Enhancing medical language models with big data technologies

Ayoub Allali[1], Ibtihal Abouchabaka[2], Najat Rafalia[1]
[1]Department of Computer Science, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco
[2]Mohammed VI University of Sciences and Health, Casablanca, Morocco

Article Info

Article history:

Received Jun 1, 2025
Revised Jan 1, 2026
Accepted Jan 22, 2026

Keywords:

Big data
DeepSeek-R1
GPT 4o-mini
LLMs
Llama 3
Retrieval-augmented generation
Vector database

ABSTRACT

In this study, we present an end-to-end, big-data–driven framework for continuously enriching and fine-tuning large language models (LLMs) with the latest professional and scientific medical knowledge. Streaming updates from premier sources such as The New England Journal of Medicine (NEJM) are ingested via an Apache Kafka cluster for low-latency delivery and durably archived in a three-node Apache Hadoop (Hadoop distributed file system (HDFS)) system. Each new article is preprocessed into high-dimensional embeddings and indexed in a Milvus vector database to enable sub-second semantic retrieval over millions of records. At query or batch time, our retrieval-augmented generation (RAG) module retrieves the top-k relevant embeddings from Milvus and injects them into prompts for DeepSeek-R1, GPT-4o-mini, and Llama 3, models which are hosted, fine-tuned, and served via Ollama on an NVIDIA GeForce RTX 3050 Ti GPU for efficient inference and continual learning. The enriched outputs are seamlessly delivered to end users through a Telegram bot programmed in Python using the Telebot library, linking the RAG-enhanced LLMs to an intuitive chat interface. Our Kafka, HDFS, Milvus, RAG, LLM, or Telegram bot pipeline demonstrably improves factual accuracy and topical currency of AI-generated medical insights across clinical decision support, patient engagement and education, drug discovery and development, virtual health assistants, and mental health support, laying the groundwork for truly intelligent, responsive, and data-driven healthcare solutions.

Corresponding Author:

Ayoub Allali
Department of Computer Science, Faculty of Sciences, Ibn Tofail University
Kenitra, Morocco
Email: ayoub.allali@uit.ac.ma

1. INTRODUCTION

The healthcare sector is experiencing a paradigm shift driven by advances in artificial intelligence and big-data technologies. In particular, large language models (LLMs) have shown tremendous promise across medical domains, from clinical decision support to virtual health assistants. But their real-world impact depends on access to high-quality, up-to-date knowledge. With the volume of medical literature, clinical studies, and expert commentary growing exponentially, it remains a pressing challenge to ingest, store, process, and surface this ever-expanding body of information in ways that keep pace with emerging discoveries and evolving standards of care.

To meet this challenge, we introduce a fully integrated, big-data architecture that couples real-time streaming, scalable storage, semantic indexing, and retrieval-augmented generation (RAG) to continuously enrich and fine-tune LLMs for medical applications. First, feeds from leading sources such as the New England Journal of Medicine (NEJM) and the Lancet are captured via an Apache Kafka cluster, ensuring sub-second delivery of newly published articles. Raw documents are durably archived in a three-node

Hadoop distributed file system (HDFS) for fault-tolerant, distributed storage. Each incoming record is then preprocessed into high-dimensional embeddings and indexed in a Milvus vector database, enabling millisecond-scale semantic retrieval across millions of medical records.

At query or batch time, our RAG module retrieves the top-k relevant vectors from Milvus and injects their content into prompts for three state-of-the-art LLMs, DeepSeek-R1, GPT-4o-mini, and Llama 3, each hosted and fine-tuned on an NVIDIA GeForce RTX 3050 Ti GPU via Ollama for optimized inference and continual learning. Finally, clinicians, researchers, and patients interact with the enriched LLM outputs through a Python-based Telegram bot (built with the Telebot library). Providing an intuitive chat interface that delivers evidence-backed insights directly to end users' mobile devices.

The primary goal of this research is to demonstrate the potential of big data technologies in improving the reliability, accuracy, and real-time adaptability of AI-driven medical applications, our proposed framework is designed to support a wide range of healthcare use cases, including:

i) Clinical decision support: assisting healthcare professionals in diagnosing conditions, recommending treatments, and identifying potential risks based on the latest medical literature.
ii) Patient engagement and education: providing personalized, evidence-based responses to patient queries, improving health literacy, and self-care management.
iii) Drug discovery and development: accelerating pharmaceutical research by analyzing vast datasets of clinical trials, drug interactions, and biomedical studies.
iv) Virtual health assistants: enhancing telemedicine services with AI-powered chatbots capable of understanding and responding to medical queries with up-to-date knowledge.
v) Mental health support: leveraging AI to provide conversational support, detect early signs of mental health conditions, and recommend appropriate interventions.

Burgan *et al.* [1] present RamChat, an AI chatbot designed to help Shepherd University students navigate their student handbook, developed in Python. RamChat integrates both API-based and local LLMs using the LangChain framework and a vector store system. The chatbot leverages OpenAI's text-embedding-3-small model for embeddings and initially used OpenAI's davinci-002 model, later replaced with gemma, a local LLM based on Google's Gemini model. The Ollama framework enables automatic LLM selection based on user prompts. The development process involved testing different LLMs, debugging, and optimizing RamChat's performance. Their conference presentation will cover the methodology, challenges, and insights gained from developing this AI-powered student assistant.

Mao *et al.* [2] discuss the challenges of updating LLMs with long-tail or outdated knowledge due to their vast number of parameters, making fine-tuning impractical. Instead, they highlight the effectiveness of black-box RAG, which enhances LLMs without modifying their parameters. Existing black-box RAG methods often fine-tune the retriever to align with LLM preferences but face two key issues: ignoring factual information, which can mislead the retriever, and inefficient token usage due to concatenating all retrieved documents.

Schiele *et al.* [3] examine the impact of information and communication technologies (ICTs) on political engagement, particularly in the context of voting decisions, with the rise of issue-based voting in western democracies. There is a growing need for transparent and unbiased voting advice applications (VAAs) like Switzerland's Smartvote and Germany's Wahl-O-Mat. The authors propose that integrating LLMs with RAG techniques could enhance VAAs by improving fairness, impartiality, and transparency.

While these studies demonstrate advancements in AI-powered applications using LLMs and RAG, they largely overlook the critical role of big data technologies in ensuring scalability, efficiency, and real-time processing. Their reliance on static embeddings, API-based models, and fine-tuned retrievers limits their ability to handle large-scale, continuously evolving datasets. Unlike these approaches, our study leverages Apache Hadoop for distributed data storage and Apache Kafka for real-time data streaming, enabling dynamic updates, high-throughput processing, and improved responsiveness. By integrating big data frameworks, our research addresses key challenges in large-scale AI applications that these studies fail to consider, ensuring a more scalable, and data-driven solution.

Traditional LLMs like Med-PaLM, BioGPT, and PubMedGPT are typically trained on static datasets, leading to knowledge cutoffs that may be several months or even years old. This static nature limits their ability to incorporate and respond to the latest medical research promptly as shown in Table 1 [4]. In contrast, RAG systems enhance LLMs by integrating dynamic retrieval mechanisms, allowing them to access and utilize up-to-date information from external databases. This approach significantly reduces the latency in incorporating new medical knowledge into the model's responses.

This paper explores the implementation details, challenges, and advantages of using big data pipelines in conjunction with RAG-enhanced LLMs for medical applications. We focus on the impact of real-time data streaming, distributed storage, and frequent incremental training, which enable our model to incorporate newly published medical research on a daily basis. By contrast, existing models such as

Med-PaLM, BioGPT, and PubMedGPT rely on static pretraining and require months to years to update their knowledge. Integrating these technologies allows our system to provide healthcare professionals with accurate, timely, and evidence-based responses, significantly improving the performance, responsiveness, and relevance of AI-driven solutions in clinical decision-making and patient care.

Table 1. Knowledge update latency for new research papers

Model	Knowledge update method	Typical latency for new paper
Med-PaLM/Med-PaLM-2	Static pretrained+fine-tuned	Months-years
BioGPT	Static pretrained model (PubMed snapshot)	Months-years
PubMedGPT	Static pretrained model (PubMed snapshot)	Months-years

2. METHOD

To bridge big-data technologies with LLMs in a medical setting, we designed a four-stage pipeline as shown in Figure 1. Data collection and streaming, where professional and scientific medical news are ingested in real-time via Apache Kafka. Distributed storage, which archives incoming documents in a multi-node Hadoop HDFS cluster for scalable, fault-tolerant persistence. Semantic encoding and RAG integration, transforming each article into high-dimensional embeddings, indexing them in a Milvus vector database, and augmenting LLMs (DeepSeek, GPT-4o-mini, and Llama 3) via RAG. Interactive AI-driven text generation, delivering ask/answer queries and generated medical insights through a user-facing interface (Telegram bot) [5].

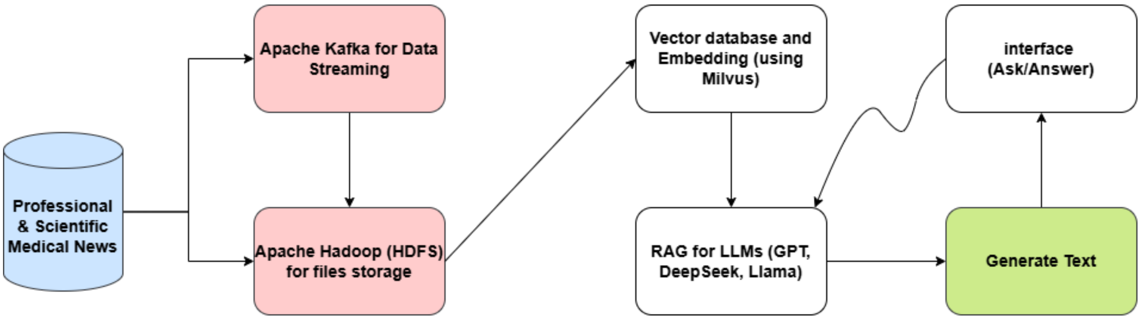

Professional & Scientific Medical News
Apache Kafka for Data Streaming
Apache Hadoop (HDFS) for files storage
Vector database and Embedding (using Milvus)
RAG for LLMs (GPT, DeepSeek, Llama)
interface (Ask/Answer)
Generate Text

Figure 1. Proposed methodology

To create a comprehensive and continuously updated medical dataset for enhancing LLMs, we implemented an automated data collection pipeline leveraging really simple syndication (RSS) feeds and Apache Kafka. We focused on obtaining medical literature from the NEJM, a highly reputable medical journal, by utilizing its publicly available RSS feeds as shown in Figure 2. Through these feeds, we systematically monitored and retrieved the latest published articles in real-time. Each new article identified via the RSS feed was automatically downloaded, preserving its original layout, figures, and textual content to ensure data integrity and completeness [6].

The real-time streaming component was managed using Apache Kafka, configured within our three-node Hadoop cluster (one NameNode and two DataNodes). Kafka's producers continuously monitored the NEJM RSS feed for updates, automatically downloading new articles as soon as they became available. Upon successful retrieval, Kafka brokers distributed these articles across designated partitions within the Kafka topics, facilitating balanced load management and ensuring fault-tolerance during data ingestion, as shown in Figure 3 [7]–[9].

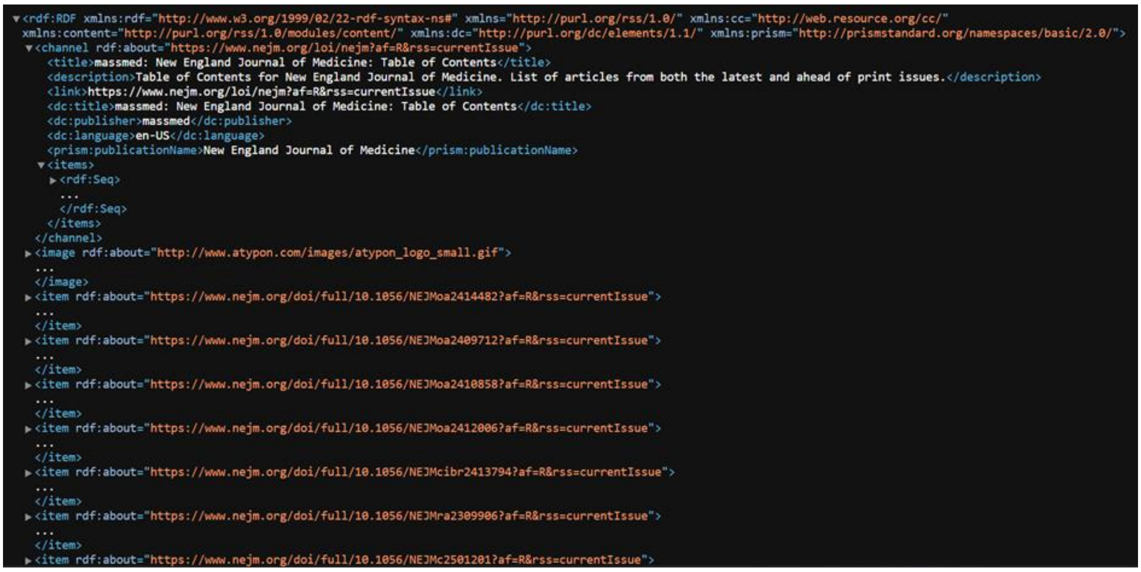
Kafka consumers then systematically processed these articles, extracting relevant textual content through dedicated parsing mechanisms. This extracted text was structured into a format suitable for subsequent storage in the HDFS. The integration of Kafka ensured seamless, uninterrupted streaming of medical articles directly into the HDFS, maintaining an up-to-date, structured, and easily retrievable repository, as shown in Figure 4.

This enhanced data collection method allowed us to achieve robust, real-time ingestion of high-quality medical literature, laying a strong foundation for training and refining our RAG-enhanced LLMs. To store the continuously collected medical data, we deployed a Hadoop cluster with three nodes, the HDFS is used to provide scalable, fault-tolerant storage for large volumes of unstructured text data, Apache

Kafka seamlessly integrates with Hadoop, where Kafka consumers pull the streamed data and store it in HDFS, the Hadoop cluster consists of:

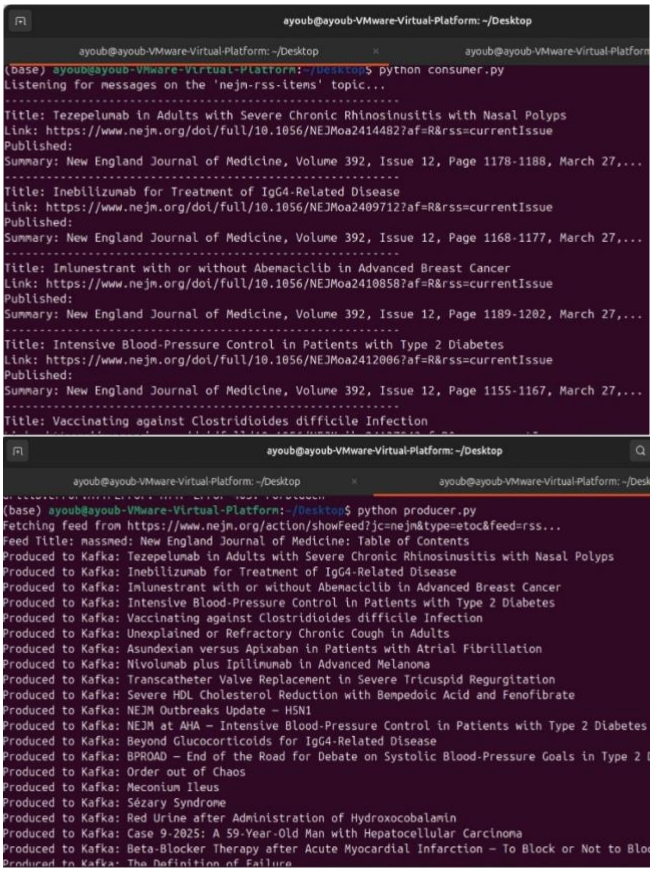
i) One NameNode: manages the metadata and directory structure of the distributed file system.

ii) Two DataNodes: store the actual raw data files and distribute the storage load for redundancy and high availability.



```
▼<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns="http://purl.org/rss/1.0/" xmlns:cc="http://web.resource.org/cc/"
 xmlns:content="http://purl.org/rss/1.0/modules/content/" xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:prism="http://prismstandard.org/namespaces/basic/2.0/">
  ▼<channel rdf:about="https://www.nejm.org/loi/nejm?af=R&rss=currentIssue">
     <title>massmed: New England Journal of Medicine: Table of Contents</title>
     <description>Table of Contents for New England Journal of Medicine. List of articles from both the latest and ahead of print issues.</description>
     <link>https://www.nejm.org/loi/nejm?af=R&rss=currentIssue</link>
     <dc:title>massmed: New England Journal of Medicine: Table of Contents</dc:title>
     <dc:publisher>massmed</dc:publisher>
     <dc:language>en-US</dc:language>
     <prism:publicationName>New England Journal of Medicine</prism:publicationName>
   ▼<items>
     ▶<rdf:Seq>
       ...
      </rdf:Seq>
     </items>
   </channel>
  ▶<image rdf:about="http://www.atypon.com/images/atypon_logo_small.gif">
   ...
   </image>
  ▶<item rdf:about="https://www.nejm.org/doi/full/10.1056/NEJMoa2414482?af=R&rss=currentIssue">
   ...
   </item>
  ▶<item rdf:about="https://www.nejm.org/doi/full/10.1056/NEJMoa2409712?af=R&rss=currentIssue">
   ...
   </item>
  ▶<item rdf:about="https://www.nejm.org/doi/full/10.1056/NEJMoa2410858?af=R&rss=currentIssue">
   ...
   </item>
  ▶<item rdf:about="https://www.nejm.org/doi/full/10.1056/NEJMoa2412006?af=R&rss=currentIssue">
   ...
   </item>
  ▶<item rdf:about="https://www.nejm.org/doi/full/10.1056/NEJMcibr2413794?af=R&rss=currentIssue">
   ...
   </item>
  ▶<item rdf:about="https://www.nejm.org/doi/full/10.1056/NEJMra2309906?af=R&rss=currentIssue">
   ...
   </item>
  ▶<item rdf:about="https://www.nejm.org/doi/full/10.1056/NEJMc2501201?af=R&rss=currentIssue">
```

Figure 2. The NEJM RSS feed



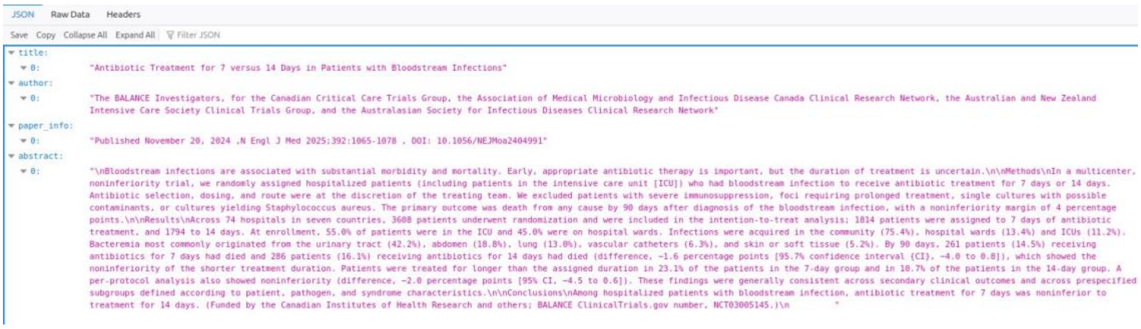
ayoub@ayoub-VMware-Virtual-Platform: ~/Desktop
ayoub@ayoub-VMware-Virtual-Platform: ~/Desktop
ayoub@ayoub-VMware-Virtual-Platform

```
(base) ayoub@ayoub-VMware-Virtual-Platform:~/Desktop$ python consumer.py
Listening for messages on the 'nejm-rss-items' topic...
--------------------------------------------------------
Title: Tezepelumab in Adults with Severe Chronic Rhinosinusitis with Nasal Polyps
Link: https://www.nejm.org/doi/full/10.1056/NEJMoa2414482?af=R&rss=currentIssue
Published:
Summary: New England Journal of Medicine, Volume 392, Issue 12, Page 1178-1188, March 27,...
--------------------------------------------------------
Title: Inebilizumab for Treatment of IgG4-Related Disease
Link: https://www.nejm.org/doi/full/10.1056/NEJMoa2409712?af=R&rss=currentIssue
Published:
Summary: New England Journal of Medicine, Volume 392, Issue 12, Page 1168-1177, March 27,...
--------------------------------------------------------
Title: Imlunestrant with or without Abemaciclib in Advanced Breast Cancer
Link: https://www.nejm.org/doi/full/10.1056/NEJMoa2410858?af=R&rss=currentIssue
Published:
Summary: New England Journal of Medicine, Volume 392, Issue 12, Page 1189-1202, March 27,...
--------------------------------------------------------
Title: Intensive Blood-Pressure Control in Patients with Type 2 Diabetes
Link: https://www.nejm.org/doi/full/10.1056/NEJMoa2412006?af=R&rss=currentIssue
Published:
Summary: New England Journal of Medicine, Volume 392, Issue 12, Page 1155-1167, March 27,...
--------------------------------------------------------
Title: Vaccinating against Clostridioides difficile Infection
```

ayoub@ayoub-VMware-Virtual-Platform: ~/Desktop
ayoub@ayoub-VMware-Virtual-Platform: ~/Desktop
ayoub@ayoub-VMware-Virtual-Platform: ~/Desk

```
(base) ayoub@ayoub-VMware-Virtual-Platform:~/Desktop$ python producer.py
Fetching feed from https://www.nejm.org/action/showFeed?jc=nejm&type=etoc&feed=rss...
Feed Title: massmed: New England Journal of Medicine: Table of Contents
Produced to Kafka: Tezepelumab in Adults with Severe Chronic Rhinosinusitis with Nasal Polyps
Produced to Kafka: Inebilizumab for Treatment of IgG4-Related Disease
Produced to Kafka: Imlunestrant with or without Abemaciclib in Advanced Breast Cancer
Produced to Kafka: Intensive Blood-Pressure Control in Patients with Type 2 Diabetes
Produced to Kafka: Vaccinating against Clostridioides difficile Infection
Produced to Kafka: Unexplained or Refractory Chronic Cough in Adults
Produced to Kafka: Asundexian versus Apixaban in Patients with Atrial Fibrillation
Produced to Kafka: Nivolumab plus Ipilimumab in Advanced Melanoma
Produced to Kafka: Transcatheter Valve Replacement in Severe Tricuspid Regurgitation
Produced to Kafka: Severe HDL Cholesterol Reduction with Bempedoic Acid and Fenofibrate
Produced to Kafka: NEJM Outbreaks Update — H5N1
Produced to Kafka: NEJM at AHA — Intensive Blood-Pressure Control in Patients with Type 2 Diabetes
Produced to Kafka: Beyond Glucocorticoids for IgG4-Related Disease
Produced to Kafka: BPROAD — End of the Road for Debate on Systolic Blood-Pressure Goals in Type 2 D
Produced to Kafka: Order out of Chaos
Produced to Kafka: Meconium Ileus
Produced to Kafka: Sézary Syndrome
Produced to Kafka: Red Urine after Administration of Hydroxocobalamin
Produced to Kafka: Case 9-2025: A 59-Year-Old Man with Hepatocellular Carcinoma
Produced to Kafka: Beta-Blocker Therapy after Acute Myocardial Infarction — To Block or Not to Bloc
Produced to Kafka: The Definition of Failure
```

Figure 3. Apache Kafka producer/consumer



```
▾ title:
  ▾ 0:      "Antibiotic Treatment for 7 versus 14 Days in Patients with Bloodstream Infections"
▾ author:
  ▾ 0:      "The BALANCE Investigators, for the Canadian Critical Care Trials Group, the Association of Medical Microbiology and Infectious Disease Canada Clinical Research Network, the Australian and New Zealand
            Intensive Care Society Clinical Trials Group, and the Australasian Society for Infectious Diseases Clinical Research Network"
▾ paper_info:
  ▾ 0:      "Published November 20, 2024 ,N Engl J Med 2025;392:1065-1078 , DOI: 10.1056/NEJMoa2404991"
▾ abstract:
  ▾ 0:      "\nBloodstream infections are associated with substantial morbidity and mortality. Early, appropriate antibiotic therapy is important, but the duration of treatment is uncertain.\n\nMethods\nIn a multicenter,
            noninferiority trial, we randomly assigned hospitalized patients (including patients in the intensive care unit [ICU]) who had bloodstream infection to receive antibiotic treatment for 7 days or 14 days.
            Antibiotic selection, dosing, and route were at the discretion of the treating team. We excluded patients with severe immunosuppression, foci requiring prolonged treatment, single cultures with possible
            contaminants, or cultures yielding Staphylococcus aureus. The primary outcome was death from any cause by 90 days after diagnosis of the bloodstream infection, with a noninferiority margin of 4 percentage
            points.\n\nResults\nAcross 74 hospitals in seven countries, 3608 patients underwent randomization and were included in the intention-to-treat analysis; 1814 patients were assigned to 7 days of antibiotic
            treatment, and 1794 to 14 days. At enrollment, 55.0% of patients were in the ICU and 45.0% were on hospital wards. Infections were acquired in the community (75.4%), hospital wards (13.4%) and ICUs (11.2%).
            Bacteremia most commonly originated from the urinary tract (42.2%), abdomen (18.8%), lung (13.0%), vascular catheters (6.3%), and skin or soft tissue (5.2%). By 90 days, 261 patients (14.5%) receiving
            antibiotics for 7 days had died and 286 patients (16.1%) receiving antibiotics for 14 days had died (difference, −1.6 percentage points [95.7% confidence interval {CI}, −4.0 to 0.8]), which showed the
            noninferiority of the shorter treatment duration. Patients were treated for longer than the assigned duration in 23.1% of the patients in the 7-day group and in 10.7% of the patients in the 14-day group. A
            per-protocol analysis also showed noninferiority (difference, −2.0 percentage points [95% CI, −4.5 to 0.6]). These findings were generally consistent across secondary clinical outcomes and across prespecified
            subgroups defined according to patient, pathogen, and syndrome characteristics.\n\nConclusions\nAmong hospitalized patients with bloodstream infection, antibiotic treatment for 7 days was noninferior to
            treatment for 14 days. (Funded by the Canadian Institutes of Health Research and others; BALANCE ClinicalTrials.gov number, NCT03005145.)\n       "
```

Figure 4. Data example

Data stored in HDFS is preprocessed and structured to improve retrieval efficiency during AI model training and inference. This architecture allows efficient scaling. Ensuring that as the volume of medical literature grows, the system can handle increased data loads without compromising performance [10]–[13].

In our pipeline, the vector database plays a central role in enabling fast, semantically aware retrieval of medical knowledge. After ingesting and archiving raw articles, each document is preprocessed, tokenized, cleaned, and passed through a transformer-based encoder to produce fixed length embedding vectors that capture contextual meaning. These embeddings are then ingested into Milvus. We configure an inverted file with product quantization (IVF-PQ) index to balance search speed and memory footprint, and we periodically rebuild index shards to accommodate new data without service interruption. At query time, top-k nearest neighbor searches retrieve the most relevant article embeddings in sub-second latency. These retrieved vectors are then decoded back into document passages and fused into prompts for our RAG module. By leveraging Milvus's scalable architecture and advanced indexing techniques, our system maintains millisecond-scale semantic retrieval performance even as the medical corpus grows into the millions of records, ensuring that LLMs always draw on the most pertinent and up-to-date information [14].

To enhance the capabilities of LLMs in generating medically accurate and contextually relevant text, we implement RAG, instead of solely relying on pre-trained LLM knowledge, RAG enables models to dynamically retrieve relevant medical information from the HDFS-stored dataset before generating responses [15]. We fine-tune three state-of-the-art LLMs (DeepSeek-R1, GPT-4o-mini, and Llama 3) using the curated medical dataset stored in HDFS, the fine-tuning process involves:

i) Dataset preparation: extracting key medical articles and structuring them for fine-tuning.
ii) Model training: using Ollama, an open-source platform optimized for efficient LLM deployment, to fine-tune models on an NVIDIA GeForce RTX 3050 Ti GPU, this setup ensures faster training times and improved inference performance.
iii) RAG implementation: combining a retrieval mechanism with LLMs, where models first search the HDFS-stored dataset for relevant information before generating responses, this reduces hallucinations and enhances the factual accuracy of generated medical insights.

To rigorously evaluate the effectiveness of our retrieval pipeline, we employed standard information retrieval metrics, namely Recall@k and mean reciprocal rank (MRR). Recall@k measures the proportion of relevant documents captured within the top-k retrieved results, reflecting the breadth of coverage for a given medical query. MRR, in contrast, emphasizes how early the first relevant document appears in the ranked list, rewarding systems that surface highly pertinent information at the top of the results. Using a benchmark set of clinical and biomedical questions, we compared our daily updated RAG system against static baselines such as Med-PaLM, BioGPT, and PubMedGPT. The results demonstrate that our approach consistently achieves higher Recall@k and MRR scores as shown in Table 2, indicating superior coverage of newly published articles and faster access to the most relevant evidence. This validates that continuous ingestion and indexing of medical literature substantially improve retrieval quality and directly enhance the reliability of downstream answer generation.

Table 2. Evaluation results of models

Model	Recall@10	Recall@20	MRR@10	MRR@20
Our daily RAG model (GPT 4o mini)	0.82	0.91	0.74	0.81
Med-PaLM	0.45	0.60	0.39	0.52
BioGPT	0.48	0.62	0.42	0.55
PubMedGPT	0.50	0.65	0.44	0.57

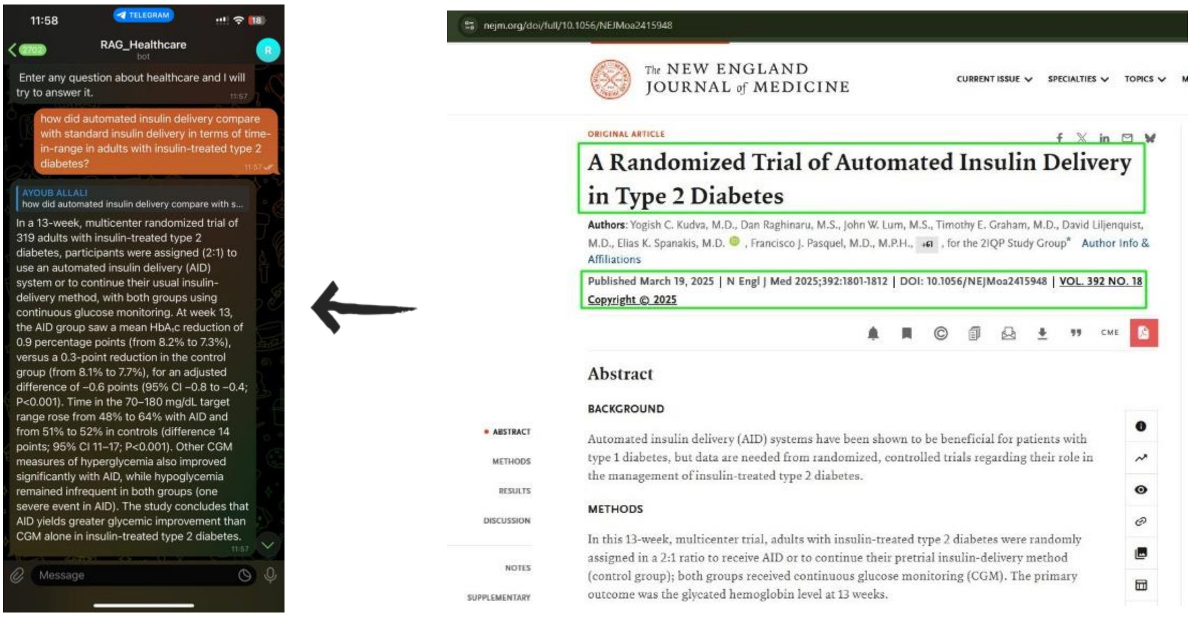
3. RESULTS AND DISCUSSION

3.1. AI-driven text generation for medical applications

Once the fine-tuned models are optimized with RAG, they are deployed to generate medical insights in various applications, as described in the following.

3.1.1. Clinical decision support

Our RAG-enhanced LLMs continuously pull in the latest peer-reviewed studies and guidelines from Milvus, synthesize the most relevant findings, and present concise, evidence-backed summaries to clinicians. By embedding this capability into the Telegram bot interface, doctors can query complex cases, such as emerging therapeutic protocols or rare adverse events, and receive substantiated recommendations in seconds. Ensuring treatment decisions are grounded in the most current medical literature, as shown in Figures 5 and 6.



Enter any question about healthcare and I will try to answer it. 11:57

how did automated insulin delivery compare with standard insulin delivery in terms of time-in-range in adults with insulin-treated type 2 diabetes? 11:57

AYOUB ALLALI
how did automated insulin delivery compare with s...

In a 13-week, multicenter randomized trial of 319 adults with insulin-treated type 2 diabetes, participants were assigned (2:1) to use an automated insulin delivery (AID) system or to continue their usual insulin-delivery method, with both groups using continuous glucose monitoring. At week 13, the AID group saw a mean HbA_1c reduction of 0.9 percentage points (from 8.2% to 7.3%), versus a 0.3-point reduction in the control group (from 8.1% to 7.7%), for an adjusted difference of −0.6 points (95% CI −0.8 to −0.4; P<0.001). Time in the 70–180 mg/dL target range rose from 48% to 64% with AID and from 51% to 52% in controls (difference 14 points; 95% CI 11–17; P<0.001). Other CGM measures of hyperglycemia also improved significantly with AID, while hypoglycemia remained infrequent in both groups (one severe event in AID). The study concludes that AID yields greater glycemic improvement than CGM alone in insulin-treated type 2 diabetes. 11:57

ORIGINAL ARTICLE

A Randomized Trial of Automated Insulin Delivery in Type 2 Diabetes

Authors: Yogish C. Kudva, M.D., Dan Raghinaru, M.S., John W. Lum, M.S., Timothy E. Graham, M.D., David Liljenquist, M.D., Elias K. Spanakis, M.D. , Francisco J. Pasquel, M.D., M.P.H., +61 , for the 2IQP Study Group* Author Info & Affiliations

Published March 19, 2025 | N Engl J Med 2025;392:1801-1812 | DOI: 10.1056/NEJMoa2415948 | VOL. 392 NO. 18
Copyright © 2025

Abstract

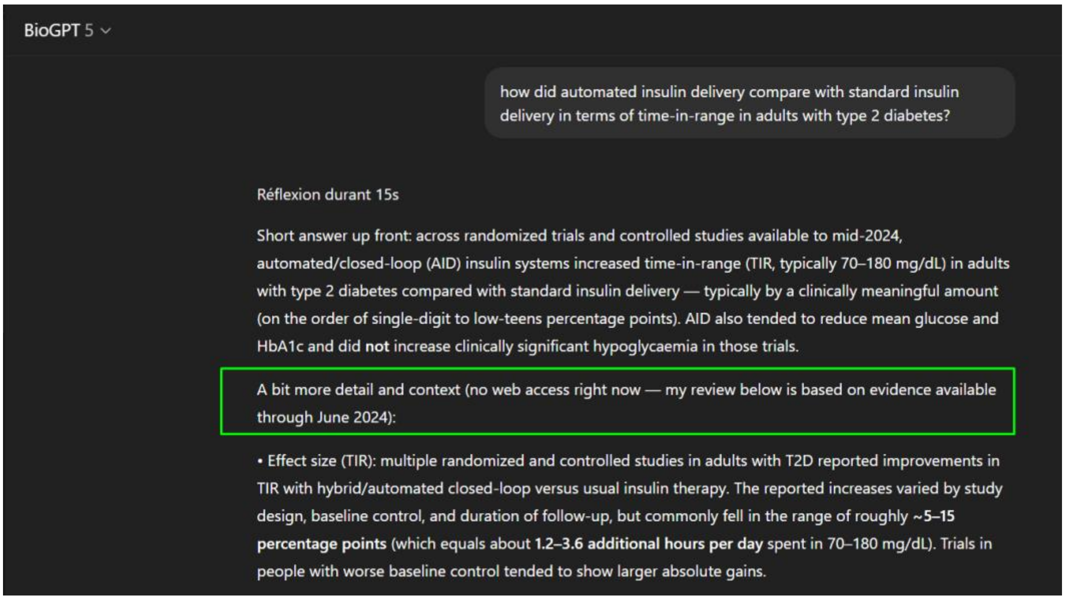
BACKGROUND

Automated insulin delivery (AID) systems have been shown to be beneficial for patients with type 1 diabetes, but data are needed from randomized, controlled trials regarding their role in the management of insulin-treated type 2 diabetes.

METHODS

In this 13-week, multicenter trial, adults with insulin-treated type 2 diabetes were randomly assigned in a 2:1 ratio to receive AID or to continue their pretrial insulin-delivery method (control group); both groups received continuous glucose monitoring (CGM). The primary outcome was the glycated hemoglobin level at 13 weeks.

Figure 5. Clinical decision support system powered by GPT 4o mini with RAG



how did automated insulin delivery compare with standard insulin delivery in terms of time-in-range in adults with type 2 diabetes?

Réflexion durant 15s

Short answer up front: across randomized trials and controlled studies available to mid-2024, automated/closed-loop (AID) insulin systems increased time-in-range (TIR, typically 70–180 mg/dL) in adults with type 2 diabetes compared with standard insulin delivery — typically by a clinically meaningful amount (on the order of single-digit to low-teens percentage points). AID also tended to reduce mean glucose and HbA1c and did **not** increase clinically significant hypoglycaemia in those trials.

A bit more detail and context (no web access right now — my review below is based on evidence available through June 2024):

- Effect size (TIR): multiple randomized and controlled studies in adults with T2D reported improvements in TIR with hybrid/automated closed-loop versus usual insulin therapy. The reported increases varied by study design, baseline control, and duration of follow-up, but commonly fell in the range of roughly **~5–15 percentage points** (which equals about **1.2–3.6 additional hours per day** spent in 70–180 mg/dL). Trials in people with worse baseline control tended to show larger absolute gains.

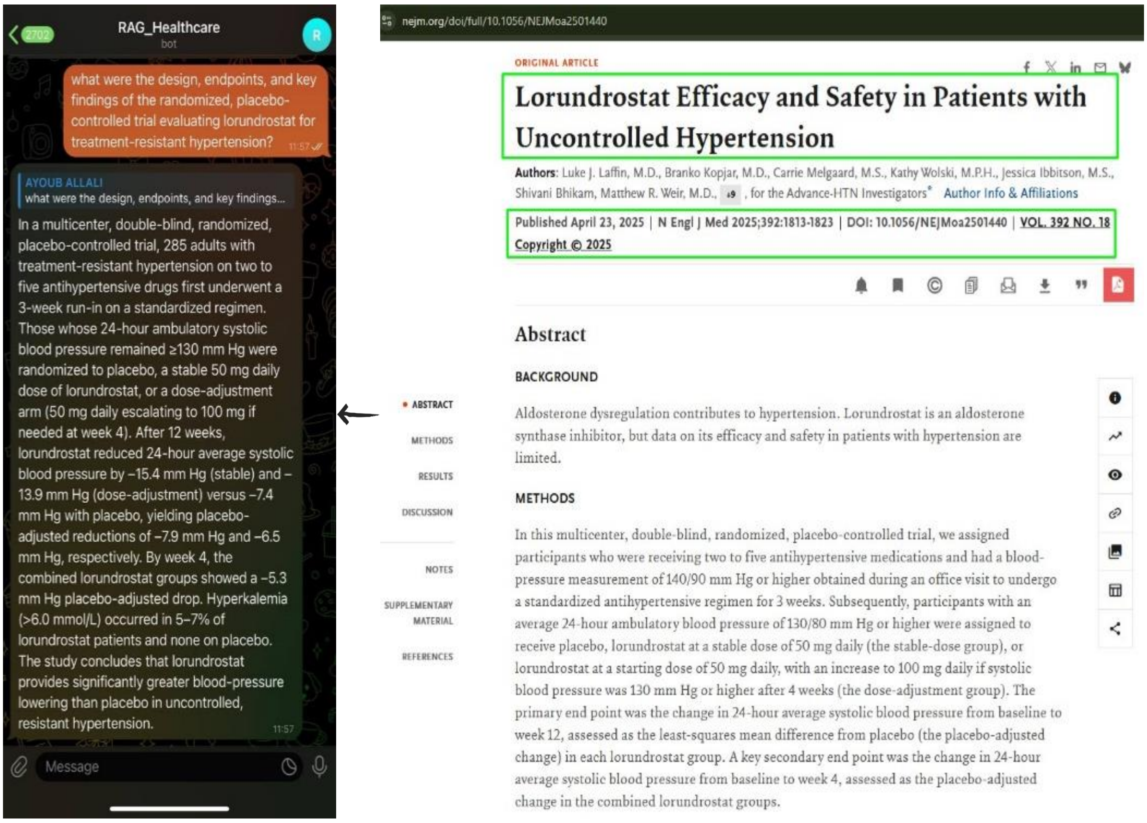
Figure 6. Clinical decision support system powered by BioGPT 5

3.1.2. Patient education and engagement

The system leverages the same semantic retrieval pipeline to translate dense clinical research into clear, layperson-friendly explanations tailored to individual patient concerns. These explanations address specific needs, whether medication side effects, lifestyle modifications, or preventative care. The bot delivers personalized, up-to-date guidance that empowers patients to better understand their conditions and adhere to treatment plans.

3.1.3. Drug discovery and development

Pharmaceutical researchers can harness our platform to navigate vast volumes of trial data, drug–drug interaction reports, and molecular studies. The LLMs, enriched via RAG, highlight promising compound interactions, flag safety signals, and summarize clinical trial outcomes. This dramatically accelerates hypothesis generation and enables more informed decisions on candidate selection and trial design, as shown in Figure 7.



what were the design, endpoints, and key findings of the randomized, placebo-controlled trial evaluating lorundrostat for treatment-resistant hypertension? 11:57

AYOUB ALLALI
what were the design, endpoints, and key findings...

In a multicenter, double-blind, randomized, placebo-controlled trial, 285 adults with treatment-resistant hypertension on two to five antihypertensive drugs first underwent a 3-week run-in on a standardized regimen. Those whose 24-hour ambulatory systolic blood pressure remained ≥130 mm Hg were randomized to placebo, a stable 50 mg daily dose of lorundrostat, or a dose-adjustment arm (50 mg daily escalating to 100 mg if needed at week 4). After 12 weeks, lorundrostat reduced 24-hour average systolic blood pressure by −15.4 mm Hg (stable) and −13.9 mm Hg (dose-adjustment) versus −7.4 mm Hg with placebo, yielding placebo-adjusted reductions of −7.9 mm Hg and −6.5 mm Hg, respectively. By week 4, the combined lorundrostat groups showed a −5.3 mm Hg placebo-adjusted drop. Hyperkalemia (>6.0 mmol/L) occurred in 5–7% of lorundrostat patients and none on placebo. The study concludes that lorundrostat provides significantly greater blood-pressure lowering than placebo in uncontrolled, resistant hypertension. 11:57

nejm.org/doi/full/10.1056/NEJMoa2501440

ORIGINAL ARTICLE

Lorundrostat Efficacy and Safety in Patients with Uncontrolled Hypertension

Authors: Luke J. Laffin, M.D., Branko Kopjar, M.D., Carrie Melgaard, M.S., Kathy Wolski, M.P.H., Jessica Ibbitson, M.S., Shivani Bhikam, Matthew R. Weir, M.D., +9, for the Advance-HTN Investigators* Author Info & Affiliations

Published April 23, 2025 | N Engl J Med 2025;392:1813-1823 | DOI: 10.1056/NEJMoa2501440 | VOL. 392 NO. 18
Copyright © 2025

Abstract

BACKGROUND

Aldosterone dysregulation contributes to hypertension. Lorundrostat is an aldosterone synthase inhibitor, but data on its efficacy and safety in patients with hypertension are limited.

METHODS

In this multicenter, double-blind, randomized, placebo-controlled trial, we assigned participants who were receiving two to five antihypertensive medications and had a blood-pressure measurement of 140/90 mm Hg or higher obtained during an office visit to undergo a standardized antihypertensive regimen for 3 weeks. Subsequently, participants with an average 24-hour ambulatory blood pressure of 130/80 mm Hg or higher were assigned to receive placebo, lorundrostat at a stable dose of 50 mg daily (the stable-dose group), or lorundrostat at a starting dose of 50 mg daily, with an increase to 100 mg daily if systolic blood pressure was 130 mm Hg or higher after 4 weeks (the dose-adjustment group). The primary end point was the change in 24-hour average systolic blood pressure from baseline to week 12, assessed as the least-squares mean difference from placebo (the placebo-adjusted change) in each lorundrostat group. A key secondary end point was the change in 24-hour average systolic blood pressure from baseline to week 4, assessed as the placebo-adjusted change in the combined lorundrostat groups.

Figure 7. Drug discovery and development by GPT 4o mini with RAG

3.1.4. Virtual health assistants

Our Python-driven Telegram bot serves as an intelligent front-line aide, engaging users in real time to triage symptoms or answer routine health inquiries. By coupling natural-language dialogue with instant access to the Milvus-indexed knowledge base, the assistant handles everyday questions autonomously while escalating critical issues to human providers. This approach increases care accessibility and reduces clinician workload.

3.1.5. Mental health support

Through empathetic conversational flows powered by RAG-augmented LLMs, the system offers on-demand mental health check-ins, coping strategies rooted in cognitive-behavioral principles, and early alerts for concerning language patterns. This always-available chat interface provides an additional layer of support. It encourages users to seek further care when needed, while ensuring interventions are informed by the latest psychological research [16].

3.2. Impact of real-time data streaming and distributed storage

The integration of real-time data streaming and distributed storage plays a crucial role in ensuring the efficiency, scalability, and reliability of AI-driven medical applications. Apache Kafka, as a real-time data streaming platform, enables continuous ingestion of professional and scientific medical news from trusted sources such as Medscape, the NEJM, and the Lancet. By leveraging Kafka's publish-subscribe model, the system ensures that newly published medical research is promptly collected, processed, and made available for AI model training and inference. This real-time ingestion capability is critical in the medical domain, where up-to-date knowledge is essential for accurate diagnoses, treatment recommendations, and patient support. Additionally, Hadoop HDFS provides a robust, distributed storage solution that efficiently manages large volumes of medical text data while maintaining high availability and fault tolerance [17].

This architecture enhances the performance of RAG by allowing LLMs to dynamically access up-to-date information, significantly reducing the risk of outdated or incorrect AI-generated responses. Furthermore, the use of a distributed file system enhances scalability, ensuring that the system can handle increasing volumes of medical literature without degradation in performance. The combined power of Kafka's real-time data ingestion and HDFS's distributed storage enables a continuously evolving knowledge base, improving the accuracy and relevance of LLM-generated medical insights [18].

3.3. Effectiveness of RAG in reducing hallucinations

The integration of RAG significantly improves the accuracy and reliability of LLMs by reducing hallucinations—a common issue where AI models generate misleading or incorrect information due to limitations in their pre-trained knowledge. Traditional LLMs rely solely on pre-existing training data, which can become outdated or lack domain-specific details, particularly in dynamic fields like medicine. RAG mitigates this by incorporating a retrieval mechanism that accesses the most recent, contextually relevant medical literature from Hadoop HDFS, powered by real-time data streamed from Apache Kafka, ensuring factually grounded and up-to-date responses.

In our implementation, all three LLMs (DeepSeek-R1, GPT-4o-mini, and Llama 3) demonstrated improved accuracy and contextual relevance when enhanced with RAG, particularly in clinical decision support, drug discovery, and patient education use cases. The models were able to reference the latest medical studies and guidelines, significantly reducing speculative or incorrect responses. Additionally, RAG enhances explainability by allowing models to cite specific documents or sources, thereby improving trustworthiness in critical medical applications [19].

3.4. Performance across different LLMs

The performance of the three fine-tuned LLMs (DeepSeek-R1, GPT-4o-mini, and Llama 3) varied across different medical applications, highlighting the strengths and trade-offs of each model in handling complex healthcare-related queries. GPT-4o-mini demonstrated exceptional natural language fluency, making it particularly effective in-patient engagement, mental health support, and virtual health assistant applications where conversational coherence and emotional intelligence are crucial, its ability to generate context-aware responses with high readability made it the preferred choice for scenarios requiring human-like interaction and empathetic communication. DeepSeek-R1, on the other hand, excelled in medical research-oriented tasks, such as summarizing clinical trial data, extracting key findings from medical literature, and identifying potential drug interactions, its strong analytical capabilities allowed for more structured, information-dense outputs, making it highly suitable for scientific and pharmaceutical research applications. Llama 3 provided a balanced performance across multiple use cases, demonstrating robust contextual understanding in clinical decision support while maintaining reasonable fluency in patient-oriented dialogues, its efficiency in real-time RAG workflows ensured that medical insights were consistently accurate and well-referenced.

Integrating these models with Apache Kafka and Hadoop HDFS enabled continuous updates and fine-tuning using fresh medical data, reducing reliance on static knowledge bases. However, model performance was influenced by computational constraints, with GPT-4o-mini consuming more resources due to its advanced reasoning capabilities, while DeepSeek-R1 and Llama 3 offered better efficiency-accuracy balance. Additionally, handling ambiguous medical queries posed challenges, as differences in each model's training architecture affected document prioritization from HDFS [20], [21].

3.5. Challenges and limitations

Despite the promising results of integrating big data technologies with RAG-enhanced LLMs, several challenges and limitations must be addressed to optimize their real-world application in the healthcare domain. One of the primary challenges is computational resource constraints, as fine-tuning and deploying LLMs on large-scale medical datasets require significant GPU power and memory. While our system utilized an NVIDIA GeForce RTX 3050 Ti, training larger models on high-dimensional medical data remains computationally expensive and time-intensive, limiting the feasibility of real-time fine-tuning for smaller healthcare organizations. Data quality and bias pose critical concerns, as AI models depend on the reliability of their training and retrieval data. Even though medical literature from NEJM was used as a primary data source, inherent biases in clinical research, disparities in patient demographics, and outdated medical findings could potentially skew model predictions [22], [23].

Ensuring data diversity, de-biasing methodologies, and continuous validation by medical professionals is essential to mitigate these risks, another limitation is the retrieval latency within the Apache Hadoop HDFS ecosystem, particularly when handling large-scale unstructured text data, while RAG improves factual accuracy, inefficient retrieval mechanisms could slow down response times, affecting usability in time-sensitive applications like clinical decision support. Regulatory and ethical concerns remain significant barriers to deployment, as AI-generated medical insights must comply with healthcare regulations such as health insurance portability and accountability act (HIPAA) and general data protection regulation (GDPR) to ensure data privacy and security, the risk of misinterpretation and over-reliance on AI-generated recommendations also highlights the need for explainability and interpretability frameworks, allowing doctors to verify AI suggestions before making critical medical decisions. Lastly, handling ambiguous medical queries remains a challenge, as LLMs may struggle with vague symptoms, rare diseases, or conflicting medical opinions, future improvements should focus on hybrid retrieval models combining keyword-based and semantic search techniques, optimized indexing strategies for faster access to stored medical data, and collaborative AI-human decision-making frameworks to maximize reliability, addressing these challenges will be crucial in enhancing the scalability, accuracy, and trustworthiness of AI-driven healthcare solutions in clinical practice [24]–[26].

4. CONCLUSION

This research demonstrates that integrating scalable big data infrastructures with RAG, enhanced LLMs can dramatically improve the relevance, accuracy, and timeliness of AI-driven medical applications by streaming and archiving professional medical news, semantically indexing millions of documents in Milvus, and fine-tuning state-of-the-art LLMs for sub second, evidence-backed insights. To advance this framework, future work must drive ultra-low-latency retrieval through optimized vector indexing and hybrid search strategies, extend model capabilities by fusing multi-modal data such as radiology images, electronic health records, and genomics into a unified embedding space, and embed explainability via explainable artificial intelligence (XAI) modules that trace each recommendation back to its source. Equally critical is ensuring ethical, compliant deployment, implementing privacy safeguards, human-in-the-loop verification, and adherence to HIPAA, GDPR, and emerging AI regulations, to mitigate bias and over-reliance on automation; finally, democratizing access and supporting continual learning across institutions will require scalable, distributed training approaches, including cloud-based platforms, federated learning, and edge-distributed GPU clusters. By addressing these challenges, we can transform today's proof-of-concept into a globally deployable, real-time, and trustworthy AI-powered medical decision-support ecosystem that elevates patient care and accelerates biomedical discovery.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ayoub Allali	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Ibtihal Abouchabaka		✓				✓			✓	✓	✓			
Najat Rafali	✓	✓	✓	✓	✓		✓	✓		✓		✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in NEJM RSS FEED at https://www.nejm.org/rss-feed/.

REFERENCES

[1] C. Burgan, J. Kowalski, and W. Liao, "Developing a retrieval augmented generation (RAG) chatbot app using adaptive large language models (LLM) and LangChain framework," *Proceedings of the West Virginia Academy of Science*, vol. 96, no. 1, 2024, doi: 10.55632/pwvas.v96i1.1068.

[2] Y. Mao, X. Dong, W. Xu, Y. Gao, B. Wei, and Y. Zhang, "FIT-RAG: black-box RAG with factual information and token reduction," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–27, Mar. 2025, doi: 10.1145/3676957.

[3] D.-I. M. Schiele, Y. Gittmann, S. Ilchmann, A. Gojsalić, D. Jurinčić, and P. Klempt, "Voting advice applications: implementation of RAG-supported LLMs," *TechRxiv*, Jul. 2024, doi: 10.36227/techrxiv.172115156.64500701/v1.

[4] K. Singhal *et al.*, "Toward expert-level medical question answering with large language models," *Nature Medicine*, vol. 31, no. 3, pp. 943–950, Mar. 2025, doi: 10.1038/s41591-024-03423-7.

[5] A. Y. Alan, E. Karaarslan, and Ö. Aydın, "Improving LLM reliability with RAG in religious question-answering: MufassirQAS," *Turkish Journal of Engineering*, vol. 9, no. 3, pp. 544–559, Jul. 2025, doi: 10.31127/tuje.1624773.

[6] Z. Xiao, X. He, H. Wu, B. Yu, and Y. Guo, "EDA-Copilot: a RAG-powered intelligent assistant for EDA tools," *ACM Transactions on Design Automation of Electronic Systems*, vol. 30, no. 6, pp. 1–24, Nov. 2025, doi: 10.1145/3715326.

[7] K. Soman *et al.*, "Biomedical knowledge graph-optimized prompt generation for large language models," *Bioinformatics*, vol. 40, no. 9, Sep. 2024, doi: 10.1093/bioinformatics/btae560.

[8] D. C.-Nieves and L. G.-Forte, "Human-centered AI for migrant integration through LLM and RAG optimization," *Applied Sciences*, vol. 15, no. 1, Dec. 2024, doi: 10.3390/app15010325.

[9] A. Mansurova, A. Mansurova, and A. Nugumanova, "QA-RAG: exploring LLM reliance on external knowledge," *Big Data and Cognitive Computing*, vol. 8, no. 9, Sep. 2024, doi: 10.3390/bdcc8090115.

[10] X. Zhao, X. Zhou, and G. Li, "Chat2Data: an interactive data analysis system with RAG, vector databases and LLMs," *Proceedings of the VLDB Endowment*, vol. 17, no. 12, pp. 4481–4484, Aug. 2024, doi: 10.14778/3685800.3685905.

[11] J. S. Jauhiainen and A. G. Guerra, "Evaluating students' open-ended written responses with LLMs: using the RAG framework for GPT-3.5, GPT-4, Claude-3, and Mistral-Large," *Advances in Artificial Intelligence and Machine Learning*, vol. 4, no. 4, pp. 3097–3113, 2024, doi: 10.54364/AAIML.2024.44177.

[12] K. Fang, C. Tang, and J. Wang, "Evaluating simulated teaching audio for teacher trainees using RAG and local LLMs," *Scientific Reports*, vol. 15, no. 1, Jan. 2025, doi: 10.1038/s41598-025-87898-5.

[13] S. Vidivelli, M. Ramachandran, and A. Dharunbalaji, "Efficiency-driven custom chatbot development: unleashing LangChain, RAG, and performance-optimized LLM fusion," *Computers, Materials & Continua*, vol. 80, no. 2, pp. 2423–2442, 2024, doi: 10.32604/cmc.2024.054360.

[14] Y. Wang, S. Leutner, M. Ingrisch, C. Klein, L. C. Hinske, and K. Danhauser, "Optimizing data extraction: harnessing RAG and LLMs for German medical documents," *Digital Health and Informatics Innovations for Sustainable Health Care Systems*, vol. 316, pp. 949–950, Aug. 2024, doi: 10.3233/SHTI240567.

[15] R. S. M. Wahidur, S. Kim, H. Choi, D. S. Bhatti, and H.-N. Lee, "Legal query RAG," *IEEE Access*, vol. 13, pp. 36978–36994, 2025, doi: 10.1109/ACCESS.2025.3542125.

[16] A. Allali, N. Bouanani, I. Abouchabaka, and N. Rafalia, "Advancing elderly care through big data analytics and machine learning for daily activity characterization," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 36, no. 3, pp. 1969–1975, Dec. 2024, doi: 10.11591/ijeecs.v36.i3.pp1969-1975.

[17] M. Son, Y.-J. Won, and S. Lee, "Optimizing large language models: a deep dive into effective prompt engineering techniques," *Applied Sciences*, vol. 15, no. 3, Jan. 2025, doi: 10.3390/app15031430.

[18] K. E. Kannammal, M. R. K. Anirudh , K. P. Tamizhiniyal, G. Ganishkar, and C. Adrinath, "Fin-Rag a Rag system for financial documents," *International Journal of Innovative Science and Research Technology*, vol. 10, no. 4, pp. 1761–1767, Apr. 2025, doi: 10.38124/ijisrt/25apr1147.

[19] P. Pany, "Reasoning engine with pre-trained LLMs: an operation GPT," *International Journal for Research in Applied Science and Engineering Technology*, vol. 13, no. 4, pp. 2452–2463, Apr. 2025, doi: 10.22214/ijraset.2025.68761.

[20] J. Wang *et al.*, "Hierarchical index retrieval-driven wireless network intent translation with LLM," *IEEE Transactions on Mobile Computing*, vol. 24, no. 10, pp. 9837–9851, Oct. 2025, doi: 10.1109/TMC.2025.3564937.

[21] A. Sghir, A. Allali, N. Rafalia, and J. Abouchabaka, "Advanced strategies for big data resource and storage optimization: an AI perspective," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 8, 2025, doi: 10.14569/IJACSA.2025.0160896.

[22] C. Carpenter, "Zero-shot learning with large language models enhances drilling-information retrieval," *Journal of Petroleum Technology*, vol. 77, no. 1, pp. 92–95, Jan. 2025, doi: 10.2118/0125-0092-JPT.

[23] A. Allali, Z. E. Falah, A. Sghir, J. Abouchabaka, and N. Rafalia, "A comparative analysis of GPUs, TPUs, DPUs, and QPUs for deep learning with python," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 38, no. 2, pp. 1324–1330, May 2025, doi: 10.11591/ijeecs.v38.i2.pp1324-1330.

[24] W. Bi *et al.*, "Leveraging the dual capabilities of LLM: LLM-enhanced text mapping model for personality detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 22, pp. 23487–23495, Apr. 2025, doi: 10.1609/aaai.v39i22.34517.

[25] V. Malik, "Hadoop distributed file system (HDFS) with its architecture," *International Journal for Research in Applied Science and Engineering Technology*, vol. 13, no. 5, pp. 6031–6034, May 2025, doi: 10.22214/ijraset.2025.71584.

[26] S. Awasthi and N. Kohli, "Hybrid encryption for fortifying HDFS data," *International Journal of Basic and Applied Sciences*, vol. 14, no. 5, pp. 436–454, Sep. 2025, doi: 10.14419/m46fn971.

BIOGRAPHIES OF AUTHORS

Ayoub Allali has pursued an academic and research career in the fields of Computer Science and Big Data. He completed his bachelor's degree in Mathematical and Computer Sciences from the Department of Computer Science at the Faculty of Sciences, Ibn Tofail University in Kenitra, Morocco, in 2019. Continuing his education at the same institution, he earned a master's degree in Big Data and Cloud Computing in 2021. Since 2022, he has been a doctoral student at the Computer Research Laboratory (LaRI) within the Department of Computer Science at the Faculty of Sciences, Ibn Tofail University. His research interests are focused on big data and deep learning, indicating a strong commitment to advancing knowledge in these critical areas of data science. He can be contacted at email: ayoub.allali@uit.ac.ma.

Ibtihal Abouchabaka is a 5th year medical student at the Mohammed VI University of Health Sciences in Casablanca. In addition, she is an external doctor at the Cheikh Khalifa International University Hospital in Casablanca. She can be contacted at email: Ibtihal1002@gmail.com.

Prof. Dr. Najat Rafalia has established herself as a prominent figure in the fields of Computer Science and Applied Mathematics. Her academic journey began at Mohammed V University in Rabat, Morocco, where she received her bachelor's degree in Applied Mathematics from the Department of Mathematics, Faculty of Sciences, in 1992. She continued at the same institution to earn a postgraduate diploma (DEA) in Computer Science in 1994 and later a doctorate in Computer Sciences from the Department of Computer Science in 1997. Since 1997, she has served as a professor in the Department of Computer Sciences at Ibn Tofail University. Further advancing her expertise, she had her postdoctoral thesis in 2013 and completed her third doctorate in Computer Sciences at Ibn Tofail University in Kenitra, Morocco, in 2017. Her research interests are deeply rooted in big data analytic, artificial intelligence and its applications, internet of things, distributed systems, multi-agent systems, concurrent and parallel programming. Throughout her career, she has made significant contributions to her field, authoring mor than 70 papers and completing 3 theses. She can be contacted at email: najat.rafalia@uit.ac.ma.