# Advancing Healthcare with Retrieval Augmented Generation (RAG) and Large Context Models (LCM): A Comparative Study

**Dr. Rashmi Sharma**[1] and **Dr. Himani Garg**[2]
[1]Department of Information Technology, [2] Department of ECE
[1,2] Ajay Kumar Garg Engineering College, Ghaziabad, U.P. India 201019
[1]sharmarashmi@akgec.ac.in, [2]garghimani@akgec.ac.in

*Abstract* -- **Advancements in Artificial intelligence (AI) have changed the face of healthcare through innovative technology that enhance decision-making, medical research and operational efficiency. It is experiencing a paradigm shift as new generations of AI are changing the ways of diagnosis, treatment, and patient care. Of these technologies, Retrieval Augmented Generation (RAG) and Large Context Models (LCMs) appear to be particularly effective. One of the most revolutionary capabilities of RAG could provide unique, highly effective perspectives for solving varied, profound healthcare issues. LCMs stand out for their ability to analyze extensive spans of interrelated data while maintaining coherence and depth. This research paper reviews these two paradigms; proposing their central approaches, comparisons, and relevance to the health context. This work provides a comparative analysis and demonstrates the possibilities of these paradigms through specific health care applications and case studies; stressing the potential of paradigms in diagnostics, individualized medicine and patient's training.**

*Keywords: Healthcare, Retrieval-Augmented Generation (RAG), Large Context Models (LCMs), Contextual Queries, Vector database*

## I.. INTRODUCTION

THE healthcare industry increasingly relies on AI technologies to handle and analyse complex data, facilitate diagnostics, and to optimize patient outcomes. Retrieval Augmented Generation (RAG) and Large Context Models (LCMs) represent two cutting-edge approaches with distinct yet complementary capabilities:

- **RAG** combines retrieval mechanisms with reasoning to generate well-founded arguments or insights based on external data sources.
- **LCMs** leverage vast training on diverse and large volume datasets to understand, evaluate and respond to intricate contextual queries, offering a wider perspective in scenarios.

Understanding these technologies and their applicability to healthcare is crucial for advancing AI-driven medical practices.

*Retrieval Augmented Generation (RAG)***:** RAG amalgamates information retrieval and augmented generation, integrating a retrieval module with a generation model. The retrieval component query external databases or knowledge sources to fetch relevant data, while the generation module synthesizes coherent and reasoned outputs[1]. Combines the power of LLMs with efficient information retrieval techniques, like vector or lexical search, to provide answers derived from existing knowledge bases or datasets. This approach is useful if the answer to a question is in the data and then extract useful information which is then analyzed by an LLM. This gives LLMs access to information that would not fit in their context window [4].

For example, the question «What was the interest rate decision of February 2024?» is one that an LLM most likely won't have the answer to on its knowledge base. By having access to a large database of recent articles, where one might contain a paragraph describing the event, it's possible to search for keywords and present only the relevant articles or even paragraphs and feed it as context to the LLM, enabling it to provide the correct answer.

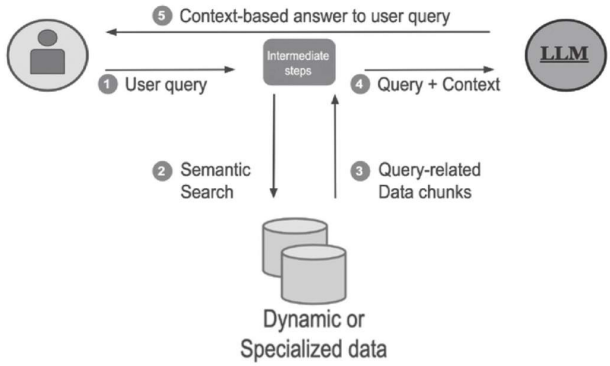


Figure 1. Simplified view of RAG

The RAG model has the following steps:

1. A user enters a query in natural language. This query is represented in vector form with the help of embedding models.

2. The resulting query embedding is then used to retrieve related or similar chunks of data within the vector database. Sometimes hybrid search is more effective.
3. Top k query-related data chunks are selected for downstream usage.
4. The prompt is generated using user query + linked data chunks and then input in LLM.
5. Context-based answers are returned to the user which are accurate and precise compared to default LLM answers on knowledge-intensive tasks.

*Key Features of RAG:*

- Data Integration: Incorporates data represented in a structured or unstructured manner, and updated in real time.
- Explainability: Offers reasons so that generated explanations are easier to comprehend.
- Focus on Specificity**:** Outperforms in specific focused domains needed to provide the best evidence-based rationale for results.

*Large Context Models (LCMs)*

Most LCMs are simply variations of LLMs which are designed to have more context so that the model can take queries and respond with greater context. LCMs are advanced NLP instruments intended to understand and produce textual data within large contexts. In fact, while traditional models are not capable of analyzing contextual data, the LCMs can handle complex dependencies between terms across large documents, multiple streams of data, or turns in a conversation. This capability makes them particularly well-suited for complex domains like healthcare, where vast amounts of interconnected data must be synthesized for accurate decision-making.

*Key Features of LCM:*

- Contextual Awareness**:** Handles large input and remains coherent at large contexts.
- Flexibility**:** Serves as a summarizer, question-answering tool, and a knowledgeable decision maker.
- Generalizability**:** Adaptable to multiple domains without requiring fine-tuned datasets.

## II.. ENHANCING HEALTHCARE INDUSTRY WITH RAG & LCM

*RAG enriching healthcare application*

Clinical Decision Support (CDS)**:** By applying AI capability, RAG can offer clinicians with the evidence-based practice advice in real-time fashion. For example, a doctor concerning their patient with unusual signs and symptoms that are difficult to diagnose consults with a RAG system which provides updated and most relevant information from journals and protocols and other similar cases.

Personalized Patient Care**:** In this case, RAG can be combined with electronic health records (EHRs) and generate individualised care programmes. For example, a RAG system could take a patient's medical history, their medications, and their genomics results and propose treatments or lifestyle changes for them [3].

Medical Research Acceleration**:** By using RAG, the researchers are able to retrieve and merge the findings gathered from a massive amount of literature within a short amount of time. This capability is especially beneficial when meta-analyses are needed by postdoctoral researchers or when new hypotheses are being considered as RAG systems allow for mapping of the current state and identification of knowledge gaps.

Health Literacy and Patient Communication**:** Health care is a profession that requires the use of appropriate communication with the patients. RAG systems can provide plain language explanations for terms used in medical practice or for certain diseases, which will in turn help the patient to make an informed decision in their case.

Drug Discovery and Pharmacovigilance**:** RAG can be instrumental in helping pharmaceutical firms find new drug targets, based on large data sets of metadata including, but not limited to: the published literature on biology, patient genetics and clinical trials. Second, RAG can be useful for pharmacovigilance with reporting adverse drug effects in real time.

TABLE 1--COMPARATIVE ANALYSIS OF RAG & LCM

| Aspect | Retrieval Augmentation Generation | Large Context Models |
|---|---|---|
| Data Dependency | Relies on external retrieval sources. | Processes contextual data internally from pre-training. |
| Explainability | Provides evidence-based arguments. | Offers broad contextual insights without explicit evidence. |
| Domain Specialization | Well-suited for narrow, evidence-heavy tasks. | General-purpose with wide-ranging adaptability. |
| Real-Time Use | Effective for real-time, evidence-required decision-making. | May lag in real-time if extensive pre-training isn't domain-specific. |
| Scalability | Limited by retrieval source scope. | Scales well across varied, large datasets. |

Telemedicine Support**:** In the remote healthcare environment, the proposed RAG-powered systems can assist healthcare workers in retrieving contextually appropriate instructions, and thus provide accurate and effective virtual consultations.

*LCM Enriching Healthcare Application*

- Comprehensive Clinical Decision-Making**:** LCMs can analyze entire patient records (diagnoses, treatment plans, progress notes etc.), LCMs are capable of providing context-aware recommendations. With the help of this integration, LCMs facilitate broader and, therefore, more accurate decisions made by clinicians [2].
- Enhanced Medical Literature Review**:** LCMs can analyze and synthesise a large amount of research papers, compare and contrast the findings, reveal the uncovered contradictions and come up with ideas for further research.
- Streamlined EHR Interactions**:** Integration with electronic health records (EHRs) allows LCMs to understand extensive, narrative entries in doctor's narratives. This capability enables physicians to query patient information using conversational prompts and receive precise, contextually relevant responses.
- Intelligent Medical Dialogue Systems**:** Telemedicine services use LCMs suitable for computing and understanding conversation length and multi destination conversations. Such models are useful to fill gaps in communication by holding information and providing reliable information throughout virtual consultations.
- Drug Discovery and Clinical Trials**:** LCMs can integrate sources of information such as trial protocols, responses from patients, regulatory documents and find new treatment approaches and forecast trial results.
- Cross-Document Inference for Rare Diseases**:** In diagnosing rare diseases, clinicians rely on some kind of integration of local and fragmented evidence. LCMs are best employed in cross-document reasoning, making it easier for clinicians to establish relationships between records, case files, and genomic info sets.
- Health Policy Analysis and Planning**:** Due to their transparency and being updated frequently, LCMs can be used by governments and healthcare organizations to review policy documents, demographic stats, and past performance of particular policies in order to create new policies that are more effective.

## III. PROPOSED FRAMEWORK USING RAG IN HEALTHCARE

Healthcare data has become fragmented and exists in silos for a number of reasons. However, many helpful and favorable laws for patients like HIPAA and GDPR only restrict the flow of data and hinder the gains made from its analysis [1]. The process is full of challenges resulting in a lot of wastes and time consumption.

The unfortunate fact is around 80% of healthcare data is unstructured like notes written by doctors, nurses, and other healthcare providers, documenting patient encounters, observations, medical reasoning, clinical trial protocols, eligibility criteria, research papers, email communications, etc [4].

Most of the data is unstructured, siloed, proprietary, or protected with regulations, and thus difficult to bake in model parameters, making a near-perfect problem space for deploying RAG-powered applications.

*Build RAG Model for Healthcare*: To meet the privacy requirement of the patient data, diagnostic history, the proposed paper uses an on-premises database and in-house LLM hosting, ensuring no data exits the organization's environment. The framework selects robust open-source models already available. These models are actually slightly behind in performance compared to proprietary ones, typically catching up within a 6-12 month timeframe, offering increased control and transparency. Cost considerations are crucial as well. Since pricing is token-based, the expenses associated with paid models can accumulate rapidly, potentially making them a costly option. However, if the focus is on ease of deployment and immediate access to state-of-the-art (SoTA) results, vendors like OpenAI and Anthropic are more suitable choices [6].

RAG can empower each stakeholder from bench to bedside in clinical workflow, including the patient.

The researchers can query through proprietary on-premise databases within their organization, or a set of publications within a specific area, reducing the time required significantly to gain knowledge or validate research ideas

Real-world data explorers trying to query free text data to get specific answers, like deriving a line of therapy from a patient's medical history notes

Clinical trial matching or eligibility validation for patients using the inclusion and exclusion criteria of the trial with patient's medical records

The physician-assistant-software can have the ability to comb through large chunks of medical history or clinical guidelines (*e.g.* NCCN) to find relevant bits of information in seconds

The patient could be a real winner here. A lot of countries have laws enabling patients to access a ton of medical information associated with them. However, considering the scale and format of these data dumps, most patients might not be able

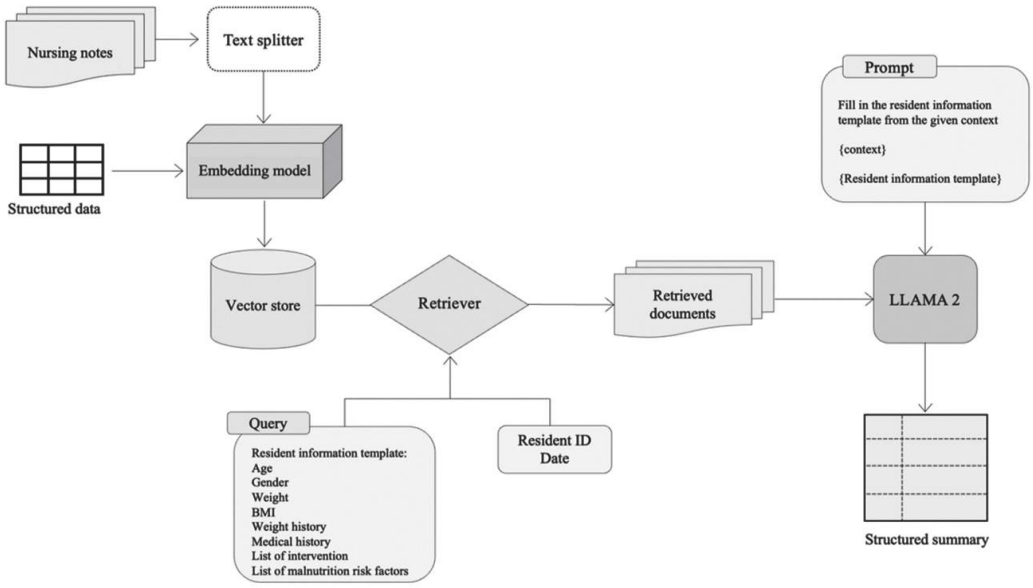


Figure 2. Proposed RAG Approach.

to make sense of it. With the power of RAG, they would have the ability to instantly find specific information from their own lab reports, other medical history, prescriptions, or even on the RCT side exploring the clinical trial protocol, informed consent form, etc leading to reduced information asymmetry we have been seeing for decades [5].

*Why does it make sense to use RAG?* : Cancer Care is always changing, and evolving. So based on the latest evidence NCCN develops cancer care recommendations covering 97% of the cancer types, used by healthcare providers worldwide. The problem in accessing these guidelines is twofold, high information density unstructured data and the pace at which it changes. There are several dozen types of cancers and these guidelines are updated almost yearly or so.

- There are around 65 types of cancer listed on NCCN (*e.g.* NSCLC, ALL, etc)
- Each type has a separate guideline document depending on the stage (*e.g.* early, local, metastasized, etc)
- And then each would have a lengthy guideline document approx 80 pages for early NSCLC

So to address this, our proposed framework must have

- Ability to extract, summarize, and cite information from free text
- For non-technical users like physicians, oncologists
- Deal with the dynamic change in the underlying guidelines (on average 3 per month)
- Be precise with answers, without hallucinations

An LLM as is, can address the first two. An LLM capable of talking with the internet could address the first three. But only a RAG-like system is more suited for all four.A RAG makes more sense to solve this dynamic data problem. Note, that the application of RAG is not limited to sensitive or private data.

## IV. CHALLENGES AND FUTURE DIRECTIONS

Both technologies face challenges, including data biases, integration complexities, and the need for robust validation in critical applications like healthcare.

- Data Privacy and Security: Ensuring compliance with regulations and data confidentiality such as HIPAA and GDPR is essential when handling sensitive medical data.
- Bias and Misinformation**:** RAG systems must be trained on diverse and verified datasets to minimize biases and avoid generating misleading information.
- Integration with Existing Systems**:** Seamlessly integrating RAG with EHRs and other healthcare systems requires robust infrastructure.
- Computational Demands: LCMs require significant computational resources, necessitating robust infrastructure.
- Bias in Data: Training LCMs on diverse, high-quality datasets is crucial to avoid biases that may compromise patient safety.

*Future Prospects***:**

- RAG**:** Improvements in retrieval algorithms and integration with domain-specific databases for enhanced precision.

- LCM**:** Continued fine-tuning for healthcare-specific applications and better handling of domain-specific jargon.

## V. CONCLUSION

RAG and LCM technologies offer unique and complementary capabilities that can transform healthcare. By enabling accurate, personalized, and efficient information retrieval, RAG can significantly enhance patient outcomes, streamline clinical workflows and accelerate medical research. While RAG is ideal for tasks requiring evidence-based insights, LCM excels in providing holistic, contextual analysis. Together, they have the potential to advance diagnostic accuracy, optimize treatment strategies, and improve patient education, ultimately driving more intelligent and compassionate healthcare systems.

## REFERENCES

[1] Mohammad Alkhalaf, Ping Yu, Mengyang Yin and Chao Deng, "Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records," *Journal of Biomedical Informatics*, Volume 156, 2024, 104662, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2024.104662.

[2] Neel Bhate, Ansh Mittal, The He, Xiao Luo, "Zero-shot Learning with Minimum Instruction to Extract Social Determinants and Family History from Clinical Notes using GPT Model," p. arXiv: 2309.05475doi: DOI: 10.48550/arXiv.2309.05475.

[3] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, Robert McHardy, "Challenges and applications of large language models," p. arXiv:2307.10169doi: DOI: 10.48550/arXiv.2307.10169.

[4] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, Pascale Fung, "A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity," p. arXiv:2302.04023doi: DOI: 10.48550/arXiv.2302.04023.

[5] V. Rawte, A. Sheth and A. Das, "A survey of hallucination in large foundation models," arXiv preprint arXiv:2309.05922, 2023.

[6] H. Alkaissi and S. I. McFarlane, "Artificial hallucinations in ChatGPT: implications in scientific writing," *Cureus*, vol. 15, no. 2, 2023.

[7] Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Sun Wang, Yucheng Xu, Hong Yu, "NoteChat: A dataset of synthetic doctor-patient conversations conditioned on clinical notes," p. arXiv:2310.15959doi: DOI: 10.48550/arXiv.2310.15959.

**Dr. Rashmi Sharma**, Ph.D. (Science - Computers) , M.Tech. in Computer Science & Engineering, Master of Computer Applications is Associate Professor in Ajay Kumar Garg Engineering College, Ghaziabad, UP India..She has more than 15 years of teaching and five years of industrial experience. She has authored 12 books, is Sun Certified Java Programmer (SCJP 2.0) in 2001, and Business English Certification (BEC) from Cambridge University in 2001. Her current research area includes Block-chain Technology, Computer Vision, Sensor IoT, Wireless Sensor Networks, Machine Learning, Data Analytics, Smart appliances. She has published more than 30 referred publications in SCI/ESCI/ Indexed journals and conferences.. She has one Granted, one Copyright, and eight published patents in her account. She is a member of many technical associations - IEEE, CSTA, IAENG, and ISTE.

**Dr. Himani Garg** received her Doctorate Degree from NIT, Kurukshetra in 2017 in the field of Signal Processing and is currently working as Professor in Ajay Kumar Garg Engineering College, Ghaziabad. She has published a number of papers in preferred Journals and chapters in books, and participated in a range of forums. She has also presented research-based papers at several national and international conferences. She has completed various research and consultancy projects and has also received the Best Teacher Award from State University. Her research interests include Wireless Sensor networks, Communication Systems, Signal Processing and Machine Learning.