Retrieval-Augmented Generation Meets Local Languages for Improved Drug Information Access and Comprehension.

Ahmad Ibrahim Ismail, Bashirudeen Ibrahim Opeyemi, Olubayo Adekanmbi, Ife Adebara
Data Science Nigeria (DSN), Lagos, Nigeria
{ahmad, olubayo, ife}@datasciencenigeria.ai,
bashirudeenopeyemi772@gmail.com

Abstract

Medication errors are among the leading causes of avoidable harm in healthcare systems across the world. A large portion of these errors stem from inefficient information retrieval processes and lack of comprehension of drug information. In low-resource settings, these issues are exacerbated by limited access to updated and reliable sources, technological constraints, and linguistic barriers. Innovations to improve the retrieval and comprehension of drug-related information are therefore poised to reduce medication errors and improve patient outcomes. This research employed open-source Retrieval-Augmented Generation (RAG) integrated with multilingual translation and Text-to-Speech (TTS) systems. Using open-source tools, a corpus was created from prominent sources of medical information in Nigeria and stored as high-level text embeddings in a Chroma database. Upon user query, relevant drug information is retrieved and synthesized using a large language model. This can be translated into Yoruba, Igbo, and Hausa languages, and converted into speech through the TTS system, addressing the linguistic accessibility gap. Evaluation of the system by domain experts indicated impressive overall performance in translation, achieving an average accuracy of 73%, and the best performance observed in Hausa and Yoruba. TTS results were moderately effective (mean = 57%), with Igbo scoring highest in speech clarity (68%). However, tonal complexity, especially in Yoruba, posed challenges for accurate pronunciation, highlighting the need for language-specific model fine-tuning. Addressing these linguistic nuances is essential to optimize comprehension and practical utility in diverse healthcare settings .The results demonstrates system's the potential to improve access to drug information, enhance comprehension, and reduce linguistic barriers. These technologies could substantially mitigate medication errors and improve patient safety. This study offers valuable insights and practical guidelines for future implementations aimed at strengthening global medication safety practices.

1 Introduction

The traditional medication information retrieval and communication has always been plagued with a myriad of issues broadly and rightly categorised as "medication errors". Medication errors are the leading cause of avoidable harm in healthcare systems around the world, together with unsafe medication practices. A medication error is defined by the United States National Coordinating Council for Medication Error Reporting and Prevention as any avoidable incident that could result in the improper use of medication or harm to a patient while the medication is in the hands of a healthcare professional, patient, or consumer (National Coordinating Council for Medication Error Reporting and Prevention).

The World Health Organization recognises that medication error occurs in prescribing, transcribing, dispensing, and administering (World Health Organization, 2023). Therefore, medical information retrieval and communication must be optimised for efficiency, effectiveness and precision.

Drug information is usually retrieved manually from multiple sources, especially in low-resource settings. A study by (Ogbonna and Okoye, 2021) found that Nigeria's most common sources of drug information are the Nigerian Essential Medicines Index (EMDEX), the British National Formulary, Pharmacopoeias and product information leaflets (PILs) included in drug packages by manufacturers.

While all this information is usually readily available for healthcare professionals, studies suggest professionals in developed climes more frequently access the best-quality information, with the only barrier reported being time (Seidel et al., 2023). However, available drug information sources in low-resource settings are usually of lower quality

Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), pages 108–114
July 31, 2025 ©2025 Association for Computational Linguistics

and mostly outdated. The retrieval process is also slow and ineffective due to factors like drug availability, slow internet access, and a large population of patients (Abdel-Latif et al., 2022).

Patients, caregivers and consumers of medications are also involved in creating medication errors. Many information sources are accessible to consumers, caregivers and patients, including physicians and pharmacists, digital platforms and resources, printed materials like PILs and Drug information centers (DICs). While people worldwide have access to these various drug information sources, the quality, availability, and effectiveness of these sources differ widely, similar to the aforementioned trend among healthcare providers. Studies show that people from around the world read PILs, less than half understand the intended information while some even reported anxiety and confusion after consulting drug information sources (Rašković et al., 2024; Al Jeraisy et al., 2023; Owusu et al., 2020).

The most prominent recommendation from all the studies cited above is to explore ways to improve communication in the different drug information sources and optimise the information retrieval process. This aligns with (Okoye and Ogbonna, 2022)'s stress of the need to include and prioritise local languages in health service delivery.

Retrieval-Augmented Generation (RAG) is an advanced technique in natural language processing that enhances the capabilities of large language models (LLMs) by integrating them with external information retrieval systems, allowing them to access and incorporate up-to-date, domain-specific knowledge during the generation process, thereby improving the accuracy and reliability of their responses. This technique has attained wide adoption as it solves the most pervasive problem of LLM hallucination. (Lewis et al.).

This research leverages RAG to allow LLMs and translation models access and work with correct and up-to-date medical information from structured and unstructured formats to produce an application that aims to improve the retrieval and enhance the comprehension of drug information by healthcare providers, patients and caregivers to reduce medication errors. This research leverages open-source models and libraries to allow interested parties to freely tweak, modify and use the source code available on the GitHub repository to progress the aim of this research.

2 Literature Review

2.1 Retrieval-Augmented Generation in Healthcare

Large Language Models (LLMs) have demonstrated remarkable capabilities in medical language tasks, even answering medical exam questions with high accuracy in the United States (Sohn et al., 2024). However, their adoption in healthcare has been limited due to their potentials to hallucinate (generate confident but incorrect outputs) and limited access to up-to-date knowledge . In the medical field, precision is essential; an incorrect fact or dosage can carry serious risks. To mitigate these issues, RAG has emerged as a key strategy. RAG systems integrate external knowledge retrieval into the LLM's generation process, providing relevant context from trusted data sources to enhance accuracy (Miao et al., 2024).

The integration of RAG with medical NLP has shown clear benefits in improving the factual accuracy of AI outputs. A 2025 systematic review and meta-analysis of RAG in biomedicine found that augmenting LLMs with retrieval significantly improved performance, yielding a pooled 1.35× increase in accuracy over base LLMs (95% CI 1.19–1.53, p = 0.001). The review which analyzed 20 studies from 2023–2024 and identified common trends in how RAG is implemented (e.g. types of knowledge sources and evaluation methods), found that many of these studies demonstrate that RAG can markedly reduce LLM hallucinations and bias, making the outputs more trustworthy for medical use. The study ultimately proposing guidelines for safe clinical deployment of RAG-powered applications (Liu et al., 2025).

A team of researchers developed Almanac, a retrieval-augmented LLM for clinical decision support, which was evaluated on 130 realistic clinical scenarios. The RAG-augmented model showed an 18% improvement in factual accuracy of its recommendations (evaluated by physicians) compared to the base model, along with gains in completeness and safety (Zakka et al., 2023). Similarly, another team introduced the Rationale-Guided RAG (RAG2), achieving up to 6.1% higher accuracy by refining retrieval queries through model-generated rationales, further reducing the risk of misinformation in medical contexts (Sohn et al., 2024). MEDIC, an LLM-driven system augmented with domain-specific guidelines was developed with the aim of substantially reducing medication errors in online pharmacies. It was able to effectively standardize prescription directions and translate complex medical jargon into clear, patient-understandable instructions. This system achieved 33% reduction in medication errors as a direct result of improved patient comprehension of drug-related communications, and medication adherence (Pais et al., 2024).

In a low-resource context, a study showcased the potential of RAG to improve drug insight generation from local medical databases by creating a chatbot called "Drug Insights". This chatbot, tailored to the needs of frontline healthcare workers in Nigeria was able to effectively bridge the gap in drug information access (Owoyemi et al., 2025). RAG serves as an “open-book” exam mode for LLMs, ensuring their answers are supported by real sources rather than just the model’s internal training data.

2.2 Cross-Lingual Applications in Healthcare

Most advanced medical NLP solutions, including RAG-augmented systems, have been developed in a handful of high-resource languages, primarily English. This poses a barrier in multilingual societies and low-resource settings, where patients and health workers often speak and read in local languages. Bridging the language gap is crucial for equitable healthcare information access.

Recent efforts in multilingual and cross-lingual NLP aim to enable medical AI systems that can operate across diverse languages, either by building multilingual models or by coupling translation modules with information retrieval. One early demonstration of such an approach developed a multilingual question-answering system for rural healthcare information access. Their prototype was a full NLP pipeline that incorporated named entity recognition (NER) on user queries, translated the queries into English (the language of the medical knowledge base), retrieved relevant information, and then generated answers which could be translated back into the user’s local language (Vinod et al., 2021). Their model was designed to be low-resource and language-agnostic, targeting “indigenous languages” spoken in rural areas of developing countries. It enabled users to ask health questions in their native language and receive answers based on global medical knowledge. Their results demonstrated that such systems could be employed in healthcare systems to provide advice on common health issues and even produce preliminary summaries of patient health records for clinician review.

Subsequent projects have continued this line of work. The AwezaMed initiative in South Africa created a speech-to-speech translator for maternal healthcare during COVID-19 pandemic, enabling communication between English-speaking doctors and patients speaking indigenous languages to disseminate timely knowledge about prevention and treatment (Hu et al., 2025).

2.3 Applications in Low- and Middle-Income Countries

The confluence of RAG and multilingual NLP opens up especially exciting opportunities for low- and middle-income countries. Many LMICs face severe shortages of healthcare professionals, and those in practice often serve multilingual populations with limited resources. AI systems that provide decision support and information in local languages could help bridge gaps in healthcare delivery (Okoye and Ogbonna, 2022).

Open access RAG tools are particularly valuable in LMICs, where cost and proprietary systems are barriers, an open framework allows local innovation and continuous improvement by the community. The inclusion of regional experts in building these tools ensures that the solutions are culturally and linguistically appropriate.

3 Methodology

3.1 Data Collection and Extraction

This research utilises data from Nigeria’s most prominent medical information sources. This information was extracted into text format using Python libraries[1] like PyMuPDF for PDF files, requests and beautifulsoup for data sourced from the web. These libraries are standard in NLP and offer the best to extract data while maintaining inherent semantic relationships. They also allowed us to store and access the metadata of the source documents to create a large corpus, an important process for our pipeline during the synthesis stage, as it allows the LLM in our work to make the best decisions.

[1]PyMuPDF is a lightweight PDF and XPS parsing library https://pymupdf.readthedocs.io/, Requests is a simple and elegant HTTP libraryhttps://docs.python-requests.org/, and BeautifulSoup is a Python library for parsing and scraping HTML and XML https://beautiful-soup-4.readthedocs.io/.

3.2 Data Preprocessing

The handy regex library was used in this research to clean data. This process involved removing excessive blank lines, fixing hyphenated words at line breaks (e.g., "exam- nple" to "example"), normalising spacing around punctuations, removing extra spaces before new lines, matching and removing unnecessary patterns like those seen in indexes and appendixes, and reconstructing broken paragraphs. The RecursiveCharacterTextSplitter class from the LangChain framework[2] was used to divide the clean text into manageable chunks. It works by splitting texts using a predefined sequence of characters, proceeding recursively through the list until the resulting segment meets the desired length criteria (chunk_size=1024). Setting the parameter 'chunk_overlap=100' allows characters to overlap between consecutive chunks, ensuring context continuity across segments. This also enables the definition and access of a structured data schema, where essential drug information, e.g., name, class, indication, interactions, contraindication, etc., defined in the source materials are maintained.

3.3 Vectorization and Vector Storage

multilingual-e5-large[3], an open-source, state-of-the-art, high-performance, multilingual text embedding model developed by Microsoft and available on Huggingface[4] was used to convert the text chunks into vector embeddings. These embeddings were then stored in Chroma[5], an open-source vector database designed to store and retrieve vector embeddings efficiently. Integrating the Chroma vector store with metadata support enables efficient management and retrieval of embeddings, facilitating accurate nearest-neighbor searches based on the default Euclidean distance metric.

3.4 Drug Information Generation

With the knowledge base in the Chroma database, the system employs a retrieval-augmented generation (RAG) approach to synthesising and generating drug-related information. When a query is submitted, a similarity search is performed within the database to retrieve the most relevant drug-related information from the indexed knowledge sources. The retrieved texts containing structured pharmaceutical details are then passed as contextual input to a large language model (OpenAI's GPT 3.5 turbo[6]), which analyses the retrieved context, extracts essential details, and synthesises a concise, coherent, and medically relevant response.

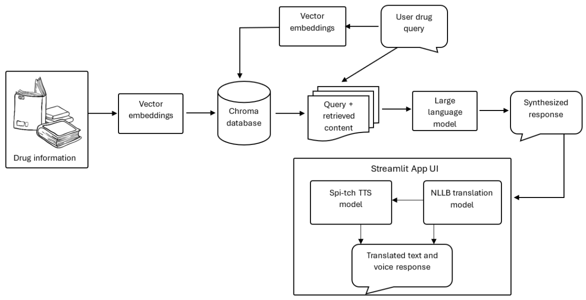

Figure 1: System flowchart

3.5 Translation

To achieve the aim of this research, the output synthesised by the LLM must be available to the end user in local languages. To facilitate this, the research employed Meta's NLLB[7], a multilingual machine translation model capable of translating between 200 languages, including many low-resource languages.

3.6 Text to Speech (TTS)

Spi-tch[8] Text-to-Speech (TTS) system was utilized to convert textual drug information into spoken output across the Yoruba, Igbo, and Hausa languages. Along with support for these languages, Spi-tch offers a selection of eight unique voices, each with distinct attributes, to enhance specific features of the synthesised speech. To ensure accurate pronunciation, especially in tonal languages like Yoruba, we applied Spitch's tone-marking feature before speech generation, allowing the model to pronounce words properly during synthesis. The audio outputs generated were in 'wav' format, facilitating seamless integration into our application.

4 Result

The output of this research is a RAG-powered chat application (built on streamlit) that leverages a cor-

[2]https://www.langchain.com/

[3]https://huggingface.co/intfloat/multilingual-e5-large

[4]Hugging Face is an AI platform that hosts open-source platform that machine learning models, datasets, and tools. http://https://huggingface.co/

[5]https://docs.trychroma.com/docs/overview/introduction

[6]https://openai.com

[7]https://huggingface.co/docs/transformers/model_doc/nllb

[8]https://docs.spi-tch.com/getting-started/welcome

pus of leading drug information data in the region. The system is built on various open-source tools to facilitate reproducibility and wrapped around a user friendly UI (see Appendix 6) using Streamlit.

The notebooks and codes used are available on GitHub

4.1 Evaluation Strategy

The system was evaluated by three independent groups of licensed pharmacists, who assessed the text and voice translation components based on a structured evaluation framework. Each group was given 20 prompts, with Group 2 generating additional domain-specific questions based on their clinical expertise. The evaluation criteria focused on three key aspects:
1. Drug Information Accuracy – The accuracy and completeness of drug-related information retained in the translated output.
2. Language Output Accuracy – The correctness of translations in the target languages (Yoruba, Igbo, and Hausa).
3. Structure of Output/Completion – The final output's coherence, grammatical structure, and completeness.

4.2 Drug Information in Local Languages

For text translation, results were recorded separately for each of the three languages (Yoruba, Igbo, and Hausa) and aggregated across the three evaluators. The overall average score for text translation was (73%), indicating high accuracy and completeness in the system's ability to translate drug-related information.
Yoruba Language had the highest scores (80%), showing strong accuracy in language fluency and drug information retention. Hausa followed closely, while Igbo had slightly lower performance, particularly in Drug Information Accuracy, where it recorded the lowest score of 60% in one evaluation.

Criteria	Yoruba	Igbo	Hausa
Language Output	8	7	7
Drug Information	7	6	9
Output Structure	7	6	9
Total Score (%)	**23 (77)**	**20 (67)**	**24 (80)**
Average Score (%)	**22 (73)**	**20.3 (68)**	**22.3 (74)**

Table 1: Evaluation Summary for language translation output.

4.3 Drug Information Voice Output (TTS)

For speech generation, the average score across all languages was 57%, indicating a moderate level of accuracy and output structure compared to text translation. Igbo (70%) was the highest-performing language, showing strong audio output accuracy. Yoruba performed the lowest, with one evaluator scoring it 40% due to pronunciation clarity and structure issues.

Evaluation Criteria	Yoruba	Igbo	Hausa
Language Output	4	7	7
Drug Information	5	6	5
Output Structure	5	6	5
Total Score (%)	**11 (36)**	**21 (70)**	**19 ()**
Average Score (%)	**14.3 (48)**	**20.3 (68)**	**18.0 (60)**

Table 2: Evaluation Summary for TTS output.

5 Discussions

This study aimed to address the critical issue of medication error from different barriers from the healthcare practitioners, patients, and caregivers by emphasising the importance of accurate, accessible and comprehensible drug information, particularly in low resource settings. It leveraged a Retrieval-Augmented Generation (RAG) system integrated with open-source language, translation and voice models. It demonstrated significant potential to enhance drug information retrieval and comprehension, thus contributing to reduced medication errors. The system can serve as an intelligent assistant for healthcare professionals, enabling them to counsel patients who speak only their local language and thereby strengthen understanding, compliance, and adherence to prescribed medications. The findings showed the robust performance of the retrieval process and the text translation component. The highest accuracy was achieved for the Yoruba language, with a better average recorded for the Hausa language, indicating effective linguistic adaptability of the multilingual model. The moderate performance of Igbo text translations, particularly in Drug Information Accuracy, underscores the need for further training or fine-tuning of the language model on domain-specific data. Conversely, the Text-to-Speech (TTS) component exhibited more varied performance. The Igbo language audio translations showed the highest accuracy, indicating effective phonetic adaptation and clarity. In contrast, Yoruba audio outputs exhibited lower performance, primarily due to pronunciation issues inherent in the tonal complexities of the language. This highlights the critical need for improving TTS models, especially for tonal languages, to enhance user comprehension and ensure accurate drug information delivery.

6 Conclusion

This research shows that implementing RAG into multilingual translation and TTS systems could enhance drug information knowledge availability, accessibility and comprehension, especially in low resource settings. This encouraging result in the accuracy of the text translation, findings from domain experts still show the need for improvement in the TTS system. However, this research has proven that an RAG-powered system is a viable tool for future efforts to improve medication information comprehension and reduce medication error.

References

Mohamed Mahmoud Abdel-Latif, Hosny Ahmed Elewa, and Amal Abd El.Moneim Soliman El-Kholy. 2022. Perceptions and attitudes of hospital' prescribers towards drug information sources and prescribing practices. *Brazilian Journal of Pharmaceutical Sciences*, 58:e20498. Accessed: 2025-03-09.

Majed Al Jeraisy, Heba Alshammari, Mashael Albassam, Kholoud Al Aamer, and Mostafa A. Abolfotouh. 2023. Utility of patient information leaflet and perceived impact of its use on medication adherence. *BMC Public Health*, 23(488). Accessed: 2025-03-09.

S. Hu and 1 others. 2025. Natural language processing technologies for public health in africa: Scoping review. *Journal of Medical Internet Research*, 27:e68720.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:200 ::contentReference[oaicite:0]index=0*.

S. Liu and 1 others. 2025. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *Journal of the American Medical Informatics Association (JAMIA)*. Ahead of print.

C. Miao and 1 others. 2024. Integration of retrieval-augmented generation (rag) in clinical practice: Enhancing clinical decision support and information retrieval in nephrology. *Medicina*, 60(3):445.

National Coordinating Council for Medication Error Reporting and Prevention. About medication errors. Accessed: 2025-03-09.

Brian O. Ogbonna and Ijeoma Mary-Ann Okoye. 2021. Utilization of drug information services in selected tertiary hospitals in enugu state, nigeria. *Journal of Current Biomedical Research*, 1(1):15–22. Accessed: 2025-03-09.

Ijeoma Okoye and Brian O. Ogbonna. 2022. Drug information services utilization in nigeria from 1980-2020: A narrative review of related studies. *Journal of Current Biomedical Research*, 2(4):380–390. Accessed: 2025-03-09.

J. Owoyemi and 1 others. 2025. "drug insights" – an open-source rag chatbot for medication lookup in africa. *medRxiv preprint*.

Frederick W. A. Owusu, Genevieve Naana Yeboah, Rahel Adutwiwah Aboagye, Cedric Dzidzor K. Amengor, and Philomena Entsie. 2020. The role of the patient information leaflet in patients' medication therapy: A case study within the kumasi metropolis of ghana. *The Scientific World Journal*, 2020:2489137. Accessed: 2025-03-09.

C. Pais and 1 others. 2024. Large language models for preventing medication direction errors in online pharmacies. *Nature Medicine*, 30(6):1574–1582.

Aleksandar Rašković, Maja Steinbach, Snežana Mugoša, Jovan Baljak, Dušan Prodanović, Stanislav Sabo, Dragica Bukumirić, Boris Milijašević, and Milica Paut Kusturica. 2024. Patient information leaflets: How do patients comprehend and understand drug information? *Archives of Pharmacy Practice*, 15(1):7–11. Accessed: 2025-03-09.

Pia Seidel, Bo Rolander, Anna L. Eriksson, Ulf Lindahl, Susanna M. Wallerstedt, Staffan Hägg, and Anders Kling. 2023. Drug information sources in professional work—a questionnaire study on physicians' usage and preferences (the drug information study). *European Journal of Clinical Pharmacology*, 79:767–774.

J. Sohn and 1 others. 2024. Rationale-guided rag for medical question answering. *arXiv preprint arXiv:2411.00300*.

V. Vinod and 1 others. 2021. Multilingual medical qa and information retrieval for rural health. In *ICLR AI4PH Workshop*.

World Health Organization. 2023. Global burden of preventable medication-related harm in health care: a systematic review. Technical report. Accessed: 2025-03-09.

C. Zakka and 1 others. 2023. Almanac — retrieval-augmented language models for clinical medicine. *NEJM AI*, 2(3).

Appendix

Appendix A: Streamlit Screenshot

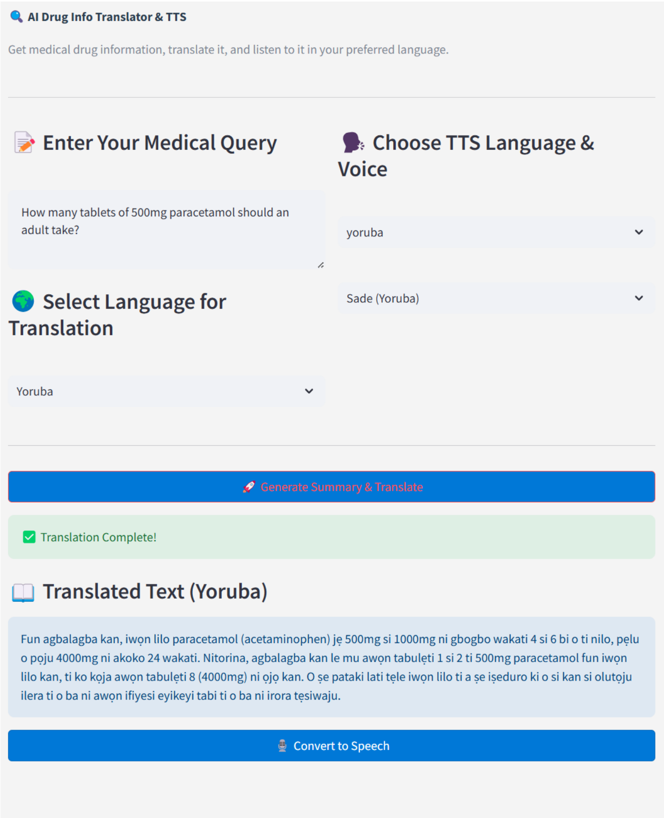

Figure 2: An image of the system running on Streamlit with sample query and output.

Appendix B: GitHub Repository

The full source code for this project is available at:
https://bit.ly/RAGDRUGINFOLANGUAGE