2nd Place Solution for Malawi Public Health Systems LLM Challenge

Daoud Saif - KHLIFI Mohamed
Ecole Polytechnique de Tunisie
Sfax, Tunisia
saif.sedaoud@gmail.com
mohamed.khelifi@ept.ucar.tn

ABSTRACT

Rapid evolution of the public health challenges and the growing complexity in the health information requires innovativeness in the solution of knowledge dissemination. This solution highlights the novel application of the 'Retrieval Augmented Generation (RAG)' model in the public health domain with the objective to enhance the capabilities of LLMs by retrieving real-time evidence from a well-curated dataset of booklets on public health. With its huge wealth of structured data contained in these authorial sources, the solution enhances, in a big way, the accuracy, relevance, and contextuality of health advice given over digital platforms.

1 OVERVIEW AND OBJECTIVES

The integration of Retrieval Augmented Generation (RAG) technology into the public health domain represents a transformative approach to enhancing the capabilities of language models. This solution aims to bridge the gap between the vast amounts of structured and unstructured public health data contained in various booklets and the need for personalized, context-rich public health advice dispensed through digital platforms.

The core purpose of this solution is to augment the response generation capabilities of language learning models (LLMs) by enabling them to retrieve and incorporate specific, relevant information from a curated dataset of public health booklets. This dataset comprises a comprehensive collection of guidelines, research findings, and health advisories published by authoritative bodies in the public health sector. By harnessing this wealth of information, the solution addresses several critical problems:

- **Limited Context Awareness:** Traditional LLMs often generate responses based solely on the input they receive, lacking the depth and context that specific queries may require, especially in nuanced fields like public health.
- **Static Knowledge Base:** The rapid evolution of public health knowledge and guidelines necessitates a dynamic approach to knowledge integration, beyond the fixed training datasets typically used to train LLMs.
- **Accessibility and Reliability:** Ensuring that users have access to reliable, evidence-based health information tailored to their inquiries, without the need for extensive searches through dense, technical documents.

The objectives of implementing this RAG solution in the public health domain are multifaceted, aiming to significantly enhance the accuracy, relevance, and reliability of the information provided to users. Key objectives include:

- **Enhancing Response Quality:** By retrieving and integrating context-specific information from health booklets, the solution aims to improve the quality and accuracy of responses provided to public health inquiries.
- **Dynamic Knowledge Utilization:** Implementing a system that can adapt to the inclusion of new information, ensuring that the advice dispensed reflects the latest public health research and guidelines.
- **Increasing User Trust and Engagement:** Providing reliable, evidence-based information tailored to user queries is expected to foster trust in digital public health platforms and encourage proactive engagement with health information.

The expected outcomes of this solution encompass a marked improvement in the capability of digital platforms to dispense personalized, accurate, and context-rich health advice. This includes:

- A measurable increase in the relevance and accuracy of responses to user inquiries, as evaluated by user feedback and engagement metrics.
- Enhanced accessibility to up-to-date public health information, directly impacting public health literacy and empowerment.
- Strengthened reliance on digital health advisories, potentially leading to more informed health decisions by the public.

By achieving these objectives, the solution aspires to set a new standard for the integration of advanced AI technologies like RAG in public health, paving the way for more informed, data-driven health communication and policy-making.

2 METHOD

In Figure 1, there is a visual representation of the pipeline used for this solution. It outlines the process from the initial user query through to the response generation by the Large Language Model (LLM), including the checks for abbreviations and the placement of information. The pipeline is designed to efficiently categorize user queries into three distinct types based on their intent and content, ensuring that each is processed in the most appropriate manner for accurate and relevant responses as shown below in Table 1.

3 ETL PROCESS

3.1 Extract

The training data provided for this competition is drawn from six booklets representing sections of the Technical Guidelines for Disease Surveillance and Response (TGs for IDSR) in Malawi. These

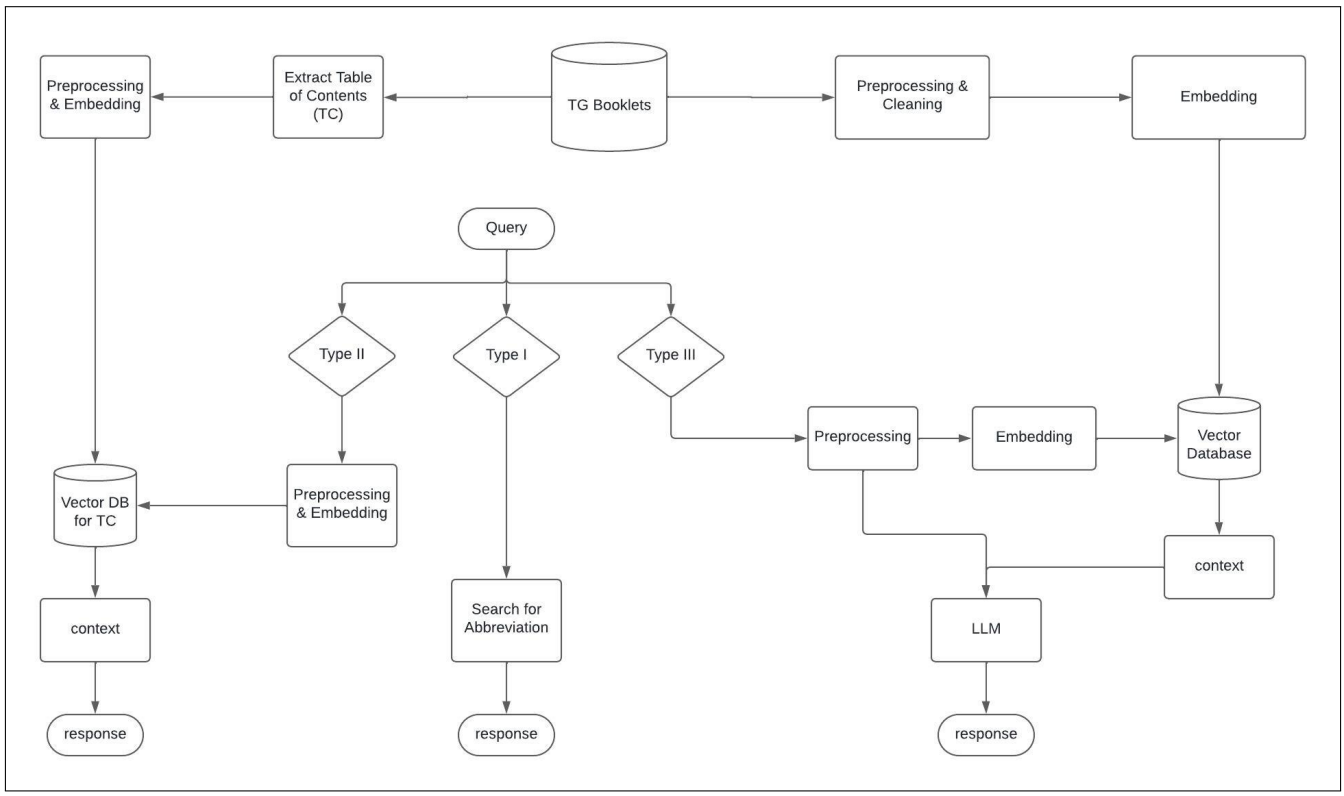

Figure 1: Architecture Diagram

booklets are available in .xlsx format, containing numbered paragraphs of the text. The distribution of the paragraphs is shown in Figure 2.

Given the uniform format of the dataset, the extraction process is streamlined to efficiently handle the Excel files. The process involves automated extraction Utilizing scripts that leverage Excel handling libraries to systematically parse and extract the numbered paragraphs. This approach ensures that the textual content is accurately retrieved, maintaining the integrity of the data.

3.2 Transform

In the transformation phase of our ETL process, we meticulously prepare the data extracted from the Technical Guidelines (TG) booklets, focusing on optimizing it for the distinct needs of our model, especially catering to the three types of user queries. This stage is critical for enhancing the model's ability to accurately interpret and respond to queries regarding abbreviations, specific information location, and general inquiries. The transformations are tailored to refine the data for its intended use:

- Standardizing the format of abbreviations, ensuring a uniform presentation of acronym-expansion pairs.
- Normalizing headings and subheadings from the table of contents to ensure consistency across different TG booklets, aiding in precise information location.

Table 1: Query types description

Query Type	Description
I	The user is looking for an abbreviation.
II	The user is seeking the location of specific information within the TG booklets.
III	The rest of the questions.

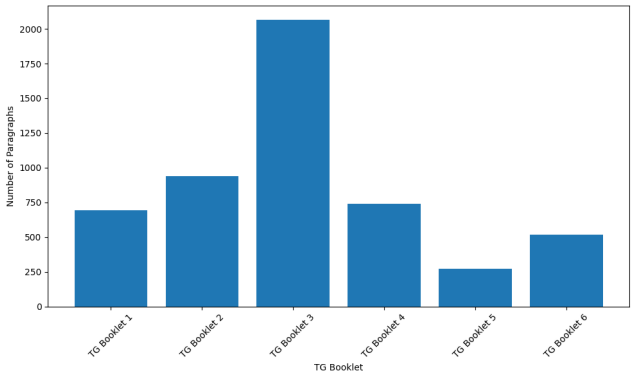

Figure 2: Number of Paragraphs in Each TG Booklet

- Unifying text format across documents to a common standard, such as lowercase transformation, to facilitate text processing and matching.

3.3 Load

Following the rigorous extraction and transformation phases, the load phase involves transferring the structured and optimized data into a storage system designed for efficient modeling and inference. This phase is crucial for ensuring the data is readily accessible and in a format conducive to rapid query processing and response generation. The loading strategy employs a combination of data storage mechanisms and advanced indexing and optimization techniques, tailored to the specific needs of our model. **Data Storage**

Mechanisms:

- **Abbreviations Dictionary:** The transformed abbreviations data is stored in a dictionary format, offering a straightforward and efficient means of accessing abbreviation - expansion pairs. This approach provides immediate lookup capabilities, essential for swiftly responding to Type I queries that seek the expansion of abbreviations.
- **Vector Databases for Table of Contents and Documents:** Two separate vector databases are utilized to store the table of contents and the entirety of the transformed documents. The distinction between these databases allows for tailored search strategies, optimizing the system's performance in locating specific information (Type II queries) and providing comprehensive responses to general queries (Type III).

Indexing and Optimization Strategies:

- **Chunking Methodology:** To effectively manage the extensive data from the documents, a robust chunking method is employed. Each document is divided into chunks of 500 characters, with a 300 character overlap between consecutive chunks. This strategy ensures that no critical information is lost or contextually isolated, enhancing the relevance and accuracy of the information retrieved in response to queries.
- **Embedding and Vectorization:** Post-chunking, the data undergoes an embedding process using the Jinaai base Sentence Transformer Model which is available on Hugging Face (https://huggingface.co/jinaai/jina-embeddings-v2-base-en), identified as the most performant encoder for our needs. This embedding transforms textual data into high-dimensional vectors that capture the semantic essence of the text. The vectorization of both the table of contents and the full document contents allows for the implementation of similarity search algorithms within the vector databases. These algorithms can rapidly identify the most relevant chunks of text in response to a query, significantly improving the system's efficiency and accuracy.
- **Optimization for Rapid Retrieval:** The vector databases are optimized for high-speed retrieval. Indexing strategies are applied to the embedded vectors, facilitating quick and efficient similarity searches. This optimization ensures that the retrieval component of our RAG system can access and return the most pertinent information with minimal latency, a critical factor in maintaining user engagement and trust.

The load phase is characterized by the strategic organization and optimization of data for use in the RAG system. By employing a dictionary for abbreviations and separate vector databases for the table of contents and document chunks, coupled with sophisticated chunking and embedding techniques, we ensure that our data is not only accurately represented but also primed for efficient retrieval and processing. This careful consideration of storage mechanisms and optimization strategies lays the groundwork for a responsive, accurate, and reliable system capable of meeting the diverse informational needs of its users.

4 DATA MODELING

Our solution employs a hybrid data model that integrates the principles of Retrieval Augmented Generation (RAG) with advanced natural language processing (NLP) techniques. This model is underpinned by two key theoretical foundations: transformer architecture for deep learning, which allows for efficient handling of sequential data, and vector space modeling for information retrieval, enabling the system to match query inputs with relevant data extracts effectively.

Assumptions:

- The textual data from Technical Guidelines (TG) booklets is sufficiently comprehensive to cover a wide array of public health queries.
- Embeddings generated from the preprocessed data capture semantic similarities in a manner that facilitates accurate retrieval and response generation.

Data preparation:

- All textual data undergoes normalization to a uniform case (lowercase) and format, ensuring consistency across the dataset.
- **Separate Vector Databases:** A distinct vector database is constructed for each TG booklet, allowing for the isolation and targeted analysis of their content.
- **Top Similarities Averaging:** For a given query, we compute the similarity scores between the query vector and the vectors representing the content of each TG booklet. The top 3 similarity scores within each booklet's vector database are averaged to determine the booklet's overall relevance to the query.
- **Maximum Similarity Retrieval:** The TG booklet with the highest average similarity score is identified as the most relevant source of information for the query. This booklet is then selected for retrieval, ensuring that the response generated for the user is grounded in the most semantically aligned content available.

This approach underpins our system's ability to accurately match user queries with the most relevant information from the extensive corpus of TG booklets. By leveraging advanced NLP techniques and a strategic similarity scoring mechanism, we ensure that our

model delivers responses that are not only informative and contextually appropriate but also deeply rooted in the specific knowledge domains represented by each TG booklet.

Pretraining and Quantization:

- **Pretrained Model:** We opted to utilize the Zephyr-7b [1] model in its pretrained state, acknowledging its extensive prior training on a broad and diverse dataset. This pretraining imbues the model with a rich understanding of language and context, crucial for the nuanced task of generating public health information.
- **Quantization:** To enhance model efficiency and reduce the computational resource requirements without significantly compromising performance, we applied a quantization process, reducing the model's precision to 4 bits. This optimization step is vital for deploying the model in resource-constrained environments, ensuring wider accessibility and faster response times.

Hyperparameters:

The model's performance and its ability to generate coherent, relevant responses are fine-tuned through the careful selection of hyperparameters:

- **MAX NEW TOKENS = 130:** Limits the maximum number of new tokens generated in each response, balancing informativeness with conciseness.
- **TEMPERATURE = 0.1:** Sets a low randomness in the prediction process, favoring more deterministic outputs for consistency in public health advice.
- **TOP K = 50:** Restricts the model to sampling from the top 50 most likely next words, focusing the generation process.
- **TOP P = 0.95:** Employs nucleus sampling to consider only the top 95% of the probability mass for the next word, enhancing the relevance of generated content.
- **DO SAMPLE = True:** Enables stochastic sampling, introducing variability in responses to cover a wider range of information and nuances in public health topics.

Training Processes:

Given the Zephyr-7b model's comprehensive pretraining, we elected not to pursue further fine-tuning. This decision is predicated on the model's already demonstrated proficiency in handling diverse and complex language tasks, aligning well with our solution's objectives to provide accurate, informative, and easily understandable public health information.

Evaluation Metrics:

ROUGE Score: To quantitatively measure the model's performance, particularly in its ability to generate text that aligns closely with human expectations and the factual content of the public health data, we employ the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scoring system. This metric assesses the overlap between the content generated by our model and reference texts, focusing on the recall and precision of 1-grams, which is crucial for evaluating the informativeness and relevance of the model's outputs.

5 INFERENCE

5.1 Model Deployment for Inference

Deployment Infrastructure:

- Our model can be deployed within a cloud-based architecture, leveraging containerization technology for scalability and reliability. Docker containers are used to encapsulate the model's environment, ensuring consistency across development, testing, and production environments. Kubernetes orchestrates these containers to manage deployment scales dynamically based on demand.
- We can also utilize cloud services that provide robust computing resources and storage capabilities, facilitating the seamless handling of high-volume queries and ensuring the system's responsiveness.

Services Used:

- API Gateway: An API gateway serves as the primary interface for receiving queries and delivering responses, ensuring secure and efficient communication between users and the model.
- Load Balancers: To distribute incoming queries evenly across multiple instances of the model, load balancers are employed, enhancing the system's ability to handle peak loads without degradation in performance.

5.2 Data Input and Output Interpretation

- **Data input:** users input their queries through a web interface or a dedicated API, designed for ease of use. The system preprocesses these queries, including tokenization and vectorization, to transform them into a format compatible with the model's input requirements.
- **Output Interpretation:** the model's output, typically in the form of vector embeddings, is mapped back to human-readable text using an inverse vectorization process. This text is then presented to the user as the response to their query. Furthermore, responses are accompanied by metadata when applicable, such as references to the TG booklet or section from which the information was derived, enhancing the transparency and usefulness of the information provided.

5.3 Model Updates

To ensure our system remains at the forefront of providing accurate and current public health information, we have implemented a dynamic model update, versioning, and retraining framework. This framework is particularly designed to automatically integrate new data, such as the addition of new TG Booklets, ensuring the model's advice remains relevant and comprehensive.

6 PERFORMANCE METRICS

To ensure the continuous improvement and operational excellence of our system, we employ a comprehensive suite of metrics and KPIs. These measures are designed to evaluate the system across multiple dimensions, including the efficiency of the ETL processes, the accuracy of the model, and the quality of inference outcomes.

Figure 3: Evaluation on booklets

Table 2: Evaluation on paragraphs based on query type

Query Type	accuracy	rouge
I	1.0	1.0
II	0.838	0.838
III	0.426	0.501
All	0.478	0.545

Table 3: Evaluation on paragraphs based on booklets

Metric	Rouge Score
TG Booklet 1	0.434
TG Booklet 2	0.443
TG Booklet 3	0.394
TG Booklet 4	0.306
TG Booklet 5	0.365
TG Booklet 6	0.719

Our evaluation has highlighted notable successes in our system's performance. Figure 3 showcases strong accuracy rates, particularly impressive in the context of TG Booklet 6. The system's adept handling of Type I queries is reflected in perfect accuracy and ROUGE scores, underscoring its reliability for straightforward information retrieval. Positive outcomes for Type II and III queries indicate a robust capability to tackle more complex inquiries. These insights emphasize the system's readiness and effectiveness as a tool for public health information.

7 ERROR HANDLING AND LOGGING

Our system implements robust error handling mechanisms at each stage of the ETL, modeling, and inference processes to ensure resilience and reliability. Errors are captured and categorized to facilitate rapid troubleshooting. Log entries are meticulously recorded, timestamped, and stored securely, allowing for comprehensive audit trails and proactive response strategies.

Table 4: Evaluation on keywords based on query type

Query Type	rouge
I	0.853
II	0.800
III	0.435
All	0.476

8 MAINTENANCE AND MONITORING

Maintenance protocols are in place to regularly assess and update the ETL workflows, models, and inference engines, ensuring they operate at peak efficiency. Performance monitoring is continuous, leveraging automated alerts to identify and address potential issues swiftly. Our scalability strategy employs cloud-native solutions, allowing the system to dynamically adjust to varying loads and ensuring consistent service quality throughout its lifecycle.

9 CONCLUSION

In conclusion, our system stands as a testament to the seamless integration of advanced technology and public health expertise. Through meticulous error handling and logging, we've constructed a robust framework that not only anticipates and mitigates issues but also learns from them. Our dedication to rigorous maintenance and vigilant monitoring ensures that the system evolves, adapts, and scales in line with the ever-changing landscape of public health needs. As we move forward, we remain committed to refining our processes, ensuring our system remains a reliable beacon in the quest for accessible and accurate public health information.

ACKNOWLEDGMENTS

We extend our heartfelt gratitude to the Zindi platform for their invaluable contribution to the success of this project. Zindi's commitment to fostering a collaborative environment has not only provided us with rich datasets but also the opportunity to engage with a vibrant community of data scientists and public health experts. Their support has been instrumental in enabling us to harness the power of AI for the betterment of public health information dissemination. We look forward to continued collaboration with Zindi as we further our endeavors in this field.

REFERENCES

[1] 2023. *ZEPHYR: DIRECT DISTILLATION OF LM ALIGNMENT.* https://arxiv.org/pdf/2310.16944.pdf.