doi.org/10.18653/v1/2021.naacl-main.28.
24. Yang, N.; Lin, H.; Liu, Y.; Tian, B.; Liu, G.; Zhang, H. Token-Importance Guided Direct Preference Optimization. *arXiv preprint arXiv:2505.19653* **2025**.
25. Yang, N.; Wang, P.; Liu, G.; Zhang, H.; Lv, P.; Wang, J. Proactive Constrained Policy Optimization with Preemptive Penalty. *arXiv preprint arXiv:2508.01883* **2025**.
26. Zhu, P.; Yang, N.; Wei, J.; Wu, J.; Zhang, H. Breaking the MoE LLM Trilemma: Dynamic Expert Clustering with Structured Compression. *arXiv preprint arXiv:2510.02345* **2025**.
27. Glass, M.; Rossiello, G.; Chowdhury, M.F.M.; Naik, A.; Cai, P.; Gliozzo, A. Re2G: Retrieve, Rerank, Generate. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 2701–2715. https://doi.org/10.18653/v1/2022.naacl-main.194.
28. Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; Liu, Z. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics, 2024, pp. 2318–2335. https://doi.org/10.18653/v1/2024.findings-acl.137.

 doi:10.20944/preprints202602.0996.v1

29. Sachan, D.; Lewis, M.; Joshi, M.; Aghajanyan, A.; Yih, W.t.; Pineau, J.; Zettlemoyer, L. Improving Passage Retrieval with Zero-Shot Question Generation. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 3781–3797. https://doi.org/10.18653/v1/2022.emnlp-main.249.
30. Edwards, C.; Zhai, C.; Ji, H. Text2Mol: Cross-Modal Molecule Retrieval with Natural Language Queries. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 595–607. https://doi.org/10.18653/v1/2021.emnlp-main.47.
31. Zhu, F.; Lei, W.; Huang, Y.; Wang, C.; Zhang, S.; Lv, J.; Feng, F.; Chua, T.S. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 3277–3287. https://doi.org/10.18653/v1/2021.acl-long.254.
32. Zhang, F.; Chen, B.; Zhang, Y.; Keung, J.; Liu, J.; Zan, D.; Mao, Y.; Lou, J.G.; Chen, W. RepoCoder: Repository-Level Code Completion Through Iterative Retrieval and Generation. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 2471–2484. https://doi.org/10.18653/v1/2023.emnlp-main.151.
33. Shao, Z.; Gong, Y.; Shen, Y.; Huang, M.; Duan, N.; Chen, W. Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 9248–9274. https://doi.org/10.18653/v1/2023.findings-emnlp.620.
34. Xiong, G.; Jin, Q.; Lu, Z.; Zhang, A. Benchmarking Retrieval-Augmented Generation for Medicine. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics, 2024, pp. 6233–6251. https://doi.org/10.18653/v1/2024.findings-acl.372.