

AI-Driven Search for Infectious Disease Research, Streamlining Information Access and Discoverability: MPOX Case Study

Leonard Boante Mensah, Samuel Driver, Daniela Moraes Morelli, Trudie Lang,
Alexis Segrt

Submitted to: Online Journal of Public Health Informatics
on: April 03, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
---------------------------------	----------

Preprint
JMIR Publications

AI-Driven Search for Infectious Disease Research, Streamlining Information Access and Discoverability: MPOX Case Study

Leonard Boante Mensah¹; Samuel Driver²; Daniela Moraes Morelli³; Trudie Lang²; Alexis Segrt²

¹ Africa Region Knowledge Hub The Global Health Network, Oxford University Oxford GB

² The Global Health Network, Oxford University Oxford GB

³ Latin America and Caribbean Region Knowledge Hub The Global Health Network, Oxford University Oxford GB

Corresponding Author:

Alexis Segrt

The Global Health Network, Oxford University
Ewert House, Banbury Rd
Summertown
Oxford
GB

Abstract

Background: The exponential growth of scientific literature presents significant challenges in infectious disease outbreak management, where rapid and reliable information retrieval is crucial for evidence-based decision-making. Traditional search methods struggle to synthesize heterogeneous and rapidly evolving data sources, leading to information overload and delays in integrating critical findings into clinical and public health strategies. The AI-driven system was developed by The Global Health Network (TGHN), to ensure accessibility and applicability of resources to strengthen global health research. The monkeypox (MPOX) outbreak exemplifies these challenges, highlighting the need for AI-driven solutions to streamline knowledge discovery and access.

Objective: Present the development and evaluation a Retrieval Augmented Generation (RAG)-based AI system for optimizing information retrieval, synthesis, and dissemination in the context of infectious disease research, using MPOX as a case study.

Methods: A descriptive case study methodology was employed to develop a pilot AI-driven system that integrates semantic retrieval and generative capabilities. The AI-driven system is designed to enhance the accessibility and relevance of scientific literature, leveraging state-of-the-art natural language processing (NLP) and machine learning techniques to provide contextually accurate, real-time responses. The AI-driven system architecture follows a microservices-based modular design. The data pipeline incorporates sources such as PubMed, medRxiv, bioRxiv, WHO, CDC reports, and curated social media feeds. Performance evaluation included: quantitative metrics and qualitative review by infectious disease specialists.

Results: The system successfully integrated traditional keyword-based retrieval with vector-based semantic search, significantly improving the relevance and contextual accuracy of retrieved information. The hybrid search model demonstrated enhanced document ranking and retrieval precision, validated through standard information retrieval metrics. The human-in-the-loop evaluation confirmed that the generated responses aligned with expert-validated sources, ensuring scientific accuracy and usability for public health professionals. The pilot implementation on MPOX research indicated the system's scalability and adaptability to various infectious disease domains.

Conclusions: AI-driven RAG architectures offer a transformative approach to knowledge discovery and synthesis in infectious disease research. By integrating retrieval-based document indexing with generative AI, this system provides a scalable, real-time, and evidence-based solution for researchers, clinicians, and policymakers. The study underscores the need for further development to enhance multilingual support, real-time data integration, and ethical AI governance. The framework has the potential to revolutionize global outbreak preparedness, facilitating rapid, reliable, and transparent access to critical health information. Clinical Trial: Not applicable.

(JMIR Preprints 03/04/2025:75436)

DOI: <https://doi.org/10.2196/preprints.75436>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/>

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Title

AI-Driven Search for Infectious Disease Research, Streamlining Information Access and Discoverability: MPOX Case Study

Introduction

The exponential growth of scientific publications has transformed the global dissemination of knowledge. However, an excess of information can hinder decision-making in both clinical and public health contexts, particularly during health emergencies such as infectious disease outbreaks. Data saturation can impede the synthesis of valid and reliable information, thereby affecting decision-making in both clinical settings and public health policy formulation [1].

Rapidly evolving outbreaks, ranging from emerging viruses to re-emerging diseases, require researchers, clinicians, and public health officials to remain up to date with the latest scientific developments. However, the overwhelming volume of scientific literature, official reports, and real-time discussions can lead to information overload, delaying and complicating decision-making in key areas such as treatment protocol development, resource allocation, public health communication, and policy formulation [2].

A major challenge in information management during epidemiological outbreaks is information overload. The excessive volume of data complicates the identification of relevant studies, thereby delaying the integration of the best available evidence into clinical practice and public health policies [2]. Furthermore, the quality of available evidence can significantly impact the effectiveness of response strategies [3], potentially leading to decisions based on incomplete or low-quality information [4,5].

Over the past decade, the volume of publications on infectious diseases has increased dramatically, with thousands of new studies published across various sources, including peer-reviewed journals, preprint repositories (e.g., bioRxiv, medRxiv), and reports from governmental agencies such as the WHO – World Health Organization (WHO) and Centers for Disease Control and Prevention (CDCs) [3]. Additionally, discussions on social media platforms and news outlets contribute to the evolution of the information landscape. However, the heterogeneity, inconsistency, and variable reliability of these sources pose significant challenges to traditional search and retrieval methods [6].

Monkeypox (MPOX) exemplifies the broader challenges in infectious disease research [7,8]. Historically confined to specific regions, recent outbreaks have demonstrated the virus's potential for wider transmission and endemicity [9]. Over the past five years, the incidence of monkeypox (mpox) in Latin America has seen a notable increase, particularly since 2022, raising public health concerns and highlighting the need for enhanced epidemiological surveillance [10].

According to the Pan American Health Organization (PAHO), between January 2022 and September 2024, a total of 109,699 confirmed cases of mpox were reported worldwide. Furthermore, in September 2024, there was an 8% increase in new cases compared to August of the same year, reflecting the persistence of transmission. During this period, the Region of the Americas accounted for 15.5% of the reported cases, making it one of the most affected areas [10,11].

A 2024 epidemiological report highlights the resurgence of mpox, linked to the emergence of a new viral clade (clade Ib), whose rapid spread has significantly impacted several regions, including Latin America [12]. This development underscores the importance of strengthening prevention strategies, early detection, and rapid response efforts to mitigate the spread of the virus across the Latin America region.

Researchers must track updates spanning epidemiological data, clinical guidelines, vaccine trials, and genomic studies across multiple platforms. Delays in synthesizing this dispersed information can undermine outbreak containment, hamper clinical care, and slow research progress.

Implementing efficient mechanisms for systematic information synthesis and access is crucial to strengthening evidence-based decision-making in healthcare [13].

Given these complexities, systematic evidence synthesis and mechanisms for ensuring efficient access to validated, high-quality information are critical to reinforcing evidence-based decision-

making in healthcare. Artificial Intelligence (AI)-driven approaches offer a promising solution by automating the ingestion, retrieval, and synthesis of heterogeneous data sources, thereby streamlining the discovery of actionable insights during outbreaks. To address these challenges, The Global Health Network (TGHN) serves as a collaborative open-access platform that facilitates equitable access to research and knowledge-sharing among scientists, healthcare professionals, and public health policymakers [14]. Its mission is to bridge the gap in global health research by providing open-access resources, training, and collaborative tools, particularly for those working in resource-limited settings [14]. Leveraging TGHN's infrastructure and knowhow, developed an AI-driven system within the TGHN framework to enhance the accessibility and applicability of infectious disease research, ensuring that cutting-edge scientific insights reach those who need them most [15].

Method: Case study, descriptive, on the pilot development of a system to optimise information access and accelerate knowledge dissemination, using Retrieval Augmented Generation (RAG) [6, 16], applied to the Monkeypox topic.

Generalized AI-Powered System Architecture

We present a RAG-based pipeline suitable for various infectious diseases scenarios. Key components include data ingestion, preprocessing, vector database indexing, and a generative model capable of returning context rich responses. While MPOX AI serves as the demonstrative case study, the architecture accommodates a broader spectrum.

The system architecture integrates several components designed to optimize data handling and query processing. The NGINX server functions as a reverse proxy, managing user requests and directing them to appropriate backend services. The API server and web server handle user interactions and API endpoints, processing user queries and facilitating communication with the backend. An open-source vector database known as Vespa is central to both the keyword-based and semantic searches, ensuring the efficient retrieval of relevant documents. We leveraged on Postgres database to manage structured data, supporting relational queries, while task queue oversees asynchronous operations such as data ingestion, preprocessing, and continuous updates to the knowledge base. We used an LLM model for question and answering to generate context-aware, coherent responses based on retrieved documents.

System Design

The system is designed to ingest, process and retrieve information from a heterogeneous corpus of documents, providing tailored responses to users with varying levels of expertise and information needs. The design emphasizes modularity, extensibility, and rigorous evaluation at each stage (Figure 1).

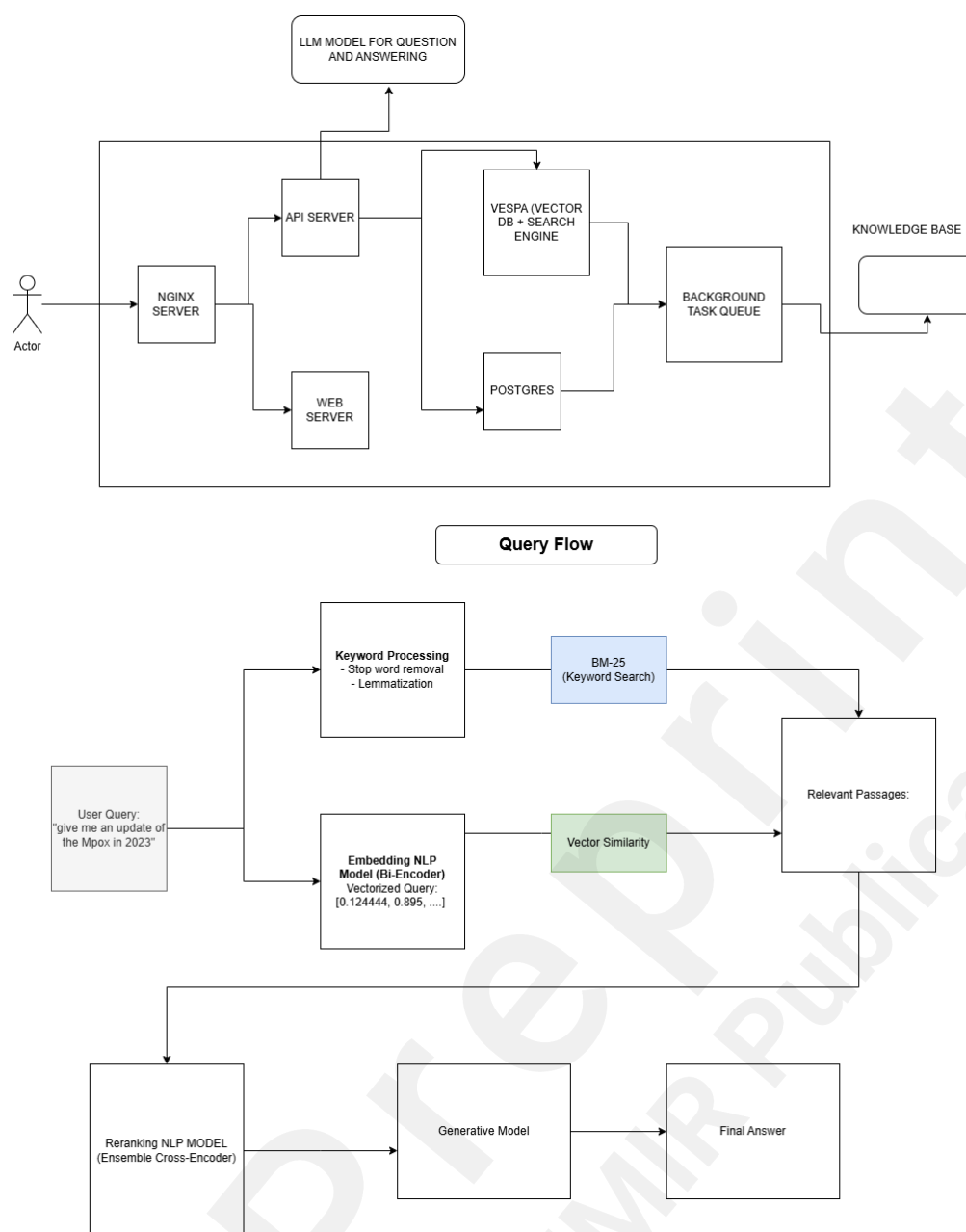


Figure 1 Architecture Overview and Query Flow

Data Collection and Preprocessing

Data collection begins with web scraping tools and APIs to extract textual content from a wide range of sources. Academic and pre-print servers such as PubMed, medRxiv, and bioRxiv form the bedrock of scientific evidence, providing peer-reviewed publications and rapidly disseminated pre-publication research findings. Automated crawlers are implemented to continuously monitor these sources for new publications, ensuring the knowledge base is constantly updated with the latest scientific advancements. Integration must address the varying levels of quality and validation inherent in pre-print literature compared to peer-reviewed publications, potentially through a scoring or flagging mechanism.

Official reports from authoritative organizations like the WHO, CDC, and local Health Ministries provide critical public health guidelines, epidemiological data, and policy recommendations. Structured data extraction and normalization processes are employed to integrate these reports effectively, alongside mechanisms for tracking updates and revisions to policy documents.

Recognizing the role of public discourse and rapid information sharing during outbreaks, carefully filtered news and social media feeds are incorporated. This necessitates robust filtering mechanisms to mitigate misinformation and noise, focusing on reputable news sources and potentially employing sentiment analysis and credibility scoring to assess the reliability of information. Ethical considerations and privacy implications are paramount in the use of social media data and must be rigorously addressed.

The extracted data undergoes comprehensive cleaning and preprocessing, which includes the removal of HTML tags, handling of special characters, and the splitting of extensive content into manageable chunks or paragraphs to enhance searchability. Chunking ensures that the data is divided into coherent units, making the subsequent retrieval process more efficient. Named Entity Recognition (NER) techniques are employed to identify key entities such as diseases, treatments, and geographic locations. This step is critical for enriching the metadata and improving the precision of search results. The preprocessing pipeline also includes tokenization, lemmatization, and, if necessary, language translation to create a unified dataset that can be effectively indexed and queried.

Semantic Embedding and Indexing

Once the data is preprocessed, it is converted into high-dimensional vector representations using advanced NLP models. Model selection is crucial, with options like BERT, RoBERTa, DistilBERT, and Universal Sentence Encoder offering robust semantic understanding capabilities. Each document chunk is transformed into a vector that captures its semantic essence.

These vectors are stored and indexed using vector similarity search in Vespa, enabling efficient nearest-neighbor searches. Alongside the vectors, metadata such as the original URL, article title, publication date, and other relevant information is stored. This metadata is essential for retrieving and displaying comprehensive search results to users.

Retrieval Augmented Generation Pipeline

The Retrieval Augmented Generation (RAG) pipeline forms the core intelligence engine of the system, enabling both accurate information retrieval and coherent, synthesized responses. This pipeline integrates state-of-the-art retrieval and generation techniques to deliver precise and contextually relevant information.

As illustrated in Figure 1, a hybrid retrieval strategy is proposed, combining traditional BM-25 keyword-based search with vector similarity search. BM-25, known for its efficiency and effectiveness in keyword based relevance ranking, can provide a rapid initial filtering of the corpus. Subsequently, vector similarity search, using the pre-computed embeddings, refines the search space by identifying documents that are semantically closest to the vectorized query. This hybrid approach aims to leverage the strengths of both methods, potentially improving both retrieval speed and relevance. The weighting and integration of BM-25 and vector similarity scores require careful tuning and evaluation.

User queries are vectorized using the same embedding model as the document corpus depicted in Figure 1. This ensures semantic consistency between queries and documents, allowing for meaningful similarity comparisons in the vector space.

The retrieved documents are provided as context to the generative model. The model is explicitly trained to ground its responses in this retrieved context, ensuring that answers are evidence-based and traceable back to the source documents. The inclusion of citations or references within the generated responses, as indicated in the design, is paramount for transparency and allows users to verify the provenance of information and delve deeper into the source literature.

To further enhance the quality of the generated answer, a reranking module, utilizing an ensemble of cross-encoders (as shown in Figure 1), is incorporated. Cross-encoders, while computationally more expensive than bi-encoders, can perform more nuanced and context-aware relevance scoring by jointly attending to the query and each candidate document. Reranking the top-retrieved documents with a cross-encoder before feeding them to the generative model can potentially improve the relevance and coherence of the generated context and, consequently, the final answer.

The final step involves the generative model, which synthesizes information from the top-ranked documents to produce comprehensive, coherent responses. These responses are enriched with citations and links to the original sources, providing users with traceable and verifiable information.

Post-Processing and Feedback Loop

Post-processing techniques are employed to further refine search results. Users are provided with filtering options based on metadata such as publication date, and mechanisms are in place to boost popular or high-importance articles in the search results.

A feedback loop is integral to the system's continuous improvement. User interactions, such as clicks on search results, are tracked to understand which results are most relevant. This feedback data is periodically used to fine-tune the model, enhancing search accuracy and relevance over time.

Ethical Considerations

This study did not involve human participants, patient data, or personally identifiable information. The development and evaluation of the AI-powered information retrieval system were conducted using publicly and open access available datasets from scientific literature, governmental health agencies, and open-access repositories. The guidelines and the research ethics policies, this case study does not require formal ethics committee approval.

Pilot implementation – MPOX topic

The *artificial intelligence-based question-answering system Table 1 (Annex 1 – Supplement material)* was implemented in a pilot applied to the MPOX topic to evaluate its performance, accuracy, and usability in the context of infectious disease research. The architecture follows a modular and scalable approach, based on microservices deployed in containers, enabling flexible integration with various data sources and information retrieval methodologies.

Table 1: Modern, open-source, based on artificial intelligence-based question-answering system

Component	Technology	Justification
Front-End	Next.js (React)	Server-side rendering (SSR) and static site generation (SSG) capabilities for improved performance and SEO. Component-based architecture facilitates code reusability and maintainability. React's ecosystem provides extensive libraries for UI development.
Reverse Proxy	Nginx	High-performance web server and reverse proxy. Handles SSL/TLS termination, load balancing, and static content serving, improving security and responsiveness.
API Gateway	FastAPI	Modern, high-performance web framework. Supports asynchronous request handling, crucial for interacting with the LLM and vector database. Built-in data validation (via Pydantic) ensures API robustness.
Vector Database & Search	Vespa	Hybrid search engine supporting both keyword-based and vector-based retrieval. Optimized for low-latency, high-throughput similarity search, crucial for efficient retrieval of relevant documents.
Relational Database	PostgreSQL	<i>Robust, open-source relational database. Provides ACID properties for reliable data management. Used for storing user data, metadata, and system configurations.</i>
Large Language Model	LLAMA	A powerful, pre-trained LLM capable of generating

		coherent and contextually relevant answers.
Task Queue	Celery	Distributed task queue for asynchronous processing of long-running tasks.
Message Broker	Redis	Facilitates communication between the API server and Celery workers.
Containerization	Docker	Provides a consistent and portable runtime environment for all system components. Simplifies deployment and scaling
Orchestration	Docker Compose	Docker Compose is used for local development and smaller-scale deployments.

The system evaluation was conducted through quantitative and qualitative metric analysis, considering the relevance of the generated responses, information retrieval time, and user perception regarding the system's utility. Internal validation approaches were employed to measure the coherence and accuracy of the results, comparing the system's responses with reference sources from scientific literature and official public health reports.

The system follows a microservices architecture, with each component running as an independent service within a Docker container. This promotes modularity, fault isolation, and independent scaling. The key components and their interactions are described in the *Table 2 (Annex 1 – Supplement material)*.

Table 2: System microservices architecture: key components

Front-End (Next.js)	The user interface is implemented as a Next.js application, providing a responsive and interactive user experience. It handles user authentication, query input, and result display, communicating with the API Server via HTTP requests.
Reverse Proxy (Nginx)	Nginx acts as a gateway, routing incoming requests to the appropriate service (Front-End or API Server). It also handles SSL/TLS termination, improving security and performance. It can be configured to provide load balancing and caching, enhancing scalability and resilience.
API Server (FastAPI)	This is the central orchestration layer, handling user authentication, query processing, document retrieval, and answer generation. It implements RESTful endpoints for all system functionalities. The key steps in the query processing pipeline include: • Request Handling: Receives the user query from the Front-End. • Authentication and Authorization: Verifies user credentials and permissions (using JWT or similar). • Query Preprocessing: Applies preprocessing steps like tokenization, lemmatization, and NER. • Retrieval Request: Sends a request to Vespa with the preprocessed query. • Response Processing: Receives the top-k relevant documents from Vespa. • Generation Request: Constructs a prompt for the LLM with the query and retrieved documents. • Response Generation: Receives the generated answer from the LLM. • Response Formatting and Return: Formats the answer with citations and returns it to the Front-End.
Large	The LLM generates natural language answers based on the user query

Language Model (LLM)	and retrieved context. Open-source LLaMA models were used for this implementation, chosen for their strong performance in natural language understanding and generation tasks. These models were fine-tuned to adapt to the domain-specific requirements of infectious disease research, ensuring the generated answers are coherent, contextually relevant, and scientifically accurate. The LLaMA models can be deployed as self-hosted services, offering greater control over data privacy and latency compared to third-party APIs. This setup allows for flexible scaling and customization to meet specific project needs.
Vespa (Search & Vector Database)	Vespa serves as the core search and retrieval engine, responsible for storing document embeddings and executing approximate nearest neighbor (ANN) searches to identify semantically similar documents. The system leverages Vespa's hybrid retrieval capabilities, combining traditional keyword-based search techniques with advanced vector similarity searches to improve both accuracy and relevance. The keyword-based search, powered by BM-25, efficiently filters documents based on textual relevance, while vector similarity search refines these results using semantic embeddings. Vespa's support for real-time indexing and query execution allows for low-latency retrieval, making it highly effective for dynamic datasets. Additionally, Vespa's scalability features facilitate handling large volumes of documents, while its support for custom ranking functions enables fine-tuning of retrieval strategies to match specific domain requirements.
PostgreSQL	PostgreSQL serves as the backbone for managing relational data within the system. It stores user data, including authentication credentials, preferences, and activity logs, ensuring secure access and personalized experiences. System configurations, such as API keys, service endpoints, and access control lists, are maintained within PostgreSQL to facilitate centralized management and consistent behavior across microservices. Additionally, PostgreSQL handles document metadata, including document IDs, source information, publication dates, and indexing statuses, enabling efficient query execution and robust data tracking. The database's ACID (Atomicity, Consistency, Isolation, Durability) properties guarantee data integrity and reliable transactional support, making it well-suited for applications requiring high availability and fault tolerance. Advanced indexing techniques and optimized query execution plans further enhance performance, supporting complex queries and large-scale data operations.
Background Task Queue (Celery)	Celery handles asynchronous tasks like document ingestion and embedding generation, allowing the API Server to remain responsive. The document ingestion workflow involves: • Task Submission: New document submitted for ingestion. • Task Enqueueing: API Server enqueues the task. • Task Execution: Celery worker picks up and processes the task. • Preprocessing: Cleaning, text extraction, NER, and normalization. • Embedding Generation: Vector embedding generated using a pre-trained model. • Indexing: Document and embedding sent to Vespa for indexing. • Metadata Storage: Relevant metadata stored in PostgreSQL.

Deployment and Infrastructure	The system is deployed on Google Cloud Platform (GCP), leveraging its robust infrastructure for scalability, reliability, and performance. Containerization using Docker ensures consistent environments across development, testing, and production stages. Kubernetes, managed via Google Kubernetes Engine (GKE), automates deployment, scaling, and management of the containerized microservices. This setup provides dynamic resource allocation, high availability, and seamless integration with GCP's suite of monitoring and security tools, enabling efficient management of computational resources and ensuring the system can handle varying workloads effectively.
Security Considerations	• Authentication and Authorization: Implemented using OAuth 2.0 and JWT, with role-based access control (RBAC). • Data Encryption: Data encrypted in transit (HTTPS) and at rest (database encryption). • Input Validation: All user inputs validated to prevent security vulnerabilities.

Data Analysis

The evaluation of the AI-driven MPOX research application employed a combination of quantitative metrics and qualitative feedback to assess the system's performance and practical utility. Standard information retrieval metrics such as precision, recall, F1-score, and Mean Reciprocal Rank (MRR) were utilized to measure the effectiveness of the retrieval mechanisms. These metrics provided insights into the system's ability to retrieve relevant documents and rank them appropriately accurately.

In addition to quantitative metrics, human-in-the-loop evaluations were conducted to capture domain-specific nuances that automated metrics might overlook. Experts in infectious disease research and public health reviewed the generated responses, assessing their clarity, coherence, and factual accuracy. This dual approach ensured a comprehensive evaluation, balancing technical performance with real-world applicability.

Results

The system leverages a sophisticated query flow that integrates both traditional and contemporary natural language processing (NLP) techniques to enhance the retrieval and generation of information. The process begins with keyword processing, where stop words are removed, and lemmatization is applied to standardize user queries, ensuring consistency and improving search relevance. Following this, the BM-25 algorithm—a classical retrieval model—identifies potential matches by analyzing keyword frequency and contextual relevance within the indexed documents.

To move beyond simple keyword matching, the system employs an embedding NLP model based on a Bi-Encoder architecture. This model converts the processed queries into vector representations, facilitating semantic similarity searches. These vectorized queries are then compared against a vector database to identify documents that are contextually aligned with the user's intent. To further refine the search results, a re-ranking process is conducted using an ensemble cross-encoder NLP model, which prioritizes the most pertinent passages based on deeper semantic relationships.

Once the most relevant documents are identified and ranked, a generative model synthesizes the extracted information to produce comprehensive, coherent answers. This model not only compiles relevant data but also ensures that the final response is contextually accurate and informative. The output provided to the user includes detailed answers enriched with citations and direct links to the original sources, thus facilitating further exploration, verification, and validation of the information.

The effectiveness of this query flow is validated through rigorous evaluation metrics, including precision, recall, F1-score, and Mean Reciprocal Rank (MRR). Beyond these quantitative measures, human-in-the-loop evaluations were conducted to assess the system's practical applicability and the

quality of generated responses. Experts in infectious disease research and public health assessed the relevance, clarity, and completeness of the system's outputs, providing qualitative feedback that informed iterative improvements.

The AI-driven pilot system [15] can be explored on The Global Health Network platform through the following open-access link MPOX AI [17]

Discussion

The imperative for rapid knowledge dissemination during infectious disease outbreaks is underscored by recent global health crises like Ebola, Zika, COVID-19, etc [18]. Delays in accessing and integrating evolving research can significantly hinder the effectiveness of treatment protocols, public health interventions, and policy decisions. Studies have demonstrated a direct correlation between the speed of knowledge translation and improved outbreak control, highlighting the time-sensitive nature of evidence-based decision-making in these contexts [1]. In rapidly evolving situations, the ability to quickly synthesize and apply the latest findings is paramount for mitigating disease spread and optimizing patient outcomes. The advent of an AI-driven information retrieval system offers a potential paradigm shift, enabling stakeholders to surface critical insights in near real-time and supporting proactive and adaptive outbreak management strategies (Hussain Khan et al., n.d.). Furthermore, the integration of AI techniques, such as machine learning and natural language processing, into infectious disease management has shown significant promise in accelerating the detection and response to outbreaks, as evidenced during the COVID-19 pandemic, where AI facilitated rapid genome sequencing and real-time surveillance [3].

AI technologies, particularly those based on Large Language models (LLMs) such as ChatGPT, have demonstrated substantial potential in enhancing information retrieval processes in healthcare. Yousif M et al. conducted a comparative analysis of ChatGPT and Google Bard, revealing that these AI tools can provide rapid access to critical information, which is essential for healthcare professionals, particularly in high-stakes environments like nursing [2]. The ability of these models to generate coherent and contextually relevant responses can significantly aid healthcare providers in making informed decisions regarding patient care and treatment options.

The Retrieval Augmented Generation technique represents a significant advancement in AI-driven search technologies, combining generative capabilities with the retrieval of external knowledge. This hybrid approach enhances the accuracy and relevance of information provided to healthcare professionals. According to a study by Lewis et al., n.d., integrating AI in healthcare communication has improved the precision of information dissemination. However, it also raises concerns regarding data security and patient confidentiality [6]. The potential for inaccuracies in content generation necessitates ongoing scrutiny and validation of AI outputs to maintain trust in these systems.

Furthermore, the role of AI in enhancing patient safety cannot be overlooked. The systematic review by Radaelli et al [19] indicates that AI can significantly reduce the incidence of HAIs by providing healthcare professionals with timely and accurate information regarding infection control measures. This proactive approach to patient safety underscores the potential of AI-driven search technologies to transform infectious disease management.

Although RAG architectures stand poised to revolutionize infectious disease data management, a number of open questions remain regarding system scalability, the integration of real-time updates from multiple data silos, and the reliability of outputs in critical clinical or policy scenarios. Traditional AI-driven outbreak tools have tended to concentrate on epidemiological modeling—particularly in forecasting disease trajectories—without offering robust solutions for live, evidence-based literature synthesis [20]. This leaves policy makers and clinicians reliant on partial or outmoded information when rapid decisions are essential. Other sophisticated models, including certain large-scale generative transformers, can produce fluent text but often lack transparent citations to source materials, thereby undermining trust and verifiability. In high-stakes environments, confidence in generative outputs is contingent upon the model's capacity to clearly anchor its statements to authoritative evidence, especially when formulating guidance on emergent

treatments or public health measures [21].

In light of these developments, the present study sets out to design and evaluate a holistic RAG-based system tailored to the demands of infectious disease management, with a particular emphasis on accommodating real-time data flows, domain-specific language embeddings, and robust back-end infrastructures for indexing and retrieval. By bridging the gap between existing predictive tools and purely generative models, this system aims to provide a practical solution for researchers, clinicians, and policy makers who require up-to-date, reliable, and easily interpretable information. It is anticipated that its success will rest on the synergy between specialized retrieval modules, carefully curated data pipelines, and a generative layer that integrates retrieved evidence in a transparent and actionable manner. This combined effort, drawing from the latest advancements in vector-based document indexing and transformer-based language models, has the potential to catalyze a new era of data-driven readiness in the face of rapidly evolving infectious diseases [22].

While our system enhances information retrieval and synthesis, it also raises important ethical considerations. Using diverse data sources, including social media and pre-print servers, necessitates rigorous filtering to mitigate misinformation and ensure data reliability. Furthermore, the potential for algorithmic bias in embedding models underscores the need for ongoing audits and transparency in model development. Ensuring compliance with data privacy regulations (e.g. HIPAA, GDPR) is essential to maintaining public trust and upholding ethical standards in public health research. By situating our work within the broader NLP and AI landscape, acknowledging the limitations of existing approaches, and addressing ethical concerns, we aim to contribute a robust, scalable solution for infectious disease research that balances technological innovation with responsible data use.

In positioning Retrieval Augmented Generation as a unified solution to the burgeoning challenges of infectious disease literature discovery, this study underscores how swift, AI-facilitated knowledge access can revolutionize outbreak response. Mpox serves as a concrete proof-of-concept, illustrating the efficiency and practicality of an approach that seamlessly integrates preprint updates, official bulletins, and peer-reviewed content [15]. By tailoring generation models to domain-specific embeddings, researchers and clinicians gain timely, trustworthy insights that can inform patient care, treatment protocols, and strategic public health interventions. As the system evolves, the pursuit of scalability, interpretability, and ethical rigor will ensure that its outputs remain robust across emerging disease threats. Ultimately, by marrying advanced retrieval with responsible AI stewardship, the proposed framework delineates a pathway toward more agile and informed responses to the ever-evolving landscape of global health emergencies.

Deploying AI-driven systems for infectious disease management involves addressing several interrelated challenges. Ensuring high data quality is paramount, as inconsistencies in data structure, outdated information, and missing context can lead to inaccurate responses and even propagation of misinformation. To combat this, rigorous data curation practices, including regular audits, cross-referencing with trusted sources, and robust validation pipelines are essential for maintaining the integrity of the knowledge base.

Equally critical is the performance of the information retrieval processes. As datasets expand, distinguishing between closely related topics becomes increasingly difficult, and inefficient algorithms may fail to retrieve the most relevant data. Additionally, determining the optimal chunking strategy for data segmentation is a delicate balance; overly large chunks risk incorporating irrelevant details, while chunks that are too small might omit necessary context, thereby undermining the accuracy of results. These issues are further compounded by scalability challenges. During periods of rapid data influx, such as during outbreaks, systems must manage large-scale data ingestion and maintain low latency under high query volumes, necessitating architectures optimized for horizontal scaling, parallel processing, and efficient memory management. Another significant challenge lies in the ability of large language models (LLMs) to accurately interpret healthcare terminologies and generate content tailored to the health domain. These systems must not only grasp specialized language and jargon but also format and contextualize their outputs to meet the rigorous

standards expected in clinical and public health communications.

Moreover, there is an ongoing need to address algorithmic bias and ensure fairness in outputs. Biases may be inadvertently introduced through pre-trained embeddings or skewed training data, potentially privileging certain regions, institutions, or demographics. Continuous evaluation using diverse benchmarks, alongside the deployment of interpretability tools and representative training corpora, is crucial to mitigate these issues and maintain balanced outcomes.

Finally, protecting sensitive health information is a critical component of system design. Adhering to data privacy regulations such as HIPAA and GDPR requires secure data ingestion workflows, rigorous anonymization techniques, and robust access controls. These measures not only ensure legal compliance but also foster the trust of users, health organizations, and government agencies, ultimately supporting the ethical and effective deployment of AI in public health contexts.

Future research should seek to broaden the scope of the current framework beyond its initial application to MPOX, extending its utility to other emerging pathogens such as avian influenza and novel coronaviruses. This extension is anticipated to facilitate the generalization of RAG-based retrieval systems across a wider spectrum of infectious diseases, thereby enhancing their applicability and impact in public health surveillance. Moreover, the integration of literature-derived insights with real-time data streams from hospitals and laboratories presents a promising avenue for research. Such a synthesis could lead to developing predictive analytics and early warning systems, which are critical for timely public health interventions.

Additionally, there is a pressing need to incorporate multilingual support within the framework. Enhancing cross-lingual retrieval capabilities will not only bridge language barriers but also consolidate findings from diverse global research communities, thereby enriching the collective understanding of disease dynamics.

Finally, future investigations should focus on integrating advanced analytical methodologies, including knowledge graphs, network analyses, and sophisticated statistical models. These tools are expected to bolster hypothesis generation and provide deeper insights into the complex interactions underpinning disease spread, ultimately contributing to more robust and proactive public health strategies.

References

1. Mullankandy S, Kazmi I, Islam T, Phia WJ. Emerging Trends in AI-Driven Health Tech: A Comprehensive Review and Future Prospects. *Eur J Technol.* 2024;8(2):25–40. <https://doi.org/10.47672/ejt.1888>
2. Yousif MM, Mohammed RO, Habibullah IJ, Rahim HM, Salih AM. Comparing ChatGPT and Google Bard: Assessing AI-Powered Information Retrieval in Nursing. *Barw Med J.* 2024. <https://doi.org/10.58742/hsn32c73>
3. Bullock J, Luccioni A, Hoffman Pham K, Sin Nga Lam C, Luengo-Oroz M. Mapping the landscape of Artificial Intelligence applications against COVID-19. *J Artif Intell Res.* 2020;69:807–45. <https://doi.org/10.1613/jair.1.12162>
4. Tarango J, González-Quñones F, Machin-Mastromatteo JD. Generación y uso inadecuado de información en ámbitos académicos y su influencia en entornos sociales: Desinformación, sobreinformación y misinformación. Universidad Autónoma de Chihuahua. 2021. Available at: https://ru.iibi.unam.mx/jspui/bitstream/IIBI_UNAM/138/1/13_informacion_desinformacion_javier_tarango_fidel_gonz%C3%A1lez_qui%C3%B1ones_juan_machin_mastromatteo.pdf
5. Pan American Health Organization (PAHO). Making Public Health Decisions. PAHO. 2020. Available at: https://iris.paho.org/bitstream/handle/10665.2/52952/Decisionespublicas_spa.pdf?sequence=1
6. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih WT, Rocktäschel T, Riedel S, Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint. Available from: <https://github.com/huggingface/transformers/blob/master/>
7. World Health Organization (WHO). Mpox [Internet]. Geneva: WHO; 2024 [cited 2024 Feb 19].

Available from: <https://www.who.int/es/news-room/fact-sheets/detail/mpox>

8. World Health Organization (WHO). WHO Director-General declares mpox outbreak a public health emergency of international concern [Internet]. Geneva: WHO; 2024 Aug 14 [cited 2024 Feb 19]. Available from: <https://www.who.int/es/news/item/14-08-2024-who-director-general-declares-mpox-outbreak-a-public-health-emergency-of-international-concern>
9. Hussain Khan S, Iqbal R, Naz S. A Recent Survey of the Advancements in Deep Learning Techniques for Monkeypox Disease Detection. Unpublished manuscript.
10. Organización Panamericana de la Salud (OPS). Actualización Epidemiológica Mpox en la Región de las Américas - 29 de octubre del 2024. OPS; 2024. Disponible en: <https://www.paho.org/es/documentos/actualizacion-epidemiologica-mpox-region-americas-29-octubre-2024>
11. Ministerio de Salud de Colombia. Panorama de la viruela símica en Colombia y el mundo. Ministerio de Salud y Protección Social; 2022. Disponible en: <https://www.minsalud.gov.co/Paginas/Panorama-de-la-viruela-simica-en-Colombia-y-el-mundo.aspx>
12. Ministerio de Salud de Argentina. Boletín Epidemiológico Nacional N°720, SE 35, Año 2024. Ministerio de Salud; 2024. Disponible en: https://www.argentina.gob.ar/sites/default/files/2024/04/ben_720_se_35.pdf
13. Prager EM, Chambers KE, Plotkin JL, McArthur DL, Bandrowski AE, Bansal N, Martone ME, Bergstrom HC, Bespalov A, Graf C. Improving transparency and scientific rigor in academic publishing. *J Neurosci Res*. 2019 Apr;97(4):377-390. doi: 10.1002/jnr.24340. Epub 2018 Dec 2. PMID: 30506706.
14. The Global Health Network. Advancing equity in health research [Internet]. Oxford: The Global Health Network; 2024 [cited 2024 Feb 19]. Available from: <https://tghn.org>
15. The Global Health Network. MPOX AI Resources [Internet]. Oxford: The Global Health Network; 2024 [cited 2024 Feb 19]. Available from: <https://outbreaks-aisearch.tghn.org/auth/login>
16. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Adv Neural Inf Process Syst*. 2020;33:9459-71.
17. Outbreaks AI Search [Internet]. The Global Health Network. [cited 2025 Apr 3]. Available from: <https://outbreaks-aisearch.tghn.org/auth/login>
18. Mukherjee S, Kulshreshtha M, Pancholi B, Garabadu D. Impact assessment of virus globally: special emphasis on COVID-19, Zika, and Ebola virus. *Nanostructured Drug Delivery Systems in Infectious Disease Treatment*. Elsevier; 2024. p. 25–63. <https://doi.org/10.1016/B978-0-443-13337-4.00004-5>
19. Radaelli D, Di Maria S, Jakovski Z, Alempijevic D, Al-Habash I, Concato M, Bolcato M, D'Errico S. Advancing Patient Safety: The Future of Artificial Intelligence in Mitigating Healthcare-Associated Infections: A Systematic Review. *Healthcare (Basel)*. 2024;12(19):1996. <https://doi.org/10.3390/healthcare12191996>
20. Gulia A, Arora RS, Panda PK, Raja A, Tiwari A, Bakhshi S, Salins N, Goel V, Janu A. Adapting Management of Sarcomas in COVID-19: An Evidence-Based Review. *Indian J Orthop*. 2021;55. <https://doi.org/10.1007/s43465-020-00143-1>
21. Abràmoff MD, Tarver ME, Loyo-Berrios N, Trujillo S, Char D, Obermeyer Z, Eydelman MB, Maisel WH. Considerations for addressing bias in artificial intelligence for health equity. *npj Digit Med*. 2023;6(1). <https://doi.org/10.1038/s41746-023-00913-9>
22. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, Dai Y, Sun J, Wang M, Wang H. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint*. 2023. Available from: <http://arxiv.org/abs/2312.10997>

Acknowledgements

This case study was supported by Wellcome under grant 226688/Z/22/Z. Additional support was provided by The Global Health Network, which facilitated access to resources and expertise.

Conflicts of Interest: The authors declare no conflicts of interest relevant to this study. None of the authors serve or have served on the editorial board of the journal to which this manuscript is submitted. The authors do not hold any positions in organizations that may benefit from the publication of this article, or have any financial interests in companies related to the technologies discussed. The authors confirm that they were not involved in the development of the AI-driven framework evaluated in this case study and have no direct financial benefits from its publication.

Abbreviations

AI – Artificial Intelligence
API – Application Programming Interface
BM-25 – Best Matching 25 (a ranking function for document retrieval)
CDC – Centers for Disease Control and Prevention
GCP – Google Cloud Platform
GDPR – General Data Protection Regulation
HAIs – Healthcare-Associated Infections
HIPAA – Health Insurance Portability and Accountability Act
JWT – JSON Web Token
LLM – Large Language Model
MPOX – Monkeypox
MRR – Mean Reciprocal Rank
NLP – Natural Language Processing
NGINX – Engine X (a web server and reverse proxy)
NER – Named Entity Recognition
PAHO – Pan American Health Organization
PLN – Procesamiento del Lenguaje Natural (Spanish equivalent of NLP)
RAG – Retrieval-Augmented Generation
RBAC – Role-Based Access Control
SSR – Server-Side Rendering
SSL/TLS – Secure Sockets Layer / Transport Layer Security
TGHN – The Global Health Network
UI – User Interface
WHO – World Health Organization