The Role of AI, Generative AI, AI Agents, RAG in Advancing Affordable Healthcare Systems

Arohi Patel
Newsome High School,
16550 Fishhawk Blvd,
Lithia, FL 33547, United States
arohi0patel@gmail.com

Publication History
Manuscript Reference No: IJIRAE/RS/Vol.12/Issue06/JNAE10083
Research Article | Open Access | Double-Blind Peer-Reviewed | Article ID: IJIRAE/RS/Vol.12/Issue06/JNAE10083
Received: 20, May 2025, Revised: 01, June 2025, Accepted: 09, June 2025, Published Online: 20, June 2025.
https://www.ijirae.com/volumes/Vol12/iss-06/02.JNAE10083.pdf
Article Citation: Arohi(2025),The Role of AI, Generative AI,AI Agents, and RAG in Advancing Affordable Healthcare Systems. IJIRAE: International Journal of Innovative Research in Advanced Engineering, Volume 12, Issue 06 of 2025 pages 292-300
Doi:> https://doi.org/10.26562/ijirae.2025.v1206.02
BibTeX Key: Arohi@2025AI

Abstract: Artificial intelligence (AI) is the replication of human intelligence in machines that are programmed to think, learn, and complete tasks independently. These computers replicate human characteristics such as reasoning, problem-solving, and decision-making, relying on powerful algorithms and massive amounts of data to function efficiently. Nowadays, artificial intelligence plays an important role in today's business organization transformation. The term artificial intelligence refers to computer systems able to perform normally requiring human intelligence. While receiving of Artificial intelligence is not a new concept in human society, it is a new event in society. Artificial intelligence came as an entity in 1956. But it took a long time of work to make significant progress towards developing an artificial intelligence system. Artificial intelligence can be useful in business in different ways. In business, Artificial Intelligence has a huge range of utilizations. In fact, most of us interact with artificial intelligence in some forms on a regular basis. From the normal to the breathtaking, Artificial Intelligence is already disrupting virtually every business procedure in every industry. An artificial intelligence (AI) agent is a software program that can interact with its surroundings collect data and use that data to perform self-directed tasks that meet predetermined goals. Humans set goals, but an AI agent independently chooses the best actions it needs to perform to achieve those goals. Generative AI is a type of artificial intelligence designed to create new content such as text, images, music or even code by learning patterns from existing data. These models generate original outputs that are often indistinguishable from human-created content. These models use techniques like deep learning and neural networks to generate output. As artificial intelligence technologies rapidly increase in number, they are becoming imperative for businesses that want to maintain a competitive move. While AI is transforming industries with advanced solutions, tools like the Paraphrasing tool can enhance productivity in writing by providing free, full-length reinterpretations without signup hassles. This seamless integration of technology aligns with the evolving landscape influenced by AI, adding significant educational value. 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. The present paper shows that what is artificial intelligence & how it is useful for business.

Keywords: Artificial Intelligence, Speech Recognition, Technology, Decision Making

1. INTRODUCTION

In the Historical way; Artificial Intelligence came first; AI is the Fundamental requirement by which; Machine Learning developed and Machine learning is the Fundamental need by which Deep Learning evolves. Detailed explanations of all three terms are as follows:

- Artificial Intelligence

Artificial Intelligence can be referred to as anything that developing Intelligent Machines through programming. for example useful machines on daily basis i.e Smartphone's, software relating to marketing. In broad terms, we can define AI as a programmed machine that responds as per well-defined rules and responses.

- Machine Learning

Machine learning is a subset of AI. It refers to the scientific terms in which machines that incorporated with the data and observed it and learns from it and improve with experience. Machine learning is primarily won't process huge amounts of data quickly. These sorts of AI are algorithms that appear to "learn" over time, recuperating at what they are doing the more often they are doing it. Feed a machine learning algorithm more data and its modelling will be improved.

Machine learning is useful for putting a big amount of data increasingly captured by connected devices and the internet of things into a digestible context for human beings. For example, if a person manages a manufacturing plant, their machinery is likely connected to the network. Connected devices feed a continuing stream of information about production and more to a central location. Machine learning can rapidly analyze the data as it comes in, identifying methods and anomalies. If a machine within the plant is functioning at a reduced capacity, A Machine Learning algorithm can catch it and notify decision-makers about the reduced capacity that it is time to dispatch a preventive maintenance team. But Machine Learning is additionally a comparatively broad category of Deep Learning. Artificial Intelligence, Machine Learning, and Deep Learning are interrelated terms that can be easily understood by the following diagram:

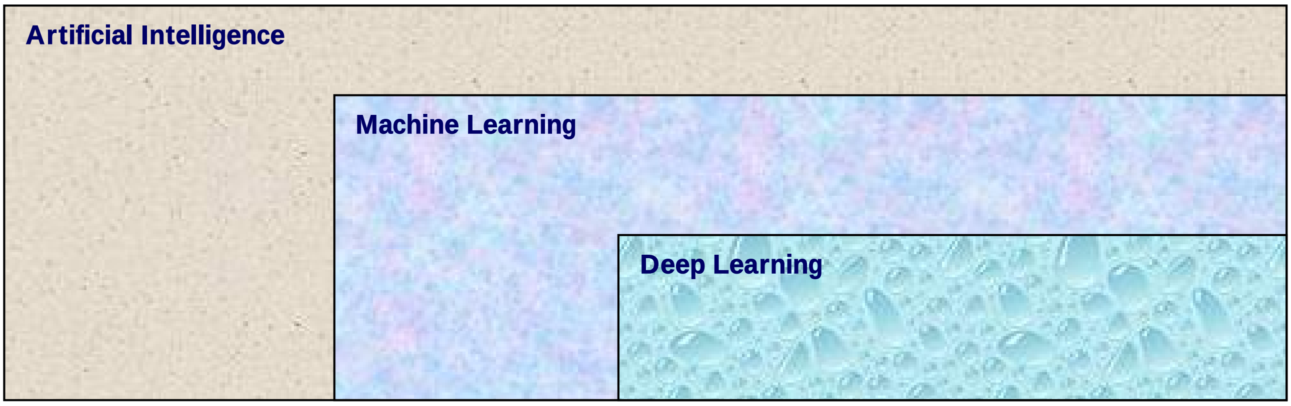

Fig 1. The difference between Artificial Intelligence, Machine Learning, and Deep Learning

Components of AI

AI encompasses various domains:

- Machine Learning (ML): Enables systems to learn from data and improve performance without explicit programming.
- Computer Vision: Empowers machines to interpret and process visual data from the world.
- Robotics: Focuses on designing intelligent robots capable of performing complex physical tasks.
- Natural Language Processing (NLP): Facilitates interactions between humans and machines through language understanding.

II. AGENTS IN AI

An artificial intelligence (AI) agent is a software program that can interact with its environment, collect data and use that data to perform self-directed tasks that meet predetermined goals. Humans set goals, but an AI agent independently chooses the best actions it needs to perform to achieve those goals.

For example, consider a customer support AI agent that wants to resolve customer queries. The agent will automatically ask the customer different questions, look up information in internal documents and respond with a solution. Based on the customer responses, it determines if it can resolve the query itself or pass it on to a human.

Key Features of AI Agents

- Autonomous: Act without constant human input and decide next steps from past data like a bookstore bot flags missing invoices.
- Goal- driven: Optimize for defined objectives like a logistics AI balancing speed, cost and fuel use.
- Perceptive: Gather info from sensors, inputs or APIs like a cyber security agent tracking new threats.
- Adaptable: Adjust strategies when situations change.
- Collaborative: Work with humans or other agents toward shared goals like healthcare agents coordinating with patients and doctor.

How do AI Agents Work?

1. Personal: Each agent is given a clearly defined role, personality and communication style along with specific instructions and descriptions of the tools it can use. A well crafted persona ensures the agent behaves consistently and appropriately for its role, while also evolving as it gains experience and engages with users or other systems.

2. Memory: Agents typically have multiple types of memory:

- Short-term memory for the current interaction
- Long-term memory for storing historical data and conversations
- Episodic memory for recalling specific past events
- Consensus memory for sharing knowledge among multiple agents

Memory enables an agent to keep context, learn from experience and adapt its behaviour over time.

3. Tools: These are functions or external resources the agent can use to access information, process data, control devices or connect with other systems. Tools may involve physical interfaces, graphical UIs or programmatic APIs. Agents also learn how and when to use these tools effectively, based on their capabilities and context. 4. Model: Agents use large language model (LLM) which serves as the agent's "brain". The LLM interprets instructions, reasons about solutions, generates language and orchestrates other components including memory retrieval and tools to use to carry out tasks.

Architecture of AI Agents

There are four main components in an AI agent's architecture:

- Profiling Module: This module helps the agent understand its role and purpose. It gathers information from the environment to form perceptions. For example: A self-driving car uses sensors and cameras to detect obstacles.
- Memory Module: The memory module enables the agent to store and retrieve past experiences. This helps the agent learn from prior actions and improve over time. For example: A chatbot remembers past conversations to give better responses.
- Planning Module: This module is responsible for decision-making. It evaluates situations, weighs alternatives and selects the most effective course of action. For example: A chess-playing AI plans its moves based on future possibilities.
- Action Module: The action module executes the decisions made by the planning module in the real world. It translates decisions into real-world actions. For example: A robot vacuum moves to clean a designated area after detecting dirt.

AI Agent Classification

An agent is a system designed to perceive its environment, make decisions and take actions to achieve specific goals. Agents operate autonomously, without direct human control and can be classified based on their behavior, environment and number of interacting agents.

- Reactive Agents: Respond to immediate environmental stimuli without foresight or planning.
- Proactive Agents: Anticipate future states and plan actions to achieve long-term goals.
- Single-Agent Systems: One agent solves a problem independently.
- Multi-Agent Systems: Multiple agents interact, coordinate or compete to achieve goals; may be homogeneous (similar roles) or heterogeneous (diverse roles).
- Rational Agents: Choose actions to maximize expected outcomes using both current and historical information.

Use Cases of AI Agents

Agents are used in a wide range of applications in artificial intelligence, including:

- Robotics: Agents can be used to control robots and automate tasks in manufacturing, transportation and other industries.
- Smart homes and buildings: They can be used to control heating, lighting and other systems in smart homes and buildings, optimizing energy use and improving comfort.
- Healthcare: They can be used to monitor patients, provide personalized treatment plans and optimize healthcare resource allocation.
- Finance: They can be used for automated trading, fraud detection and risk management in the financial industry.
- Games: They can be used to create intelligent opponents in games and simulations, providing a more challenging and realistic experience for players.

Benefits of AI Agents

- Fast and efficient operations.
- Adapt and learn from experience.
- Scalable for large or complex problems.
- Operate autonomously with minimal human input.
- Consistent, reliable task performance.

Limitations:

- Struggle with complex or unpredictable environments.
- High computational needs for learning and planning.
- Communication issues in multi-agent setups.
- Risk of bias or unintended actions.
- Challenges in designing clear goals and utility functions.

III. GENERATIVE AI

Generative AI is a type of artificial intelligence designed to create new content such as text, images, music or even code by learning patterns from existing data. These models generate original outputs that are often indistinguishable from human-created content. These models use techniques like deep learning and neural networks to generate output. Unlike discriminative AI which focuses on classifying data into categories like spam vs. not spam, generative AI creates new data such as text, images, audio or video that resembles real-world examples.

How Generative AI Works

1. Core Mechanism (Training & Inference)

Generative AI is trained on large datasets like text, images, audio or video using deep learning networks. During training, the model learns parameters (millions or billions of them) that help them predict or generate content. Here models generate output based on learned patterns and prompts provided

2. By Media Type

- Text: Uses large language models (LLMs) to predict the next token in a sequence, enabling coherent paragraph or essay generation.
- Images: Diffusion models like DALL·E or Stable Diffusion start with noise and iteratively denoise to create realistic visuals
- Speech: Text-to-speech models synthesize human-like voice by modeling acoustic features based on prompt.
- Video: Multimodal systems like Sora by OpenAI or Runway generate short, temporally coherent video clips from text or other prompts

3. Agents in Generative AI

A modern system often uses agents which are autonomous components that interact with the environment, obtain information and execute chains of tasks. These agents uses LLMs to reason, plan and act enabling workflows like querying databases, performing retrieval or controlling external APIs.

4. Training and Fine-Tuning

LLMs are trained on massive general corpora (e.g., web text) using self-supervised methods. These models become pre-trained models which can be further trained on domain-specific labeled data to adapt to specialized tasks or stylistic needs. This technique is called fine tuning and it can be done using:

- LoRA
- QLoRA
- Peft
- Reinforcement Learning from Human Feedback (RLHF)
- LLM Distilation

5. Retrieval-Augmented Generation (RAG)

Modern systems also uses RAG which enhances outputs by retrieving relevant documents at query time to ground the generation in accurate, up-to-date information, reducing hallucinations and improving factuality. The process typically involves:

- Indexing documents into embeddings stored in vector databases
- Retrieval of relevant passages
- Augmentation of the prompt with retrieved content
- Generation of grounded, informed responses

This approach preserves the base model while enabling dynamic knowledge updates

Types of Generative AI Models

1. Transformers or Autoregressive Models

- Transformers or Autoregressive Models generate sequences by predicting the next token based on all previous ones moving step by step through the text.
- The architecture relies on the transformer's self attention mechanism to capture context from the entire input so far making it highly effective for natural language and code generation.
- Popular examples include GPT models which can produce coherent, context aware paragraphs, solve coding tasks or answer complex queries.
- The autoregressive approach gives fine grained control over each output step but can be slower for long generations since tokens are generated one at a time.

2. Diffusion Models

- Diffusion models generate data such as images or audio by starting with pure random noise and gradually refining it into a coherent output through a series of denoising steps.
- Each step reverses a simulated diffusion process that added noise to real data during training.
- This iterative approach can produce highly detailed and realistic results especially in image synthesis where models like Stable Diffusion and DALL·E 3 have set benchmarks.
- Diffusion models are also versatile they can be adapted for in painting, style transfer and conditional generation from text prompts.

3. Variational Auto encoders (VAEs) and Generative Adversarial Networks (GANs)

- VAEs and GANs were among the first deep learning architectures for generative tasks.
- A VAE encodes data into a compressed latent space and then decodes it back with a probabilistic twist that encourages smooth, continuous representations. This makes them good for controllable generation and interpolation between styles.
- GANs in contrast use two networks against each other a generator that tries to produce realistic outputs and a discriminator that tries to detect fakes.
- This adversarial setup leads to sharp, lifelike images though training can be unstable and prone to mode collapse.

4. Encoder Decoder Models

- Encoder decoder architectures consist of two stages: the encoder processes the input into a dense representation and the decoder generates the desired output from that representation.
- They are widely used for sequence to sequence tasks like language translation, summarization and image captioning.
- The encoder captures the full context of the input before the decoder starts producing tokens, allowing for strong performance on tasks that require global understanding rather than token by token prediction.
- Modern encoder decoder models often use transformers for both stages as in T5, BART and many multimodal systems.

Evaluation of Generative AI

Evaluating generative AI involves multiple dimensions because outputs can vary in accuracy, style and usefulness depending on the task. Key aspects include:

1. Fact Accuracy and Hallucination Avoidance: Benchmarks like BEIR, Natural Questions assess factual correctness. Techniques like RAG and fine-tuning reduce hallucinations and ground responses in reliable data.

2. Quality Metrics: Outputs are judged on fluency, coherence, logical consistency and contextual relevance. Commonly used metrics are BLEU, ROUGE, METEOR, FID and IS.
3. Efficiency and Accuracy Trade-Offs: LoRA and QLoRA helps in balancing performance with computational cost making models faster and lighter without losing quality.
4. Resilience to Retrieval Noise: Advanced approaches like "Finetune-RAG" improve model accuracy by training the model to handle imperfect retrieval inputs hence increasing factual reliability.
5. Creativity and Diversity: Models should generate varied and original outputs rather than repetitive or biased ones.
6. Bias and Fairness: Evaluation includes checking whether outputs reflect harmful stereotypes or unfair treatment of groups. Tools like Bias Benchmark for QA (BBQ) and StereoSet measure bias levels.
7. User Experience and Usefulness: Beyond technical metrics, effectiveness is judged by how well the system supports users in real scenarios like chatbots providing relevant, actionable responses.

Relationship between Humans and Generative AI

- The relationship between humans and generative AI is collaborative and evolving.
- Generative AI serves as a powerful tool that enhances human creativity, productivity and decision making by assisting in tasks like writing, designing, coding and problem solving.
- Rather than replacing humans it augments their abilities allowing people to work faster, explore new ideas and automate repetitive tasks.
- At the same time this relationship raises important questions around ethics, authorship and dependency emphasizing the need for thoughtful use and responsible development of AI technologies.
- Ultimately the synergy between human intuition and generative AI's capabilities has the potential to transform how we create, communicate and innovate.

Generative AI Vs AI

Table 1: Here is the comparison between Generative AI and AI:

Criteria	Generative AI	Artificial Intelligence
Purpose	It is designed to produce new content or data	Designed for a wide range of tasks but not limited to generation
Application	Art creation, text generation, video synthesis and so on	Data analysis, predictions, automation, robotics etc
Learning	Uses Unsupervised learning or reinforcement learning	Can use supervised, semi supervised or reinforcement
Outcome	New or original output is created	Can produce an answer and make a decision, classify, data etc.
Complexity	It requires a complex model like GANs	It has ranged from simple linear regression to complex neural networks
Data Requirement	Required a large amount of data to produce results of high quality data	Data requirements may vary and some need little data and some need vast amounts

Applications of Generative AI

Generative AI is applied in multiple domains like:

- Text: Powers chatbots, virtual assistants, content creation, document summarization and even code generation tools like GitHub Copilot.
- Images: Used in digital art, product design, fashion, advertising and medical imaging to create visuals that are close to real-world examples.
- Audio & Speech: Enables natural-sounding voice assistants, multilingual dubbing, music composition, personalized voice cloning and accessibility tools.
- Video: Supports animation, movie special effects, gaming, marketing videos and realistic training simulations.
- Business Use Cases: Enhances customer support with AI agents, boosts knowledge discovery using RAG systems, accelerates drug discovery, assists in financial forecasting and improves data-driven decision-making.

Advantages

1. Accelerates Research and Development: In fields like science and technology, Generative AI reduces the time needed for research by generating multiple outcomes and predictions such as molecular structures in drug development. This speeds up innovation and helps solve complex problems efficiently.
2. Improves Personalization: Generative AI creates tailored content based on user preferences. From personalized product designs to customized marketing campaigns it enhances user engagement and satisfaction by delivering exactly what users need or want.
3. Empowers Non Experts: Even users without expertise can create high quality content using Generative AI. This helps individuals learn new skills access creative tools and open doors to personal and professional growth.
4. Drives Economic Growth: Generative AI introduces new roles and opportunities by fostering innovation, automating tasks and enhancing productivity. This leads to economic expansion and the creation of jobs in emerging fields.

Disadvantages

1. Data Dependence: The accuracy and quality of Generative AI outputs depend entirely on the data it is trained on. If the training data is biased, incomplete or inaccurate the generated content will reflect these flaws.

2. Limited Control over Outputs: It can produce unexpected or irrelevant results making it challenging to control the content and ensure it aligns with specific user requirements.
3. High Computational Requirements: Training and running Generative AI models demand significant computing power which can be costly and resource intensive. This limits accessibility for smaller organizations or individuals.
4. Ethical and Legal Concerns: It can be misused to create harmful content like deep fakes or fake news which can spread misinformation or violate privacy. These ethical and legal challenges require careful regulation and oversight to prevent abuse.

Applications of AI

While AI is transforming industries with advanced solutions, tools like the Paraphrasing tool can enhance productivity in writing by providing free, full-length reinterpretations without signup hassles. This seamless integration of technology aligns with the evolving landscape influenced by AI, adding significant educational value. AI has become a vital aspect of many sectors.

1. Healthcare: AI helps with disease diagnosis, drug discovery, and personalised treatment strategies. AI-powered tools, such as IBM Watson, are transforming health care.
2. Businesses: Businesses use AI for customer service through chatbots, predictive analytics for decision-making, and automation to increase efficiency.
3. Transportation: Autonomous vehicles and AI-powered traffic control technologies are transforming urban mobility.
4. Education: Artificial intelligence (AI) enables personalised learning experiences by identifying gaps in student knowledge and delivering relevant answers.
5. Entertainment: AI-powered recommendation systems on platforms such as Netflix and Spotify deliver user-specific content.

Benefits of AI

Here are some of the advantages of Artificial Intelligence that cannot be overlooked.

1. Efficiency: Artificial intelligence automates repetitive tasks, saving time and resources.
2. Accuracy: Machine learning algorithms are frequently better than humans at recognising patterns and predicting outcomes.
3. 24/7 Availability: AI systems are capable of operating continually without tiredness.
4. Problem-solving: Artificial intelligence solves hard problems in domains such as climate modeling and scientific research.

Challenges of AI

Despite its benefits, AI poses various challenges:

1. Job Displacement: Automation may cause unemployment in industries that rely on routine tasks.
2. Bias and Inequality: AI algorithms might inherit biases from training data, resulting in unfair outputs.
3. Privacy Concerns: The usage of personal data raises questions regarding surveillance and misuse.
4. Security Risks: The malicious use of AI in cyber attacks is a big hazard.

The Future of AI

AI's potential is limitless, with advances in quantum computing, 6G networks, and more realistic virtual realities. However, addressing ethical problems and ensuring responsible AI development is critical. Policymakers and technologists must work together to design policies that promote innovation while protecting societal values.

IV. RAG IN ADVANCING AFFORDABLE HEALTHCARE SYSTEMS

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.

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.

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 users which are accurate and precise compared to default LLM answers on knowledge-intensive tasks.

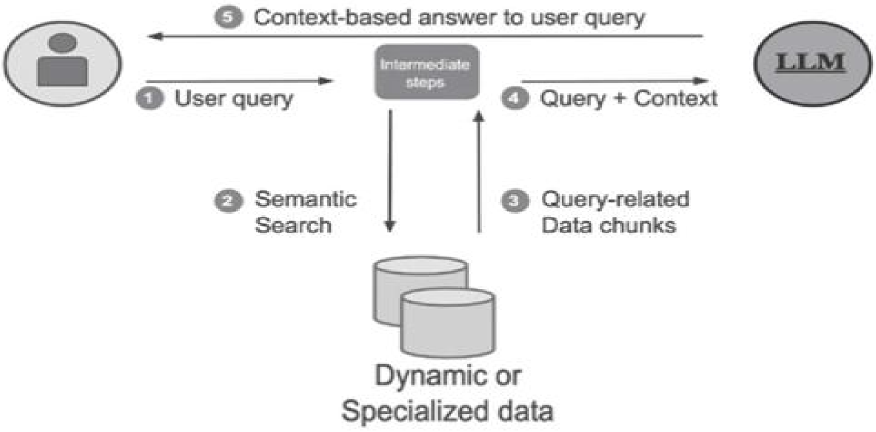

Fig 2. RAG Models

Key Features of RAG:

- Data Integration: Incorporates data represented in a structured or unstructured manner, and updated in real time.
- Explain ability: 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 contexts 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 remain 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.

V. 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 pharma covigilance with reporting adverse drug effects in real-time.

Table 2 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

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.

PROPOSED FRAMEWORK USING RAG IN HEALTHCARE

Healthcare data has become fragmented and exists in silos for a number of reasons. However, many helpful and favourable laws for patients like HIPAA and GDPR only restrict the flow of data and hinder the gains made from its analysis. 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, silted, 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.

CHALLENGES AND FUTURE DIRECTIONS

Both technologies face challenges, including data bases, 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.

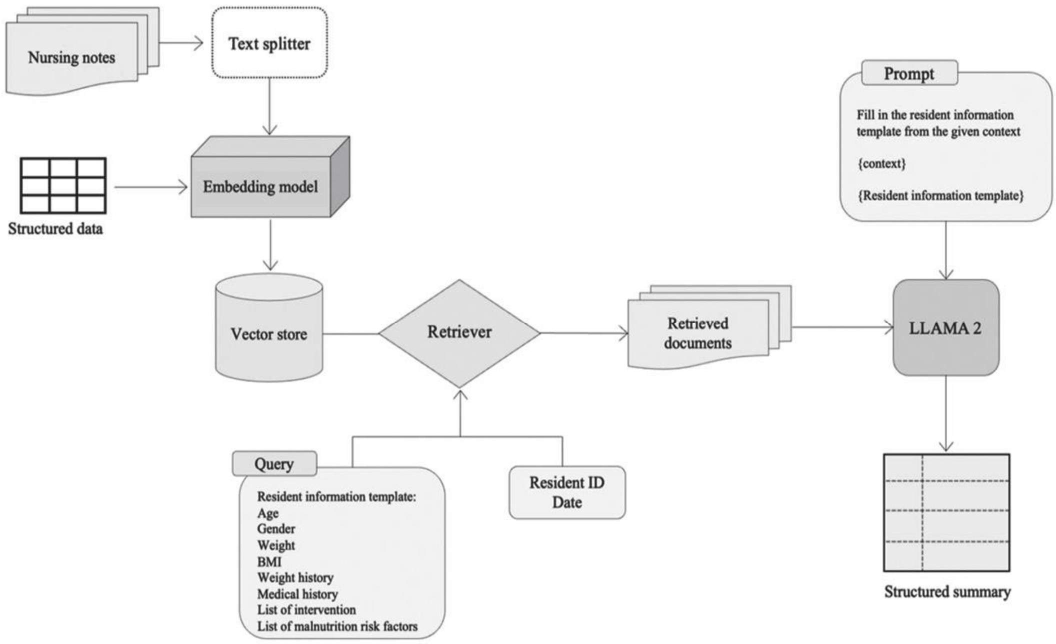

Fig 3. Proposed RAG Approach

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.

VI. CONCLUSION

Artificial intelligence is transforming the way we live and work. Including it into daily life is unavoidable, providing solutions to some of humanity's most pressing problems. However, careful execution is required to ensure it works as a tool for advancement rather than disturbance. By maintaining a balance between innovation and ethics, AI can be used to create a better future for all of humanity. After all the above discussion that the importance of Artificial Intelligence; Machine Learning and Deep Learning is increasing constantly in human life. Business Organisations and Customers are taking advantage of these technologies. In the current scenario, Artificial Intelligence is an essential part of the Business Organizations due to High Competition and in order to detect fraudulent activity as well as to detect hacking of data, etc. 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. Adam C. Uzialko. (April 22, 2019). How Artificial Intelligence Will Transform Business. Freelance Editor; Business News Daily.
2. John Dages, Sajith Wanigasinghe, Patrick Carfrey. (August 13, 2020) .Fusion Experts: The buzz about artificial intelligence, machine learning, and deep learning: An overview" Fusion Alliance.
3. Some sources interviews were conducted for getting more about personal experienced of work from Relevant Field of Marketing.
4. Tyler Danner; Artificial Intelligence, Machine Learning, and Deep Learning: Aunalytics.
5. 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
6. 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," parXiv: 2309.05475 https://10.48550/arXiv.2309.05475
7. Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, Robert McHardy, "Challenges and applications of large language models," p. arXiv:2307.10169 https://doi/10.48550/arXiv.2307.10169.
8. 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:2302.04023 https://doi:/10.48550/arXiv.2302.04023