# Vietnamsese RAG for Healthcare

**Pham Hong Tra, Nguyen Le Thanh Minh**
Faculty of Information Science and Engineering, University of Information Technology,
Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{22521495,22520875}@gm.uit.edu.vn

## Abstract

This study explores the application of Retrieval-Augmented Generation (RAG) in the healthcare domain to enhance the accuracy and mitigate hallucinations of large language models. We compare the performance of keyword-based retrieval using BM25 with sentence embedding methods in retrieving relevant context. Results indicate that BM25 performs better in the healthcare domain due to the importance of keywords. Based on this finding, we combine BM25 with a cross-reranker to optimize context selection. We evaluate the performance of large language models GPT and Gemini when provided with comprehensive context. The results demonstrate that with appropriate context, both models achieve high accuracy and significantly reduce hallucinations. This research contributes to the effective application of RAG in healthcare, particularly in providing accurate and reliable information.

**Terms**: RAG, LLM, Healthcare, Information Retrieval...

## 1 Introduction

This research paper aims to explore and develop a Vietnamese RAG (Retriever-augmented Generation) system for the healthcare domain, addressing the growing demand for accurate and reliable medical information among the public. Currently, the health of the Vietnamese population is facing numerous challenges, particularly the rise of non-communicable diseases such as cardiovascular diseases, diabetes, and cancer. According to a report from the World Health Organization (WHO), Vietnam is currently undergoing an aging population process, which creates a significant demand for healthcare information, disease prevention, and effective treatment.

However, finding accurate medical information through online channels such as websites and health Q&A forums remains a challenge, as the quality and reliability of these information sources have not been fully validated. Health Q&A websites, while having a large number of participants, are often unable to ensure the provision of relevant and accurate information for users, leading to the risk of patients receiving misleading or outdated information.

RAG (Retriever-augmented Generation) is an advanced natural language processing method that combines information retrieval from large document sources (retrieval) with answer generation based on the retrieved information (generation). This approach promises to significantly improve the ability to provide timely and accurate medical information to users, while also supporting doctors and healthcare professionals in offering appropriate and effective advice to patients.

This paper will introduce the methodology for building and applying the Vietnamese RAG system in the healthcare field, aiming to improve healthcare quality and meet the public's demand for fast and accurate medical information.

The data for this task has been collected from four reputable websites that provide health information to users across a variety of domains, including reproductive health, pediatrics, respiratory diseases, digestive health, etc.

## 2 Methodology

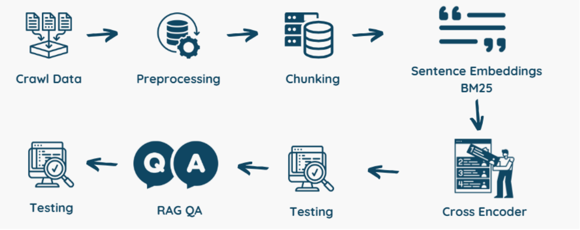


Figure 1: Overview of methodology

The implementation of the RAG system follows four main stages:

**Stage 1: Data Collection, Dataset Construction, and Preprocessing:** This stage involves gathering data from various sources, creating a suitable dataset, and preparing the data for the RAG pipeline. Preprocessing steps may include:

- Cleaning and normalizing text data (e.g., removing noise, handling special characters, lowercasing).
- Tokenization (splitting text into words or subwords).
- Indexing the dataset for efficient retrieval.

**Stage 2: RAG Module 1 - Retrieval:** This module focuses on retrieving relevant documents or passages from the dataset based on a user query. This typically involves:

- Encoding the query and documents into vector representations (embeddings).
- Calculating similarity scores between the query and documents (e.g., using cosine similarity).
- Ranking and selecting the top-k most relevant documents.

**Stage 3: RAG Module 2 - Generator:** This module uses a Large Language Model (LLM) to generate a response based on the retrieved information and the original query. The LLM aims to:

- Contextualize the retrieved information.
- Generate a coherent, informative, and relevant response.
- Minimize hallucinations by grounding the response in the retrieved context.

**Stage 4: Evaluation:** This final stage evaluates the performance of the RAG system. This may involve:

- Automatic evaluation metrics (e.g., BLEU, ROUGE, METEOR).
- Human evaluation to assess relevance, fluency, and factual accuracy.
- Comparing different RAG configurations or LLMs.

## 3 Data

To build a dataset for the RAG task in the healthcare domain, we collected data from top reputable health information-sharing websites in Vietnam, totaling 79,037 articles covering a wide range of specialized topics, such as reproductive health, pediatrics, cardiology, respiratory diseases, etc. The selection of these sources ensures the accuracy and reliability of the information, which is crucial for building the model. The websites include:

- Long Châu: Known as one of the largest pharmacy chains in Vietnam, Long Châu not only provides pharmaceuticals but also shares a wealth of health knowledge.

- Medlatec: A reputable private healthcare system with many years of experience, Medlatec offers diagnostic and treatment services, while also publishing in-depth articles on health issues.

- Tâm Anh General Hospital: A well-known private general hospital with a team of skilled doctors and modern equipment, Tâm Anh provides authoritative and up-to-date medical information on its website.

- Vinmec: A high-quality private healthcare system under the Vingroup conglomerate. Vinmec is renowned for its excellent service and top experts, offering in-depth and reliable medical information.

The articles collected from these websites are distributed as shown in the figure below:

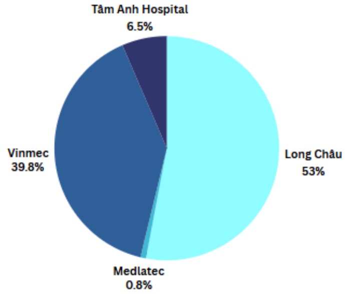
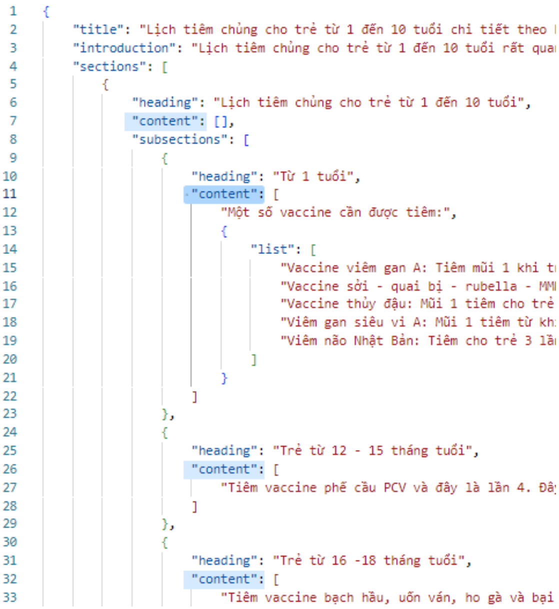


Figure 2: Distribution of articles collected from the 4 health websites

The articles are formatted in JSON files, with each article containing: "title" - the title of the article, "introduction" - a brief introduction to the article content, "heading" - the main sections of the article, and "content" - the main content of the article.

Additionally, to evaluate the model's performance, we created a ground truth Q&A dataset



```
{
    "title": "Lịch tiêm chủng cho trẻ từ 1 đến 10 tuổi chi tiết theo
    "introduction": "Lịch tiêm chủng cho trẻ từ 1 đến 10 tuổi rất qua
    "sections": [
        {
            "heading": "Lịch tiêm chủng cho trẻ từ 1 đến 10 tuổi",
            "content": [],
            "subsections": [
                {
                    "heading": "Từ 1 tuổi",
                    "content": [
                        "Một số vaccine cần được tiêm:",
                        {
                            "list": [
                                "Vaccine viêm gan A: Tiêm mũi 1 khi t
                                "Vaccine sởi - quai bị - rubella - MM
                                "Vaccine thủy đậu: Mũi 1 tiêm cho trẻ
                                "Viêm gan siêu vi A: Mũi 1 tiêm từ kh
                                "Viêm não Nhật Bản: Tiêm cho trẻ 3 lầ
                            ]
                        }
                    ]
                },
                {
                    "heading": "Trẻ từ 12 - 15 tháng tuổi",
                    "content": [
                        "Tiêm vaccine phế cầu PCV và đây là lần 4. Đâ
                    ]
                },
                {
                    "heading": "Trẻ từ 16 -18 tháng tuổi",
                    "content": [
                        "Tiêm vaccine bạch hầu, uốn ván, ho gà và bại
```

Figure 3: Example structure of an article in JSON format

based on the collected data. Initially, a Large Language Model (LLM) was used to generate potential question-answer pairs from chunks of text (smaller segments) derived from the articles. The LLM used here is Gemini 2.0. For example, from a paragraph about contraceptive methods, Gemini might generate the question "What are the common contraceptive methods today?" with the corresponding answer derived from the paragraph. From the 79,037 articles collected, 317,023 chunks were generated. We then used 500 chunks to create the ground truth dataset, resulting in 4,121 Q&A pairs.

# 4 Knowledge base

## 4.1 Retrieval-Augmented Generation (RAG)

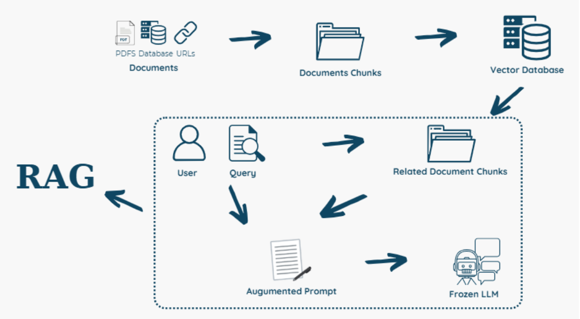


Figure 4: RAG

Retrieval-Augmented Generation (RAG) (Lewis et al., 2021) is a technique that enhances large language models (LLMs) by connecting them to external knowledge sources. Instead of relying solely on knowledge acquired during training, RAG enables LLMs to retrieve relevant information from databases or document repositories and use it to generate more accurate and informative responses. This process typically involves converting data into vectors, using these vectors to search for relevant information based on user queries, and finally employing the LLM to combine the retrieved information with the original query to produce an answer. This allows LLMs to overcome limitations of fixed knowledge and information update capabilities, while also improving the reliability and accuracy of responses.

RAG operates with two main modules:

### 4.1.1 Retrieval Module

This module retrieves relevant information from databases or document repositories based on user queries, using a combination of methods:

- **BM25 (Best Matching 25) (Murthy and Reys, 2016):** A term frequency-based ranking algorithm for quickly identifying documents containing query keywords, returning a large set of potential candidates (e.g., top 100).

- **Sentence Embeddings:** Models convert queries and documents into numerical vectors capturing their semantic meaning. Similarity between these vectors (e.g., cosine similarity) assesses semantic relevance between queries and documents, narrowing down the search from BM25 results.

- **Cross-Encoder Reranker (Sheka, 2019):** A Cross-Encoder model reranks the candidate list from BM25 and Sentence Embeddings for higher accuracy. It considers both the query and each document simultaneously, evaluating their relevance in the specific context to select a small set of the most relevant documents (e.g., top 5).

### 4.1.2 Large Language Model (LLM) Module

This module receives the original user query and the content of the retrieved documents (e.g., top 5). The LLM uses this information to generate an answer or response. BERT (Bidirectional Encoder Representations from Transformers)(Devlin et al., 2019) is used to enhance the process of retrieving and processing information from the retrieved documents. BERT helps the model capture bidirectional context, improving its ability to understand and generate more accurate responses to user queries. By integrating BERT into the query and document content processing, we are able to generate detailed, accurate, and grounded answers based on the retrieved documents, minimizing the occurrence of hallucinations or incorrect responses, which are common in models that rely solely on training data.

### 4.2 Experimental Model Setup

To evaluate the performance of RAG, we conducted experiments with specific retrieval and LLM configurations. The following subsections detail the models and preprocessing techniques used in our experiments.

#### 4.2.1 Sentence Embeddings

We employ two specific sentence embedding models:

- **dangvantuan/sentence-embedding:** A model designed for general sentence embeddings, which may implicitly handle word segmentation.
- **dangvantuan/sentence-document-embedding:** A model specifically trained to handle document-level embeddings. It is more robust to variations in text length and structure, potentially offering advantages in scenarios where explicit word segmentation is less critical or even detrimental.

By comparing the performance of these two models, we aim to understand the influence of tokenization strategies on the effectiveness of semantic retrieval within our RAG system.

#### 4.2.2 Cross-Encoder Reranker

We use two distinct reranking models:

- **cross-encoder/mmarco-mMiniLMv2-L12-H384-v1:** A multilingual cross-encoder designed to handle various languages without requiring explicit word segmentation or tokenization. This model is suitable for cross-lingual information retrieval and scenarios where preprocessing steps like tokenization might introduce errors or inconsistencies.
- **PhoRanker (Ba, 2024):** A model specifically trained for the Vietnamese language. It relies on accurate word segmentation for optimal performance. Input to **PhoRanker** is preprocessed using a Vietnamese tokenizer.

This comparison allows us to investigate the impact of language-specific training and preprocessing requirements on the effectiveness of reranking within our RAG system.

#### 4.2.3 Large Language Models

We evaluate the performance of several prominent LLMs within this RAG framework:

- **GPT-4:** A leading large language model known for its broad generalization capabilities.
- **Gemini 2.0:** A next-generation language model optimized for multilingual and contextual understanding.
- **phamhai/Llama-3.2-3B-Instruct-Frog:** An open-source model designed for generating contextually relevant and factual answers.

This comparative analysis aims to provide insights into the strengths and weaknesses of different LLM architectures and their ability to leverage retrieved information for enhanced response generation.

## 5 Experimental results

For performance evaluation, we conduct separate assessments of the two core modules: retrieval and generation.

### 5.1 Module retrieval

Table 1 demonstrates the superior performance of BM25 compared to sentence embeddings. This is attributable to the significance of keywords within the health domain and the fact that the sentence embeddings were employed without fine-tuning. Consequently, BM25 was selected as the primary retrieval method, and a cross-reranker was subsequently applied. The results presented in Table 2 indicate that the multilingual cross-reranker outperformed the Vietnamese tokenized PhoRanker.

### 5.2 Module generator

Owing to resource limitations, the evaluation of this module was conducted using a sample of 50 questions. An answer was deemed correct if its semantic meaning aligned with the answer present

| **Retrieve** | **Top 1** | **Top 5** | **Top 10** | **Top 20** | **Top 30** | **Top 50** | **Top 100** |
|---|---|---|---|---|---|---|---|
| BM25-NLTK | 0.34 | 0.58 | 0.66 | 0.74 | 0.78 | 0.83 | 0.88 |
| **BM25-pyvi** | **0.38** | **0.62** | **0.70** | **0.78** | **0.82** | **0.86** | **0.91** |
| dangvantuan/sentence-embedding-pyvi | 0.15 | 0.33 | 0.42 | 0.50 | 0.56 | 0.63 | 0.71 |
| **dangvantuan/sentence-document-embedding** | **0.36** | **0.59** | **0.69** | **0.73** | **0.77** | **0.81** | **0.86** |

Table 1: Result on retrieval module

| **Reranker** | **Top-1** | **Top-2** | **Top-3** | **Top-4** | **Top-5** | **Top-10** | **Top-20** |
|---|---|---|---|---|---|---|---|
| **BM25-cross-encoder/mmarco-mMiniLMv2-L12-H384-v1** | **0.49** | **0.6** | **0.67** | **0.7** | **0.72** | **0.79** | **0.85** |
| BM25-PhoRanker | 0.44 | 0.55 | 0.61 | 4 | 0.67 | 0.74 | 0.81 |

Table 2: Retrieval Performance for Different Rerankers

in the retrieved documents. Following document retrieval, the top 5 most relevant documents were provided as context to the large language models. The results presented in Table 3 indicate that with comprehensive context, models such as GPT and Gemini achieve near-perfect accuracy, effectively addressing instances of hallucination observed when the context lacks a direct answer.

| **RAG QA** | **Accuracy** |
|---|---|
| GPT-4o mini | 50/50 |
| Gemini 1.5 Flash | 50/50 |
| phamhai/Llama-3.2-3B-Instruct-Frog | 40/50 |

Table 3: Accuracy comparision of RAG QA models

## 6 Conclusion

This study demonstrates the effectiveness of using Retrieval-Augmented Generation (RAG) for healthcare queries in the Vietnamese language. By combining a powerful LLM with relevant external information, RAG minimizes hallucinations and produces accurate, grounded answers. The approach ensures that responses are not only precise but also based on real-world knowledge, addressing the limitations of LLMs relying solely on training data.

Future work will focus on fine-tuning embedding models and cross-encoders for better retrieval of relevant documents. Additionally, fine-tuning the LLM for RAG tasks will optimize response generation. We will also explore alternative retrieval methods and enhance the dataset to cover a wider range of healthcare topics, improving the model's ability to answer complex health-related questions in Vietnamese.

## Acknowledgement

We would like to express our sincere gratitude to Professor Nguyễn Hiếu Nghĩa and Professor Nguyễn Văn Kiệt for their invaluable guidance and support throughout the course of this project. Their constant encouragement, insightful feedback, and dedicated mentorship have been essential to the successful completion of this work.

## References

Dai Nguyen Ba. 2024. Phoranker: A cross-encoder model for vietnamese text ranking. https://huggingface.co/itdainb/PhoRanker.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Sameer Murthy and Valentin Reys. 2016. Single-centered black hole microstate degeneracies from instantons in supergravity. *Journal of High Energy Physics*, 2016(4):1–33.

Elena F. Sheka. 2019. Spin chemistry of sp2 nanocarbons.