# MedStudio: A Comprehensive Intelligent Medical Platform for Vietnamese Users Integrating Natural Language Processing and Computer Vision*

Minh-Trieu Truong[1], Quang-Khai Hoang[1], Van-Dung Hoang[1*]

[1]Faculty of Information Technology, HCMC University of Technology and Education

*Corresponding author: dunghv@hcmute.edu.vn

**Abstract**

This paper introduces MedStudio - A Comprehensive Intelligent Medical Platform for Vietnamese Users, integrating advanced deep learning techniques, including large language models (LLMs) and vision language models (VLMs). MedStudio tackles key medical information challenges: medical information retrieval, question answering, and specialized machine translation (MT). It uniquely combines medical image classification, lesion segmentation, and medical visual Q&A. Critically, MedStudio strongly supports Vietnamese, facilitated by English-Vietnamese machine translation. We also introduce VietMeD: Vietnamese Pharmaceuticals Dataset, a high-quality Vietnamese drug dataset foundational for future expansion. The system's multi-modular, multi-modal architecture emphasizes efficient data handling, intelligent query processing, and an Advanced Retrieval-Augmented Generation (RAG) approach for accurate, context-aware responses.

*This manuscript received AI assistance solely for translating content from Vietnamese to English. However, at least 98% of the manuscript was authored entirely by us, WITHOUT the use of any AI tools or plagiarism.

# 1 Introduction

## 1.1 Problem Statement and Motivation

Information retrieval and question answering (QA) systems using natural language are some of the most actively researched areas in natural language processing (NLP). With rapid advancements, QA systems are increasingly applied in daily life for efficient and natural access to vast amounts of information. To achieve higher accuracy in information retrieval, these models must comprehend the underlying meaning and context of queries, going beyond traditional keyword-based methods, which often fail to retrieve relevant content. This understanding of context and semantics allows models to retrieve information more effectively from massive datasets.

However, despite significant breakthroughs in AI impacting various sectors from manufacturing and economics to education and healthcare, a substantial disparity exists in the applicability of these advancements across languages. While computer vision applications can often be deployed globally without language barriers, NLP heavily depends on language-specific resources. Consequently, many groundbreaking NLP innovations, particularly in tasks like text generation, language modeling, and information retrieval, are primarily trained and evaluated on large English datasets. This leads to excellent computational power and language modeling capabilities for English but poor performance when applied to other languages, such as Vietnamese. Thus, despite rapid progress, the practical impact of NLP on Vietnamese tasks remains limited, often requiring robust methods for cross-lingual transfer or mapping between English and Vietnamese. This underscores the critical role of machine translation (MT).

Machine translation, especially neural machine translation (NMT), serves as a linguistic bridge, enabling the application of advanced NLP achievements to language-dependent tasks. To effectively leverage generative models and large language models, Vietnamese source texts often need to be translated into English. Similarly, for information retrieval and semantic similarity-based document retrieval tasks that utilize word embeddings, most effective embedding models are trained and processed predominantly in English. Applying these models directly to languages like Viet-

namese, which are either untrained or have very limited pre-training data, yields poor results. Consequently, a promising approach involves translating all queries from the source language to an intermediate language like English for processing (e.g., query/retrieval), and then back-translating the results to the source language for the user. While multilingual pre-trained models are emerging and showing state-of-the-art performance, they present significant challenges, including complex and resource-intensive data collection and preprocessing, the need for exceptionally high-quality and large-scale bilingual datasets for various language pairs, and extremely high computational resources and training times. Therefore, the approach of translating source language queries to a popular intermediate language like English still demonstrates significant potential and strong applicability. If the translation model's performance is sufficiently high, the overall system's effectiveness can be substantially improved. For these reasons, this research focuses on investigating, analyzing, and evaluating machine translation models, particularly neural-based models, for English-Vietnamese language pairs. We believe that robust MT modules will play a core role as a bridge for transferring and applying impressive results from general NLP techniques to language modeling tasks specifically for Vietnamese speakers.

Furthermore, applying MT to medical data presents unique challenges. Medical translation involves specialized terminology, such as disease names, drug names, and symptoms. The quality of the translation heavily depends on the accurate rendition of these specific terms. This characteristic means that even MT models that perform exceptionally well in general domains may struggle significantly or fail entirely when translating medical texts. Therefore, in addition to investigating the network architectures of NMT models, we will also focus on translation quality and techniques to improve the quality of medical translations. Specifically, we will dedicate significant effort to researching and enhancing MT models for medical translation, focusing on medical data and experimenting with large English-Vietnamese bilingual datasets in fields like diseases, drugs, and clinical information.

## 1.2 Research Objectives

This research aims to investigate, analyze, and understand the network architectures of language models, large language models, and machine translation models, as well as their integration to address issues related to medical information retrieval in a specialized linguistic domain. This work will serve as a foundational basis for other NLP tasks focusing on the medical field, exemplified by the medical question answering task.

Our research also focuses on general machine translation-based solutions that can be fine-tuned and improved for various linguistic domains. Currently, we are experimentally implementing English-Vietnamese translation and, for application purposes, aiming to build and develop an intelligent medical platform to solve medical-related problems for Vietnamese users.

Additionally, we will explore and consider medical image classification and lesion segmentation models to develop an application capable of automatically delineating lesions and integrating the aforementioned question answering module. This integration will enable users to query information about lesions in provided images. Finally, we will investigate several Visual Language Models (VLMs) to build a medical visual question answering module. This is a novel, challenging, yet highly promising field for the future. Research and development of VLMs can significantly advance medical visual QA, allowing doctors, students, or users to inquire about all information embedded in medical images. Both of these modules will be researched, developed, and integrated into a unified platform with the medical QA and MT modules. Our aspiration is to develop an application that leverages and connects advanced techniques from both NLP and computer vision to solve medical-related problems, serving as a bridge for applying the rapid advancements in artificial intelligence, especially deep learning, to healthcare, thereby promoting the development of medical technology.

## 1.3 Scope and Methodology

### 1.3.1 Research Subjects

The primary research subjects in this study will focus on language models, large language models, text generation models, visual language models, and neural machine translation models. Furthermore, the collection, analysis, and evaluation of results on experimental datasets will also be among our contributions. The main experimental datasets in this study will focus on information retrieval and machine translation tasks.

### 1.3.2 Research Scope

In this study, we will research and build an intelligent medical platform that addresses key problems in medical-related natural language processing and computer vision. Specifically, we will focus on:

- Medical information retrieval techniques.
- Language models and large language models.
- Models for neural machine translation.
- Medical image classification models.
- Medical image lesion segmentation models.
- Models for medical visual question answering.

### 1.3.3 Research Methodology

The main research methodology for this study will involve theoretical research, experimentation, analysis, and evaluation of results:

- **Theoretical Research Method:** Through national and international scientific papers, literature, and related studies.

- **Experimental Method:** Developing programs, implementing them in practice, and adjusting them based on results to achieve the highest accuracy.
- **Analysis and Evaluation Method:** Analyzing experimental results, calculating program accuracy through practical implementation, and proposing solutions to improve accuracy and performance.

## 1.4 Key Efforts and Contributions

Our main efforts and contributions in this research include:

- Researching and developing a medical information retrieval module.
- Researching and developing a medical machine translation module.
- Researching and developing a medical question answering module.
- Researching and developing a medical visual question answering module.
- Researching and developing a medical image processing module, including sub-modules for:
  - Medical image classification.
  - Medical image lesion segmentation.
- Researching and constructing the ”Vietnamese Pharmaceuticals Dataset” (VietMeD).

## 1.5 Project Goals and Expected Outcomes

This project will focus on investigating and researching:

- Language models and machine translation models to solve medical-related language processing problems.
- Image classification and segmentation models to solve medical-related computer vision problems.

- Visual language models to solve medical visual question answering problems.

Subsequently, we will apply these models to build and develop solutions for medical-related natural language processing problems for Vietnamese users. The expected outcomes of this project include the development of an all-in-one platform called MedStudio, an open-source Vietnamese-English machine translation library named VietTranslate, and a Vietnamese pharmaceutical dataset called VietMeD, where:

- MedStudio is an open-source intelligent online medical platform integrating advanced language models and deep learning techniques.
- VietTranslate is an open-source neural machine translation library including implementations of several pre-trained models and some fine-tuned models developed by us.
- VietMeD is a high-quality, standardized Vietnamese pharmaceutical dataset containing information on approximately 200 types of drugs and functional foods, manually constructed and preprocessed from reputable Vietnamese pharmaceutical sources.

Both the MedStudio platform and the VietTranslate library are open-source, while the VietMeD dataset currently has restricted access, available only for research and development purposes.

## 2 Proposed Method and Our MedStudio System

In this research, we have developed and deployed an intelligent online medical platform: MedStudio—an online intelligent medical platform.

MedStudio is an intelligent online medical platform integrating advanced deep learning techniques. The core of the system leverages the power of large language models and state-of-the-art neural machine translation models to address natural language processing tasks related to medical data, including:

- Medical information retrieval.

- Medical question answering.
- Specialized medical machine translation.

Furthermore, we have researched and developed two modules integrating Vision Language Models (VLMs) to address hybrid NLP and computer vision tasks, specifically:

- Medical visual question answering.
- Medical image classification combined with medical lesion segmentation, with result generation integrated with disease information retrieval for medical disease question answering.

With the integration of neural machine translation architectures, MedStudio aims to support specialized user languages, with fine-tuning capabilities for different national linguistic domains. Within the scope of this research, we will primarily focus on English-Vietnamese machine translation, positioning the system as an intelligent medical platform specifically for Vietnamese users with robust support for the Vietnamese language.

Additionally, as part of our efforts in this research, we have collected, preprocessed, and manually constructed the ”VietMeD: Vietnamese Pharmaceuticals Dataset.” This is a standardized dataset in Vietnamese containing information on approximately 200 types of drugs and functional foods, built and preprocessed manually from reputable Vietnamese pharmacopoeia sources. The Vietnamese Pharmaceuticals Dataset is a high-quality dataset, and its construction lays the groundwork for building a larger and more foundational medical dataset in Vietnamese in the future.

## 2.1 System Overview Architecture

MedStudio’s overall architecture is a multi-module platform capable of a multi-modal approach. Specifically, the MedStudio system architecture comprises the following modules:

- Medical Information Retrieval Module.
- Medical Machine Translation Module.

- Medical Question Answering Module, including the following sub-modules:
  - Medical Data Collection and Preprocessing Module.
  - Data Indexing and Storage Module.
  - Medical Machine Translation Module.
  - Query Pre-processing and Processing Module.
  - Medical Information Retrieval Module.
  - Answer Generation Module from Data and Query Context.
- Lesion Segmentation and Medical Visual Question Answering Module, including the following sub-modules:
  - Medical Visual Question Answering Module.
  - Medical Image Classification Module.
  - Lesion Segmentation in Medical Images Module.
- Drug Information Searching and Exploring Based on the VietMeD dataset.

## 2.2 Detailed System Architecture

### 2.2.1 Medical Question Answering Module

The Medical Question Answering (QA) Module is one of the largest modules in this study. Its overall architecture is illustrated in Figure 1.

This module leverages the powerful semantic search capabilities of the information retrieval module and the natural language understanding and semantic analysis prowess of language models to generate context-specific answers. Specifically, the information extraction module receives user queries, processes them, and returns relevant passages or question-answer pairs from the data repository that exhibit high semantic similarity to the query. These retrieved results serve

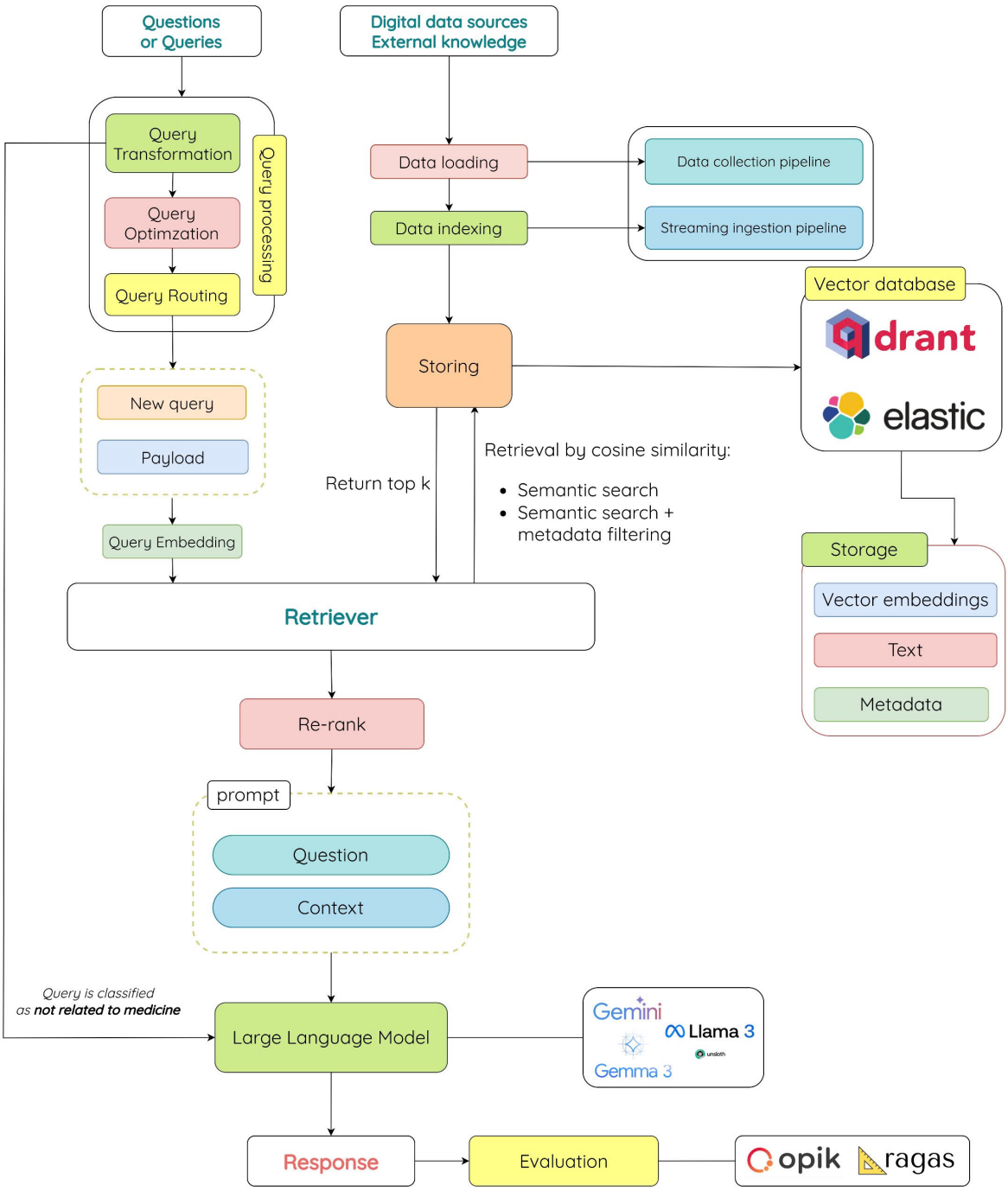


Figure 1: Architecture of the Medical Q&A Module.

as context, combined with the original user query, to be fed into large language models (LLMs) pre-trained on extensive datasets to generate the final answer for the user.

For Question Answering (QA) and Text Generation systems, two particularly prominent approaches are systems using the Retrieval-Augmented Generation (RAG) architecture and those that fine-tune pre-trained large language models (LLMs) on domain-specific datasets. Furthermore, many recent studies have proposed novel and promising methods and techniques, such as GraphRAG, or Medical GraphRAG for the medical data domain, as well as Small Language Models (SLMs) and state-space models, exemplified by Mamba. However, most published solutions remain largely at the research level, with limited practical application or effectiveness confined to specific task groups and data domains. This is primarily due to these methods not yet being robust enough for high efficacy in real-world applications, partly because they face numerous challenges, including computational cost, resource consumption, and deployment complexity.

Through our research, we have determined that the RAG architecture remains the most optimal choice. We believe that current language models are highly powerful, possessing excellent language understanding and semantic analysis capabilities. Moreover, these models are continuously being improved and developed daily, trending towards even greater completeness and linguistic comprehension. Therefore, we posit that if these models are provided with accurate information and context, they can generate highly accurate and semantically natural answers. Thus, we identify the key factor for our system's strong performance as the high semantic similarity and relatedness of the retrieved results (used as context) to the user's query. Achieving this depends on two influencing factors: first, the user's query must be truly clear and encapsulate the full intent or information the user seeks; second, the information retrieval capability of the retrieval block.

To ensure that user queries are genuinely clear and fully encapsulate the user's information and intent, we have developed a sub-module for query processing and preprocessing to normalize and refine the original user query before retrieval. Furthermore, to enhance the information retrieval capabilities of the information retrieval module, in addition to using semantic search techniques to boost retrieval performance, we have also built a module for data collection and preprocessing,

and a module for indexing and loading data into vector databases.

Additionally, we have integrated the Medical Machine Translation Module as a sub-module within this Medical Question Answering Module. This allows for the translation of user queries and questions from Vietnamese into English for processing, information retrieval, and answer generation. Subsequently, the results are back-translated to the source language before being sent back to users interacting with the platform in languages other than English, specifically focusing on Vietnamese users.

Furthermore, in addition to the main modules introduced above, the Medical Question Answering Module we built also contains other supplementary components such as rerank and evaluation components. These blocks are designed to improve information retrieval results or evaluating the overall system's performance to plan for further system improvement, development, and refinement.

**Medical Data Collection, Preprocessing, and Loading Module** In this module, all medical data from digital sources, datasets, and external knowledge bases will be collected, aggregated, preprocessed, and loaded into NoSQL databases. This process primarily involves cleaning and loading usable data into databases, also known as data repositories. In this research, we will utilize both digital and external data sources, including public datasets, data we have collected, and the drug dataset we have built. Our data will comprise three categories: Question-Answer (QnA) pairs, Medical passages (which may be single paragraphs or multi-paragraph texts), and the Drug data we collected. For each data type, we employ an ETL pipeline (Extract, Transform, Load) to process the data. Key processing techniques include data cleaning and normalization. Finally, after the initial raw data passes through the ETL pipelines, we obtain high-quality processed data. We then proceed to load this data into our primary data stores, which are NoSQL databases.

A unique aspect here, and in this research, is that we have identified a "source" characteristic in our original dataset. This characteristic indicates that each data sample or record in our dataset belongs to a different topic, such as neurological diseases or digestive system diseases. Specifically,

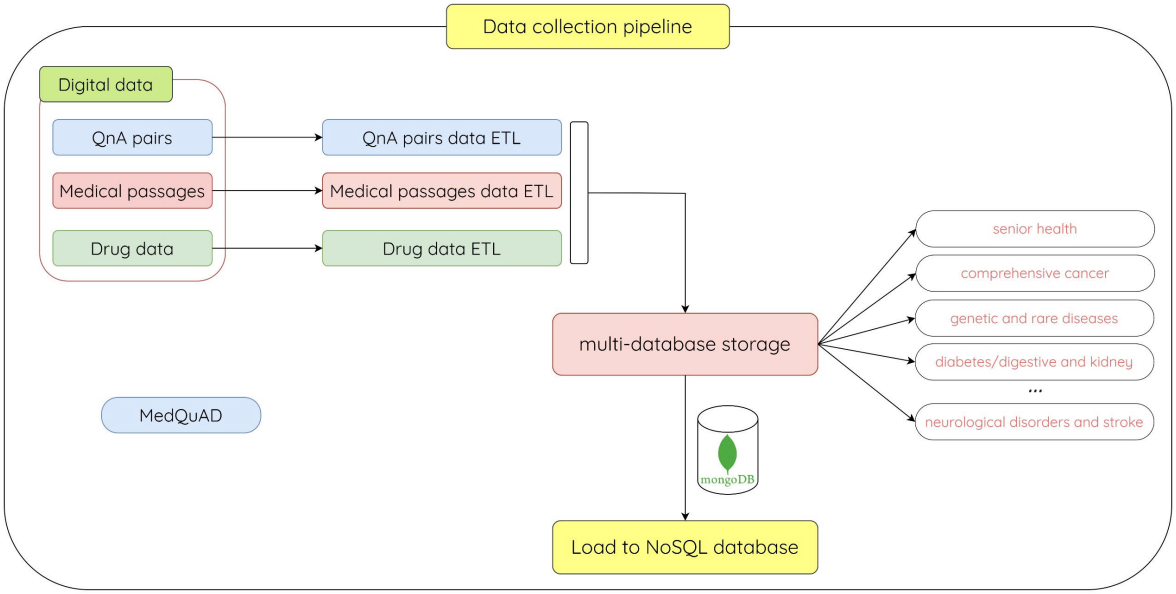


Figure 2: Pipeline of the Medical Data Collection Module.

after data analysis, we found that each data sample can be categorized into one of nine topics:

- Senior Health.
- Comprehensive Cancer.
- Medline Plus, Health Topics.
- Genetic and Rare Diseases (GARD).
- Disease Control and Prevention (CDC).
- Heart, Lung, and Blood (NHLBI).
- Neurological Disorders and Stroke (NINDS).
- Growth Hormone Receptor (GHR).
- Diabetes and Digestive and Kidney Diseases (NIDDK).

Therefore, we decided to load and store data in the database using a multi-database storage mechanism, where each topic (e.g., senior health, genetic and rare diseases) will be stored in

a unique and separate database instance. In this study, we selected MongoDB as our NoSQL database, with each database instance stored in a distinct collection. This storage mechanism significantly optimizes the performance of our information retrieval module. Firstly, if we can classify the user's query by topic or if the user explicitly wants to query information on a specific topic (which is very useful for doctors or medical professionals interested in a particular specialty), these queries will be processed for topic classification or will be routed based on user-selected topics. Subsequently, the query will be directed to query only one appropriate database instance for that topic. This significantly reduces query time due to a smaller query sample space and substantially increases retrieval efficiency as it focuses on a specific specialty without being affected by data noise from other specialties. However, to achieve this, the challenge lies in accurately classifying the original data into topic categories and precisely classifying the queries. Its overall pipeline is illustrated in Figure 2.

**Data Indexing and Storage Module** In this module, the preprocessed data, currently stored in NoSQL databases, will be loaded, processed, with important data features selected, and then indexed (or "chunked") to create vector embeddings for each content segment. These vector embeddings will then be loaded and stored in vector databases, along with necessary metadata and data features for the information retrieval and extraction module. Its overall pipeline is illustrated in Figure 3.

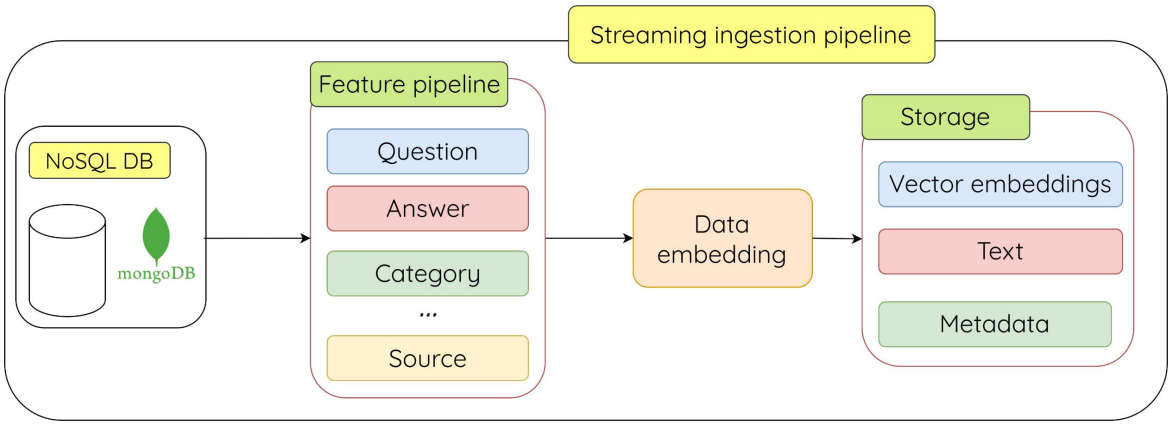


Figure 3: Pipeline of the Medical Data Streaming Ingestion Q&A Module.

A particular aspect of the data features is that the generated embedding vectors, after being embedded by embedding models, will be stored as dense vectors. When loading this data into databases, configuring the vector data type and its dimensions is essential and must be predefined. Therefore, we need to clearly determine the dimension of the embedding vector to be stored. The vector's dimension will depend on the embedding models used, as each model produces output vectors of different sizes (e.g., 384, 768, or 1024). Additionally, storing metadata is crucial. Combining semantic search based on embedding vectors with metadata filtering will achieve significantly improved efficiency compared to implementing each search technique independently.

**Medical Machine Translation Module** In this module, user queries or questions made in Vietnamese will be translated into English for processing through the system's steps, such as information retrieval and result generation. Subsequently, the retrieved results will be back-translated to the source language before being sent back to users who interact with the platform using source languages other than English. Since this module is also used to build a standalone specialized medical translation system within the MedStudio platform, a more detailed description of this module will be provided in subsequent sections.

**Query Pre-processing and Processing Module** To ensure that user queries are clear and accurately represent the user's intent and information need, we have designed a dedicated pipeline to preprocess and process input queries. This pipeline includes three main stages:

- **Query Transformation:** This stage normalizes user input and determines whether the query is medically relevant. The overall process is illustrated in Figure 4.
- **Query Optimization:** Medically relevant queries are further refined. Complex or multi-part queries are split into simpler, atomic ones (query expansion). Conversely, overly simple queries are expanded with additional contextually similar queries (query augmentation). See Figure 5.

- **Query Routing:** Refined queries are categorized by topic and intent using logical and semantic routing. See Figure 6.

**Query Transformation** Query Transformation is the first stage in the pipeline and focuses on handling raw user input. It comprises two layers: **Query Normalization** and **Query Classification**.

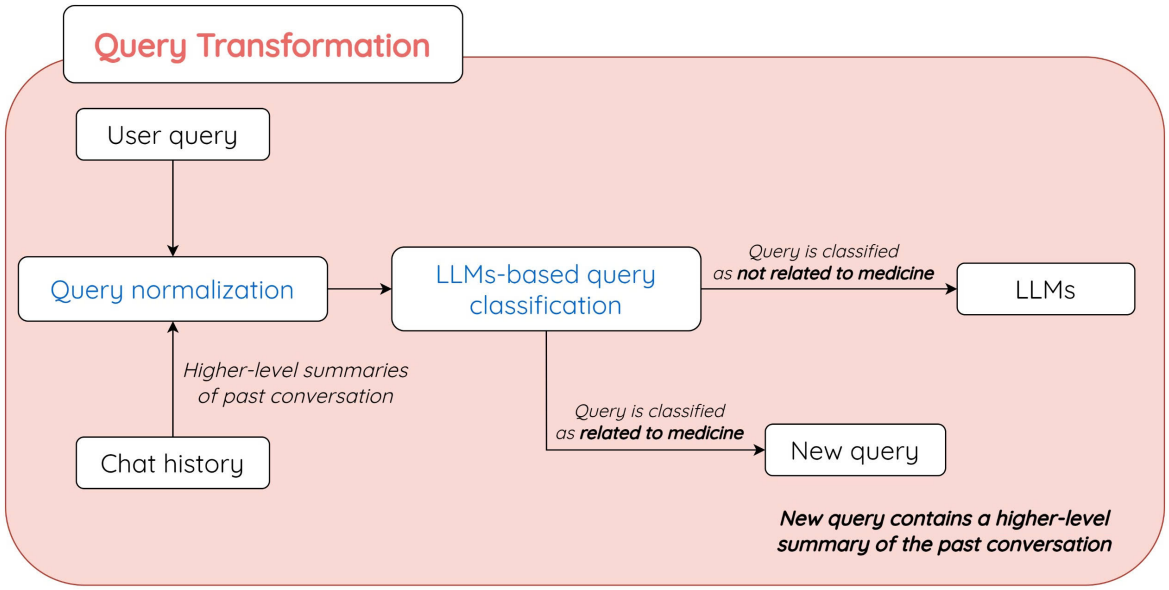


Figure 4: Pipeline of the Query Transformation Module.

**Query Normalization** standardizes the structure and content of the query. It may use prior conversation history to resolve ambiguity or incomplete information.

**Query Classification** determines if the query is medically relevant. It assigns one of three labels:

- `disease_info`: Clearly medical and disease-specific queries.
- `diagnosis_query`: Medical-related but general diagnostic questions.
- `general`: Non-medical or unrelated queries.

Only queries labeled as `disease_info` or `diagnosis_query` are passed to the next module. For example:

- Input: ”Give me some information about Glaucoma.”
- After normalization: ”What is Glaucoma?”
- Classification label: `disease_info`

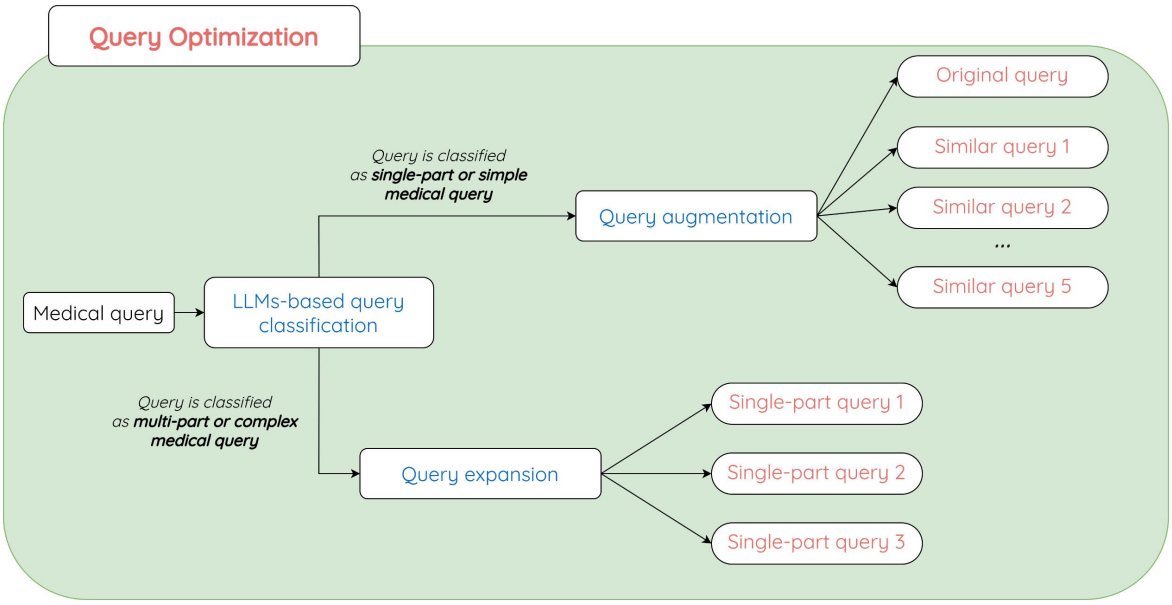


Figure 5: Pipeline of the Query Optimization Module.

**Query Optimization** In this stage, normalized and relevant queries are either expanded or augmented depending on their complexity. . It comprises two layers: **Query Expansion** and **Query Augmentation**.

**Query Expansion** splits complex queries into atomic sub-queries.

**Query Augmentation** enhances simple queries by generating semantically similar alternatives.

**Example – Query Expansion:**

- Input: ”What is dengue fever, what are its symptoms, how can it be diagnosed, what is the treatment, and what are the complications it leaves behind?”
- Output:

- ”What is dengue fever?”
- ”What are the symptoms of dengue fever?”
- ”How is dengue fever diagnosed?”
- ”What is the treatment for dengue fever?”
- ”What are the complications of dengue fever?”

**Example – Query Augmentation:**

- Input: ”What is thalassemia?”
- Output:
  - ”What is thalassemia?”
  - ”What are the symptoms of thalassemia?”
  - ”How is thalassemia treated?”
  - ”What is the treatment for thalassemia?”

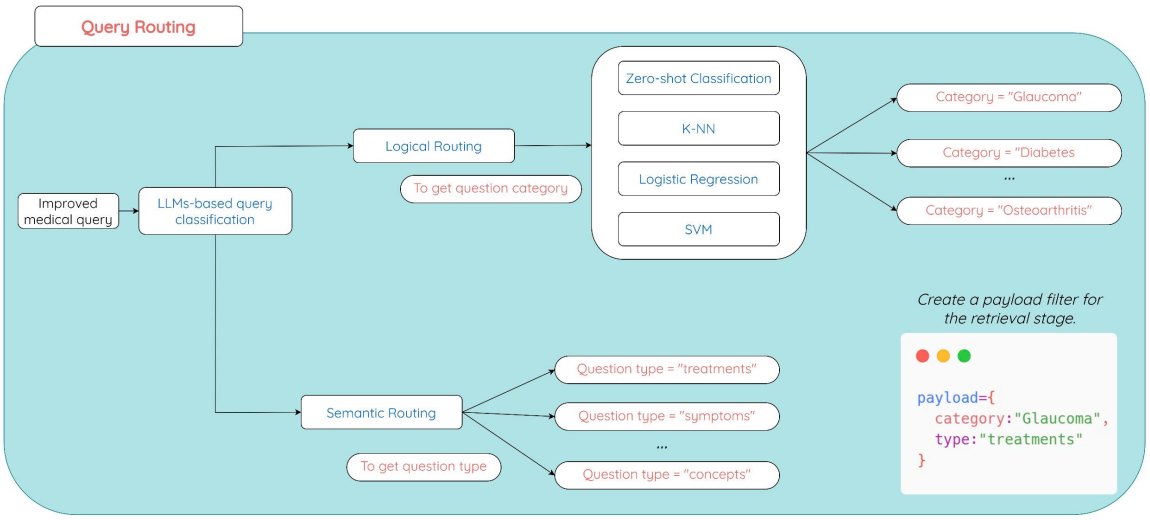


Figure 6: Pipeline of the Query Routing Module.

**Query Routing** It comprises two layers: **Logical Routing** and **Semantic Routing**.

**Logical Routing** categorizes the query by disease/topic using classification models such as Zero-shot Classification, Logistic Regression, SVM, and K-NN.

**Semantic Routing** identifies the query's intent (e.g., treatment, symptoms, complications) based on sentence similarity using cosine similarity with predefined templates.

**Examples – Semantic Routing:**

- Input: "What are the treatments for Hemolytic Uremic Syndrome in Children?"
- Output: Intent: `treatments`, Confidence: 0.6342
- Input: "Can you tell me about the treatments for Hemolytic Uremic Syndrome in Children?"
- Output: Intent: `treatments`, Confidence: 0.6691
- Input: "What are the complications of Mineral and Bone Disorder in Chronic Kidney Disease?"
- Output: Intent: `complications`, Confidence: 0.7323

**Medical Information Retrieval Module** This module will be used to retrieve information relevant to the user's query from the medical data sources stored in the corpus. Semantic search techniques will be employed in this module, and its input will be the processed, optimized, and medically relevant queries. Since this module is also used to build a separate function for medical information retrieval and lookup within the MedStudio platform, a more detailed and clearer description of this module will be provided in subsequent sections.

**Answer Generation Module from Retrieved Data and Context** In this module, language models and large language models (LLMs) are used for generation based on the information retrieved from the retrieval module and the original input query from the user. These two pieces of information are combined to help the language models produce appropriate responses to the input request.

The input query processing step requires several sub-steps to proceed to subsequent stages or to conclude a final answer generation process for the user. The information will be integrated and added to the prompt. In this context, a prompt is a part of the language model, used as input content comprising requests, queries, and synthesized information. The language model then relies on these requests and synthesized content to generate a suitable answer based on the input information from the prompt.

The results from the language models can be used to analyze information for subsequent processing steps. There are some considerations for prompt writing to avoid generating undesirable responses for subsequent processing steps. Therefore, providing specific contexts or examples is necessary to guide the model to generate results in the desired direction. For prompt content used for generation based on retrieved information, the prompt must explicitly request that the answer only be generated based on the provided content.

### 2.2.2 Medical Information Retrieval Module

This module is responsible for querying and retrieving information about diseases, drugs, and related medical data from the medical data repository based on user queries. The approach of this research is to use vector databases to store the medical corpus and employ semantic search techniques for information retrieval. In this module, two vector databases, Elasticsearch and Qdrant, are used to store data and execute queries. These are both excellent tools for searching and querying textual data, and both support semantic search to find information based on word meanings.

Semantic search typically relies on calculations involving cosine similarity and dot product. When initializing the vector database, we are allowed to declare the choice of similarity calculation, with cosine similarity often being the optimal and most popular choice for determining the similarity between data points. These calculations compute to find the vectors closest to the input query and return the corresponding documents. For each stored document, descriptive information (content and metadata) is retrieved from the data repository, then normalized and converted into a single text segment called ”context.” This ”context” is then used as input to the answer generation

module to produce a response to the user's question. Additionally, both the Qdrant and Elasticsearch libraries support filtering to narrow down the vector search scope. Filters are used to restrict the information content of each document. In our system, filters are used to filter based on two attributes corresponding to the "question_type" or "question_category" data fields. Accurate information must be entered for this filter to ensure precise retrieval. Therefore, when using a filter, for queries where these two pieces of information can be clearly identified before being passed to the filter and search, the retrieval results will have more accurate content when creating context for the LLM. It is important to note that for open-ended or ambiguous queries within medical information categories, filters should not be used to ensure better exploratory capabilities within the medical literature.

### 2.2.3 Medical Machine Translation Module

In this study, to build the specialized medical machine translation module, we explored neural machine translation approaches, specifically focusing on leveraging and applying pre-trained models. These models are typically pre-trained on very large bilingual datasets and can be used immediately. Following this, we investigated and implemented techniques to fine-tune these pre-trained models on a specialized domain, specifically medical translation datasets between Vietnamese and English.

Specifically, after thorough research and consideration, we chose two models: envit5-translation and mBART-50 many-to-many multilingual machine translation. The envit5-translation is a fine-tuned version of the T5 model on Vietnamese-English bilingual machine translation datasets, MTet and PhoMT, from the VietAI research group. The mBART-50 many-to-many multilingual machine translation is a fine-tuned version of the mBART-large-50 model for multilingual machine translation from Facebook's neural translation research group. We then fine-tuned these two models on another Vietnamese-English bilingual dataset specifically dedicated to medical content. This dataset comprises translation pairs primarily collected and constructed within a medical context. After fine-tuning, we will conduct a quality and performance evaluation to determine which model

performs better. Additionally, we also deployed several other pre-trained Vietnamese-English bilingual machine translation models, such as VinAI-translate, OPUS-MT, and envit5-base, for experimentation, analysis, and evaluation of results.

### 2.2.4 Medical Image Lesion Classification and Segmentation Module

In this module, our research group developed an integrated pipeline combining two specialized computer vision models for image classification and image segmentation. This specialized pipeline allows users to upload lesion images, perform automatic lesion analysis and evaluation using the system, and finally, view the analysis and evaluation results returned by the system. There will be two main processes within this pipeline: lesion classification and lesion segmentation. Its overall pipeline is illustrated in Figure 7.

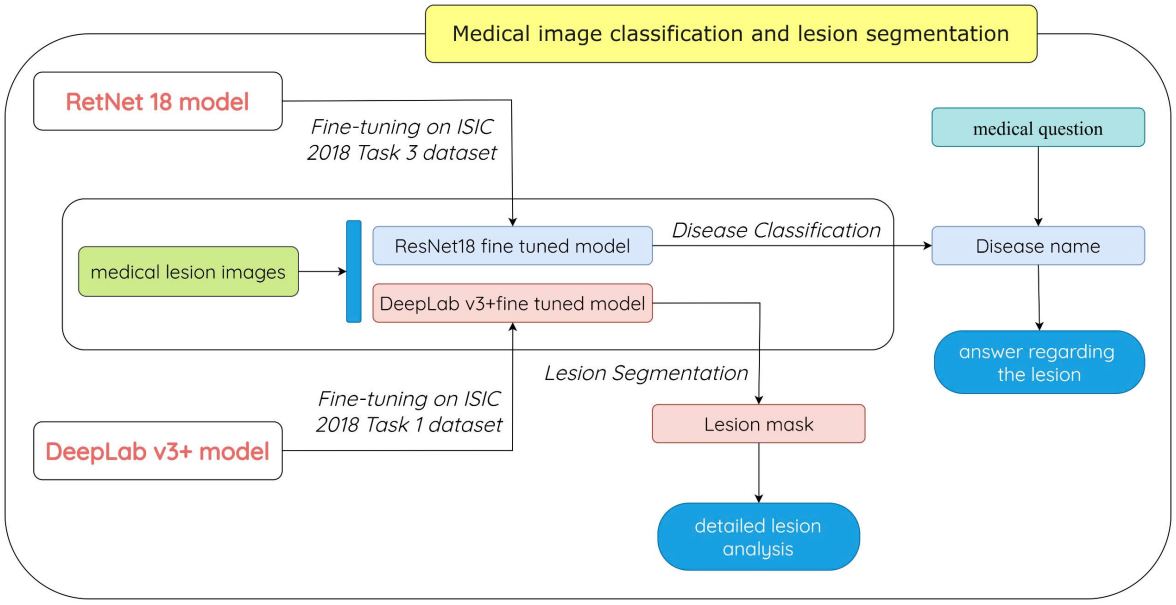


Figure 7: Pipeline of the Medical Image Classification and Lesion Segmentation Module.

For the lesion classification phase, the processing flow is as follows: first, the image will pass through the medical image classification module. Within this module, a specialized image classification model, which we have trained and fine-tuned on an image classification dataset, will be loaded and used to classify the lesion images uploaded by doctors or users. After the model com-

pletes the classification, we will obtain the class label of the image. This module will then map the class label to the name of the lesion or disease and return it to the user.

In parallel with the classification module, the processing flow for the lesion segmentation module is as follows: first, a specialized semantic image segmentation model, which we have trained and fine-tuned on an image segmentation dataset, will be loaded and used to segment lesions in the image that doctors or users have just uploaded. After the model processes and completes the segmentation, we will obtain the segmentation result of the lesion area in that image. This segmentation result is essentially in a binary mask format. Therefore, we leverage this segmentation result and developed an additional submodule to analyze and generate statistical indices for the segmented lesions.

With this module, the group aims to target two groups of users with different objectives. The first group of users we want to target are doctors and medical professionals who can use the platform, specifically this module, to perform clinical evaluations and refer to diagnostic results. The second group of users we want to target are general users, such as patients, who want to use the platform to check if they have any lesions, and if so, what type of lesion it might be and what the lesion area looks like. The processing flow for the first group of users is as follows: assuming doctors are examining a patient with signs of a lesion in a certain part, such as on the skin or throat, after examination, doctors will take images of the lesions and then upload them to the application for analysis. Once doctors upload the image, the system will receive it as input and proceed to analyze this image through two main stages: lesion classification and lesion segmentation. For the second group of users, the processing flow will be simpler: if a user suspects a lesion on their skin, they will upload the lesion image to the system for analysis. After the user uploads the image, the system will proceed with processing similar to that for doctors, going through two analysis processes for lesion classification and segmentation.

For doctors or medical professionals, the application development team hopes that MedStudio can serve as a clinical lesion assessment platform, providing doctors with valuable insights and information, as well as reinforcing their decisions during examination or simply serving as a third

source of information during examination. For patients or individuals with illnesses, the group hopes that the system can help them learn and find information about the lesions they may have.

**Medical Image Classification Module** For this module, we will train and fine-tune the ResNet18 model on the ISIC 2018 dataset. After training and monitoring to determine the optimal number of epochs and parameters for the model to achieve the best classification results, we will save this trained model and then integrate it into the system for image label prediction. It is important to note that when training this model, we need to determine the potential number of class labels for an image to train the model. When using the trained model for inference, it should be considered that there might be data samples that do not fall into the categories of labels used for training, and in such cases, an appropriate handling method is required. Within the scope of this research's experiments, we will use the label with the highest prediction score, and therefore the accuracy will correspond to the accuracy achieved during model training. Additionally, we will also integrate a user query option when uploading images for analysis. In the future, the predicted class label or lesion type name from the classification module can be combined with user queries to generate answers to user questions.

**Medical Image Lesion Segmentation Module** In the group's previous research, experimental evaluation results of models trained on the ISIC 2018 dataset showed that the DeepLab v3+ model achieved significant performance in the task of segmenting lesion areas in medical images and consistently outperformed another very well-known and specialized image classification model, U-Net. The detailed results and analysis are as follows:

Based on the model training process after 22 epochs for both models: DeepLab v3+ and U-Net, the training results of both models were very good. The best-performing model evaluated throughout the training process was DeepLab v3+. The DeepLab v3+ model converged very quickly and showed no signs of instability. However, the U-Net model cannot be considered poor; its results were only slightly inferior to DeepLab v3+ on this dataset. The initial epoch results for the U-Net model were quite poor, with a dice loss value exceeding 1.2, a Dice score of 0, and an IoU score

below 0.2. However, in subsequent epochs, the model learned quickly, achieved good results, and stabilized until the end of the model training process.

Therefore, based on the results of previous research, we determined that DeepLab v3+ has the best and most stable performance. Consequently, in this study, for the ISIC 2018 module, we will train and fine-tune the DeepLabv3+ model on the ISIC 2018 dataset. After training and monitoring to determine the optimal number of epochs and parameters for the model to achieve the best segmentation results, we will save this trained model and then integrate it into the system for predicting image segmentation masks.

### 2.2.5 Medical Visual Question Answering Module

**Visual Question Answering (VQA)** is a task that involves answering questions related to medical images or information implicitly contained within them. What distinguishes VQA from typical question answering is the additional image input; both the questions and answers necessitate an understanding and analysis of the image's meaning and content. Consequently, training a model that can effectively perform VQA is extremely challenging. This is because it requires models to not only comprehend natural language but also to analyze images. This particular module represents one of the most difficult and challenging aspects of our research, as it is a very new task, often explored by large research groups globally. Furthermore, the complexity of the models and architectures for this problem is high due to the need to process both images and text, corresponding to the two distinct tasks of natural language processing and computer vision. When applied to the medical domain, the complexity and difficulty increase significantly, as the inherent complexity of medical images and questions demands specialized data sources and in-depth expertise for accurate answers. Nevertheless, this problem holds immense potential for powerful development, especially when leveraged and applied within the medical field to address medical visual questions and issues.

Fundamentally, the VQA task takes an image and a related question (or a question about information hidden within the image) as input, and outputs an answer derived from the image's content.

Because it requires processing both image and text, the general architecture of models for VQA typically comprises two main blocks:

**Image encoder**: This block is responsible for encoding the image or extracting features from the original user-provided input image.

**Text encoder**: This block is responsible for encoding the text or extracting text features from the original user-provided input question. Additionally, model architectures often include a supplementary block responsible for combining all extracted image and text features (often called Features Fusion or Multimodal Fusion) and generating the final answer returned to the user. Given that the architecture for these tasks involves both image and text analysis, it is quite similar to other multimodal models.

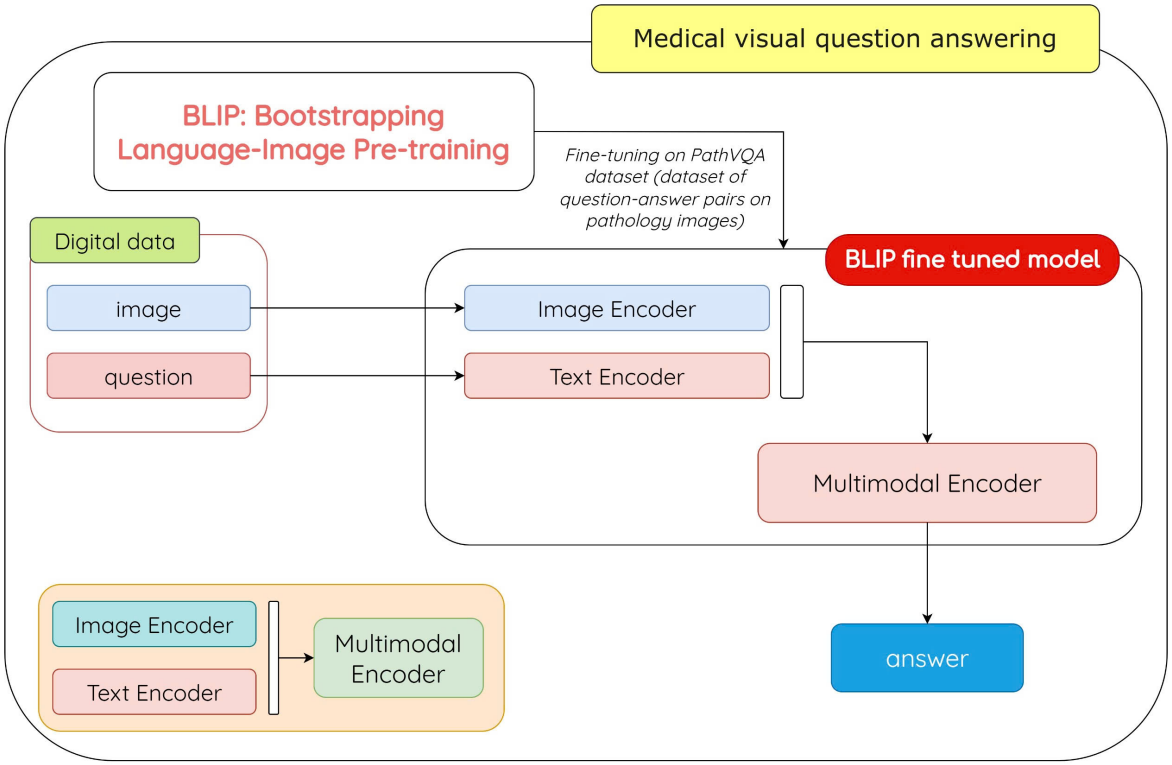


Figure 8: Pipeline of the Medical Visual Question Answering Module.

In this research, for the medical visual question answering module, we will leverage the pre-trained **BLIP: Bootstrapping Language-Image Pre-training** model, which has been pre-trained on datasets covering Visual Question Answering, Image-Text retrieval (Image-text matching), and

Image Captioning. Subsequently, we will train and fine-tune this model on the PATH-VQA dataset (a dataset of question-answer pairs on pathology images). Its overall pipeline is illustrated in Figure 8. During the training and fine-tuning process, we will monitor and determine the optimal number of epochs and hyperparameters for the model to achieve the best results. Once fine-tuning is complete, we will save the fine-tuned model along with its best hyperparameters and integrate it into the system to build the medical visual question answering module.

### 2.2.6 Drug Information Searching and Exploring Based on the VietMeD dataset

This module will be developed from the medical question answering module introduced earlier in this chapter. The unique aspect here is that this module will load and process a separate dataset from the one used by the main QA module. This dataset will be in Vietnamese and will focus exclusively on information about drugs and functional foods collected and built by our team from the Vietnamese Pharmacopoeia.

Our team developed this module with the expectation that it will become an open data repository for drug information in Vietnam. When integrated into our MedStudio platform, users can use it to look up and search for drug information they are interested in. Furthermore, our specialized drug question answering system can answer user queries and problems with clear, specific, and accurate supporting evidence.

Although our ”Vietnamese Pharmaceuticals Dataset” is still small, containing information on only about 200 types of drugs and functional foods, we anticipate that this will serve as a foundational basis. In the future, we will continue to build upon and complete this dataset. At that point, the drug information retrieval and Q&A module based on the Vietnamese Pharmacopoeia will have a richer data source and be able to answer a wider variety of questions about the drugs included in our dataset.

# 3 Experimental Datasets

## 3.1 MedQuAD Dataset

The **MedQuAD** (Medical Question Answering Dataset) is a dataset consisting of 16,413 medical question-answer pairs collected from ten official NIH websites (e.g., cancer.gov, niddk.nih.gov, GARD, MedlinePlus Health Topics). Each pair is associated with specific topics, covering 5,127 unique subjects (e.g., Glaucoma, High Blood Pressure, Paget's Disease of Bone). After preprocessing and cleaning by our team, approximately 16,345 samples remain.

This dataset emphasizes healthcare-related content, featuring various types of questions related to concepts, treatments, diagnoses, side effects of diseases, medications, and other medical entities (e.g., lab tests). Its diversity allows the model to handle a broad range of users, from children to the elderly, and to improve accuracy in treatment and diagnosis-related tasks.

## 3.2 VietMeD Dataset

**VietMeD**, or the Vietnamese Pharmaceuticals Dataset, is a high-quality dataset curated by our team, focusing on medications and dietary supplements. It is manually collected and cleaned from trusted Vietnamese medical information websites, with all data ultimately sourced from the Vietnamese National Pharmacopoeia (2002, 2012, 2017, and 2022 editions).

The dataset is stored in CSV format, with approximately 200 entries corresponding to over 200 pharmaceutical items. Each entry is also available in Markdown format to support applications analyzing structured content like images and tables.

**Attributes of the VietMeD dataset:**

This dataset serves as a centralized and structured pharmaceutical resource, enhancing our retrieval system (MedStudio) for answering drug-related queries.

Table 1: Attributes of the VietMeD Dataset

| Column Name | Description |
|---|---|
| index | Index of the data sample |
| count | Index within the category of the monograph |
| category | Monograph category |
| name | Name of the pharmaceutical item |
| description | Description of the pharmaceutical item |
| origin_source | Original data source |
| source_page | Web page from which data was scraped |
| url | URL to the pharmaceutical product |

## 3.3 ISIC 2018 Dataset

The ISIC 2018 dataset was released as part of the **ISIC Challenge 2018**, which includes three tasks:

- Task 1: Lesion Segmentation
- Task 2: Lesion Attribute Detection
- Task 3: Disease Classification

In this research, we utilize datasets from Task 1 and Task 3.

### 3.3.1 ISIC 2018 Task 3 Dataset

Task 3 focuses on image classification. The dataset consists of approximately 10,000 dermoscopic images labeled with one of seven lesion types using one-hot encoding. A challenge with this dataset is the class imbalance — some classes contain fewer than 200 samples, while others exceed 5,000 samples.

The seven skin lesion categories include:

- Melanoma
- Melanocytic nevus
- Basal cell carcinoma

- Bowen's disease carcinoma
- Benign keratosis
- Dermatofibroma
- Vascular lesion

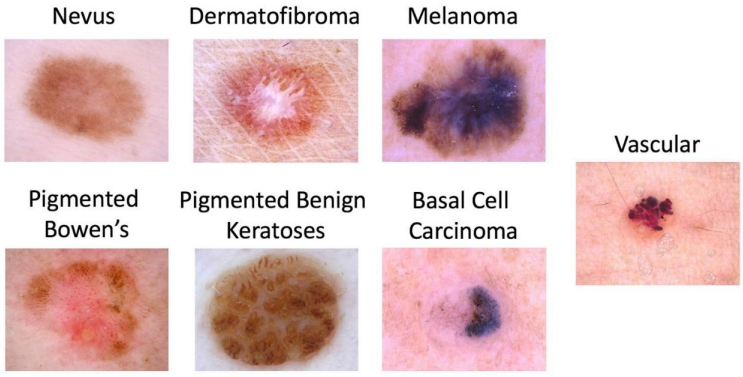


Figure 9: Sample image from ISIC 2018 Task 3 dataset (lesion classification).

### 3.3.2 ISIC 2018 Task 1 Dataset

Task 1 involves binary segmentation of dermoscopic images to distinguish between lesion and non-lesion regions. The original training set includes 2,075 images, which we randomly split into 80% for training and 20% for testing due to the limited public availability of labels.

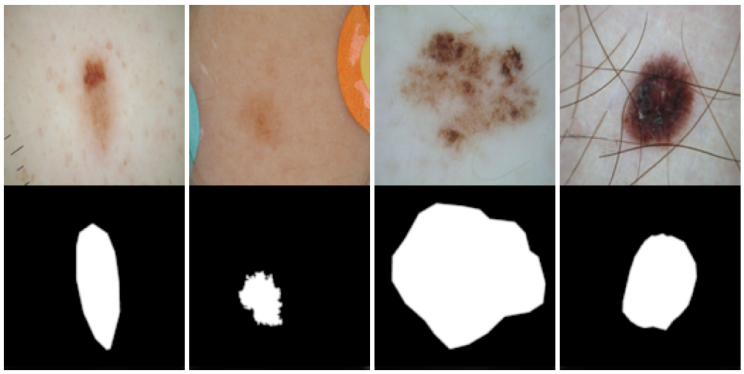

Figure 10: Sample image from ISIC 2018 Task 1 dataset (lesion segmentation).

## 3.4 PathVQA Dataset

The **PathVQA** dataset is a visual question-answering dataset built specifically for the pathology domain. It consists of 5,004 pathology images and 32,795 Q&A pairs, including both open-ended and yes/no questions. The dataset was compiled from two pathology textbooks—"Textbook of Pathology" and "Basic Pathology"—and from the open-access PEIR (Pathology Education Informational Resource) digital library.

**Sample open-ended question:**

- Question: What is present?
- Image: Figure 11
- Answer: gastrointestinal

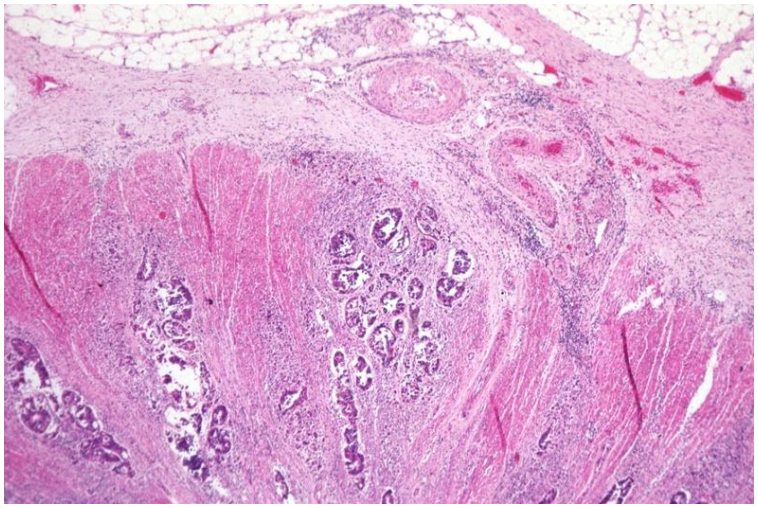

Figure 11: Sample open-ended question from PathVQA.

**Sample yes/no question:**

- Question: Is endoscopic view of a longitudinally-oriented Mallory-Weiss characterized by nests of closely packed glands?
- Image: Figure 12
- Answer: No

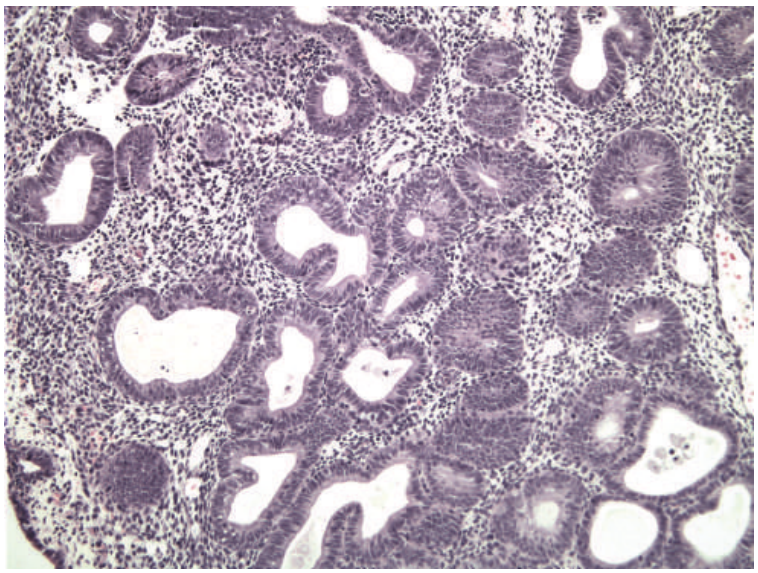

Figure 12: Sample yes/no question from PathVQA.

## 3.5 MedEV Dataset

**MedEV** is a bilingual Vietnamese-English machine translation dataset for the medical domain, developed by the VinAI research team. It aims to address the lack of high-quality bilingual medical corpora, supporting the development of more accurate and domain-specific translation models.

**Dataset Overview:**

- Domain: Medical
- Language Pair: Vietnamese – English
- Size: Approximately 360,000 sentence pairs
- Purpose: To support medical machine translation model training
- Training Set: 340,897 pairs
- Validation Set: 8,982 pairs
- Test Set: 9,006 pairs

# 4 Experimental Results

## 4.1 Medical Question Answering Module

We evaluated two different approaches for our medical question answering system, each combining different retrieval and generation components.

### 4.1.1 Elasticsearch with Llama 3.2

The first approach combines Elasticsearch for document retrieval with Llama 3.2 for answer generation. As shown in Table 2, this configuration achieved strong performance across all metrics.

Table 2: Performance of Elasticsearch + Llama 3.2 configuration

| Metric | Score |
|---|---|
| Context Relevance | 0.778 |
| Answer Relevance | 0.775 |
| Diversity | 0.736 |
| Average Answer Length | 114.4 |
| Average Ground Truth Length | 212.9 |

### 4.1.2 Qdrant with Cohere

Our second approach uses Qdrant for vector-based retrieval and Cohere's language model for generation. Table 3 presents the evaluation results, showing comparable performance to the first approach with slightly better context relevance but lower answer relevance.

Table 3: Performance of Qdrant + Cohere configuration

| Metric | Score |
|---|---|
| Context Relevance | 0.780 |
| Answer Relevance | 0.730 |
| Diversity | 0.729 |
| Average Answer Length | 103.1 |
| Average Ground Truth Length | 212.9 |

## 4.2 Medical Machine Translation Module

We evaluated several machine translation models on medical text translation tasks. Table 4 shows the SacreBLEU scores across different configurations.

Table 4: Machine translation performance (SacreBLEU scores)

| Model | Q&A pairs | SacreBLEU | Steps |
|---|---|---|---|
| VietAI/envit5-translation | 250 | 31.26 | 63 |
| our-team/envit5-translation-fine-tuning/checkpoint-2500 | 250 | 42.67 | 63 |
| our-team/envit5-translation-fine-tuning/checkpoint-7500 | 250 | 49.07 | 63 |
| facebook/mbart-large-50-many-to-many-mmt | 20 | 27.87 | 5 |
| our-team/mbart-large-50-many-to-many-mmt-finetuned-vi-to-en | 20 | 71.41 | 5 |
| our-team/mbart-large-50-many-to-many-mmt-finetuned-vi-to-en | 250 | 71.28 | 63 |

## 4.3 Medical Image Analysis Module

### 4.3.1 Lesion Classification

For lesion classification, we preprocessed the ISIC 2018 dataset as shown in Table 5. The validation results in Table 6 demonstrate consistent performance across different batch sizes.

Table 5: Balanced dataset for lesion classification

| Class | Train | Validation |
|---|---|---|
| MEL (Melanoma) | 1,113 | 223 |
| BKL (Benign keratosis) | 1,099 | 220 |
| NV (Melanocytic nevus) | 1,000 | 200 |
| BCC (Basal cell carcinoma) | 514 | 103 |
| AKIEC (Actinic keratosis) | 327 | 65 |
| VASC (Vascular lesion) | 142 | 28 |
| DF (Dermatofibroma) | 115 | 23 |

### 4.3.2 Lesion Segmentation

The lesion segmentation model achieved excellent performance during training, as shown in Tables 7 and 8.

Table 6: Batch-wise classification accuracy

| Accuracy | Correct/Total |
|---|---|
| 82.00% | 41/50 |
| 78.00% | 390/500 |
| 77.50% | 310/400 |
| 77.89% | 296/380 |

Table 7: Segmentation model performance summary

| Metric | Value |
|---|---|
| Minimum Training Loss | 0.0441 |
| Maximum Training IoU | 0.9515 |
| Maximum Training Dice | 0.8872 |
| Total Training Time | 705 minutes |

Table 8: Detailed training metrics by epoch

| Epoch | Train Loss | Train IoU | Train Dice |
|---|---|---|---|
| 1 | 0.4715 | 0.5861 | 0.8683 |
| 2 | 0.2918 | 0.7104 | 0.8567 |
| 3 | 0.2207 | 0.7737 | 0.8871 |
| 4 | 0.1773 | 0.8147 | 0.8877 |
| 5 | 0.1519 | 0.8396 | 0.8806 |
| 6 | 0.1275 | 0.8635 | 0.8837 |
| 7 | 0.1098 | 0.8814 | 0.8862 |
| 8 | 0.0942 | 0.8968 | 0.8831 |
| 9 | 0.0849 | 0.9070 | 0.8869 |
| 10 | 0.0769 | 0.9158 | 0.8791 |
| 11 | 0.0747 | 0.9187 | 0.8849 |
| 12 | 0.0694 | 0.9240 | 0.8868 |
| 13 | 0.0629 | 0.9307 | 0.8832 |
| 14 | 0.0603 | 0.9339 | 0.8870 |
| 15 | 0.0560 | 0.9382 | 0.8874 |
| 16 | 0.0540 | 0.9409 | 0.8869 |
| 17 | 0.0514 | 0.9433 | 0.8821 |
| 18 | 0.0481 | 0.9468 | 0.8838 |
| 19 | 0.0481 | 0.9472 | 0.8861 |
| 20 | 0.0463 | 0.9491 | 0.8871 |
| 21 | 0.0442 | 0.9514 | 0.8872 |
| 22 | 0.0441 | 0.9515 | 0.8859 |

# 5 Conclusions & Future Work

This research presents the development of **MedStudio**, an intelligent online medical platform that integrates advanced deep learning techniques to address natural language processing (NLP) and computer vision (CV) tasks in the healthcare domain. Our team primarily focused on two key problem areas: first, integrating language models and machine translation to handle medical information and document retrieval, and to answer medical questions. Second, we incorporated modules into MedStudio for medical lesion classification, segmentation of lesion regions in medical images, and answering medical vision questions.

By integrating large language models (LLMs) for understanding and analyzing natural language, along with visual language models for processing multimodal data—specifically text and images—MedStudio provides robust methods for medical information retrieval, question answering, and specialized medical machine translation. A significant contribution of this research is also the collection and construction of the **VietMeD** dataset: a dataset of Vietnamese pharmaceuticals. Through this dataset, our team aims to lay the groundwork for building a comprehensive knowledge base of drug information in Vietnamese.

The proposed multi-module and multimodal architecture of MedStudio, with our team's core focus on accurately addressing medical information retrieval tasks and generating contextually aware answers through a Retrieval-Augmented Generation (RAG) approach, demonstrates a feasible and effective method for delivering accurate and semantically rich medical information.

However, challenges were encountered during the development and practical deployment of MedStudio, particularly concerning limited computational resources and time constraints, which made it difficult to fine-tune models for machine translation and medical vision question-answering tasks. Nevertheless, with its strong capabilities in both natural language processing and computer vision tasks, our team hopes that MedStudio can serve as a foundation for developing into a comprehensive platform that offers accessible online intelligent medical knowledge and solutions for Vietnamese users, and further, to promote the advancement of medical technology.