National College of Ireland

Configuration Manual

MSc Research Project
Master of Science in Artificial Intelligence

Divyasree Harikrishnan
Student ID: 23291508

School of Computing
National College of Ireland

Supervisor: Professor Lavish Thomas

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Divyasree Harikrishnan		
Student ID:	23291508		
Programme:	MSc in Artificial Intelligence	**Year:**	2024-2025
Module:	MSc (Research) Practicum		
Lecturer:	Professor Lavish Thomas		
Submission Due Date:	11-Aug-2025		
Project Title:	Medical Assistant Utilizing Large Language Models with Retrieval Augmented Generation and Vector Search		
Word Count:	716	**Page Count:**	6

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Divyasree Harikrishnan

Date: 11-Aug-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

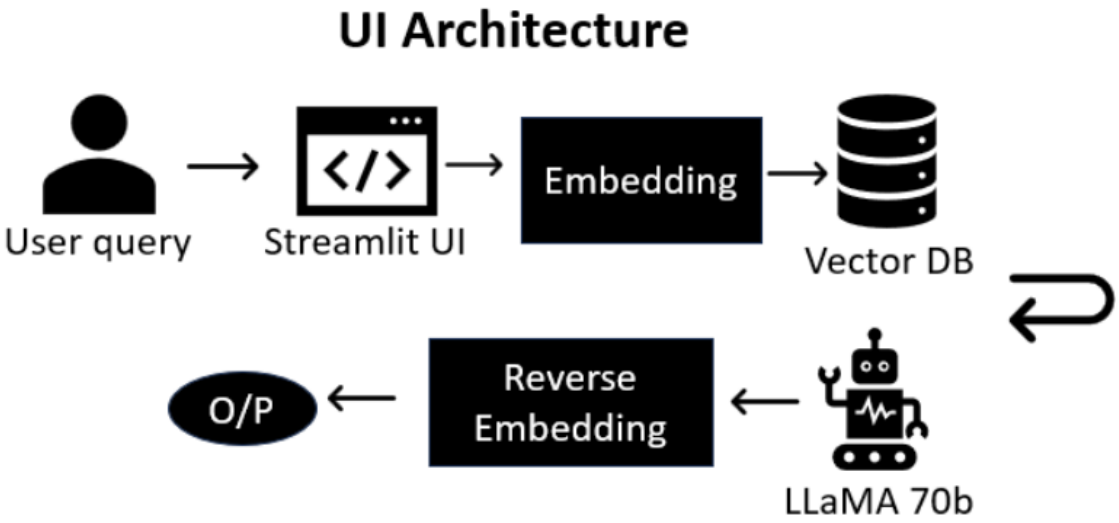
Divyasree Harikrishnan
Student ID: 23291508

1 Introduction

Medical Assistant utilizing LLM with Retrieval Augmented Generation and vector search aims at answering accurately to medical health related queries without any assumptions. The knowledge base is created based on the Gale enclyopedia of medicine (Longe, Blanchfield and Gale Research Company, 2002)

The chatbot involves using Qdrant cloud vector database (qdrant.tech, n.d.) to store the knowledge text which are converted to meaningful biomedical embeddings using PubMedBERT (Huggingface.co, 2025a) and two LLMs were used namely LLaMA 3-70B via Groq API (Groq.com, 2025) and MedGemma 4B-IT via HuggingFace (Huggingface.co, 2025a). these two models are evaluated based on the benchmarks PubmedQA Dataset (Github.io, 2019). and best model is chosen to build the UI

The application has an interactive user interface built in Streamlit, so that it can offer real-time conversation with conversations memory that will provide smooth user experience. Its first-in retrieval strategy prequalifies it as an aid in teaching and in the support of medical research and the health education needs of individuals, and it (conspicuously) does not provide medical diagnoses or individual reparatory recommendations.



2 System Requirements

2.1 Hardware

CPU: Quad-core processor - Intel i5 or equivalent

RAM: Minimum 8 GB required

Storage: 2 GB free space to store the dataset and run the project

Internet connection to access API calls to Hugging Face, Groq, and Qdrant Cloud.

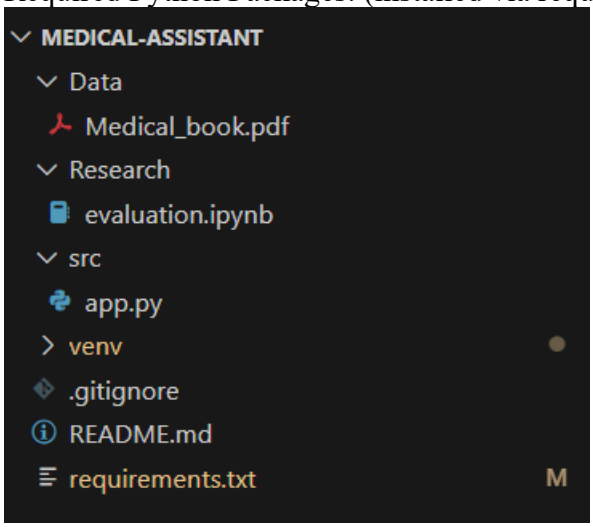
2.2 Software

Operating System: Windows 10 recommended

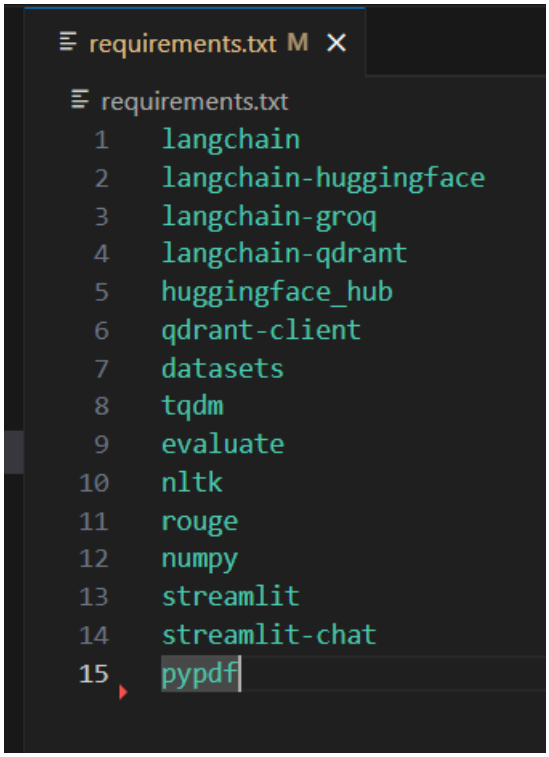
Python: Version 3.9 or later.

Git: Latest version installed.

Required Python Packages: (installed via requirements.txt)



Folder Structure



Requirements.txt

3 Program Execution

Step 1: Clone the repository or download the Zip File

```
C:\Users\dhivs>git clone https://github.com/Divyasree0780/Medical-Assistant.git
Cloning into 'Medical-Assistant'...
remote: Enumerating objects: 20583, done.
remote: Counting objects: 100% (3/3), done.
remote: Compressing objects: 100% (3/3), done.
remote: Total 20583 (delta 0), reused 0 (delta 0), pack-reused 20580 (from 1)
Receiving objects: 100% (20583/20583), 153.51 MiB | 32.17 MiB/s, done.
Resolving deltas: 100% (3221/3221), done.
Updating files: 100% (19593/19593), done.

C:\Users\dhivs>cd Medical-Assistant

C:\Users\dhivs\Medical-Assistant>
```

Step 2: Create and activate a virtual environment

```
C:\Users\dhivs\Medical-Assistant>venv\Scripts\activate
```

Step 3: Install Dependencies

```
(venv) C:\Users\dhivs\Medical-Assistant>pip install -r requirements.txt
Requirement already satisfied: langchain in c:\users\dhivs\documents\ds-
om -r requirements.txt (line 1)) (0.3.27)
```

```
(venv) C:\Users\dhivs\Medical-Assistant>python -m nltk.downloader punkt
<frozen runpy>:128: RuntimeWarning: 'nltk.downloader' found in sys.modules a
execution of 'nltk.downloader'; this may result in unpredictable behaviour
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\dhivs\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping tokenizers\punkt.zip.
```

Step 4: Set Environment Variables

- ➔ Go to Qdrant Cloud Database (qdrant.tech, n.d.) and create your own account and generate your own API Key and get the credentials
- ➔ Go to Hugging Face (Huggingface.co, 2025b) and create an account and get the Hugging Face Token
- ➔ Go to Groq Cloud (Groq.com, 2025) and create an account and get your own API Key

Go to start -> click Environment variables -> add user variables

1. qdrant_url = YOUR_QDRANT_URL
2. qdrant_api_key = YOUR_QDRANT_API_KEY
3. GROQ_API_KEY = YOUR_GROQ_API_KEY
4. HF_TOKEN = YOUR_HUGGINGFACE_TOKEN

Step 5: Run the Application

```
(venv) C:\Users\dhivs\Medical-Assistant>cd src

(venv) C:\Users\dhivs\Medical-Assistant\src>streamlit run app.py

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://192.168.0.238:8501
```

References

Longe, J.L., Blanchfield, D.S. and Gale Research Company (2002). *Gale encyclopedia of medicine*. Detroit, Mi: Gale Group.

Huggingface.co. (2025b). *NeuML/pubmedbert-base-embeddings · Hugging Face*. [online] Available at: https://huggingface.co/NeuML/pubmedbert-base-embeddings.

Groq.com. (2025). *GroqDocs - Build Fast*. [online] Available at: https://console.groq.com/docs/model/llama3-70b-8192.

Huggingface.co. (2025a). *google/medgemma-4b-it · Hugging Face*. [online] Available at: https://huggingface.co/google/medgemma-4b-it.

Github.io. (2019). *PubMedQA Homepage*. [online] Available at: https://pubmedqa.github.io/.

qdrant.tech. (n.d.). *Qdrant - Vector Database*. [online] Available at: https://qdrant.tech/.