Problem Statement

Business Context

The healthcare industry is rapidly evolving, with professionals facing increasing challenges in managing vast volumes of medical data while delivering accurate and timely diagnoses. The need for quick access to comprehensive, reliable, and up-to-date medical knowledge is critical for improving patient outcomes and ensuring informed decision-making in a fast-paced environment.

Healthcare professionals often encounter information overload, struggling to sift through extensive research and data to create accurate diagnoses and treatment plans. This challenge is amplified by the need for efficiency, particularly in emergencies, where time-sensitive decisions are vital. Furthermore, access to trusted, current medical information from renowned manuals and research papers is essential for maintaining high standards of care.

To address these challenges, healthcare centers can focus on integrating systems that streamline access to medical knowledge, provide tools to support quick decision-making, and enhance efficiency. Leveraging centralized knowledge platforms and ensuring healthcare providers have continuous access to reliable resources can significantly improve patient care and operational effectiveness.

Common Questions to Answer

1. Diagnostic Assistance: "What are the common symptoms and treatments for pulmonary embolism?"

2. Drug Information: "Can you provide the trade names of medications used for treating hypertension?"

3. Treatment Plans: "What are the first-line options and alternatives for managing rheumatoid arthritis?"

4. Specialty Knowledge: "What are the diagnostic steps for suspected endocrine disorders?"

5. Critical Care Protocols: "What is the protocol for managing sepsis in a critical care unit?"

Objective

As an AI specialist, your task is to develop a RAG-based AI solution using renowned medical manuals to address healthcare challenges. The objective is to understand issues like information overload, apply AI techniques to streamline decision-making, analyze its impact on diagnostics and patient outcomes, evaluate its potential to standardize care practices, and create a functional prototype demonstrating its feasibility and effectiveness.

Data Description

The Merck Manuals are medical references published by the American pharmaceutical company Merck & Co., that cover a wide range of medical topics, including disorders, tests, diagnoses, and drugs. The manuals have been published since 1899, when Merck & Co. was still a subsidiary of the German company Merck.

The manual is provided as a PDF with over 4,000 pages divided into 23 sections.

Installing and Importing Necessary Libraries and Dependencies

In [1]:

```
# Installation for GPU llama-cpp-python
!CMAKE_ARGS="-DLLAMA_CUBLAS=on" FORCE_CMAKE=1 pip install llama-cpp-python==0.2.45 --for
ce-reinstall --no-cache-dir -q
```

```
                                        36.7/36.7 MB 195.6 MB/s eta 0:00:00
```

```
  Installing build dependencies ... done
  Getting requirements to build wheel ... done
  Installing backend dependencies ... done
  Preparing metadata (pyproject.toml) ... done
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 62.1/62.1 kB 157.4 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 45.5/45.5 kB 282.5 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 134.9/134.9 kB 394.9 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 16.9/16.9 MB 289.0 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 43.9/43.9 kB 181.2 MB/s eta 0:00:00
  Building wheel for llama-cpp-python (pyproject.toml) ... done
ERROR: pip's dependency resolver does not currently take into account all the packages th
at are installed. This behaviour is the source of the following dependency conflicts.
torch 2.6.0+cu124 requires nvidia-cublas-cu12==12.4.5.8; platform_system == "Linux" and p
latform_machine == "x86_64", but you have nvidia-cublas-cu12 12.5.3.2 which is incompatib
le.
torch 2.6.0+cu124 requires nvidia-cuda-cupti-cu12==12.4.127; platform_system == "Linux" a
nd platform_machine == "x86_64", but you have nvidia-cuda-cupti-cu12 12.5.82 which is inc
ompatible.
torch 2.6.0+cu124 requires nvidia-cuda-nvrtc-cu12==12.4.127; platform_system == "Linux" a
nd platform_machine == "x86_64", but you have nvidia-cuda-nvrtc-cu12 12.5.82 which is inc
ompatible.
torch 2.6.0+cu124 requires nvidia-cuda-runtime-cu12==12.4.127; platform_system == "Linux"
and platform_machine == "x86_64", but you have nvidia-cuda-runtime-cu12 12.5.82 which is
incompatible.
torch 2.6.0+cu124 requires nvidia-cudnn-cu12==9.1.0.70; platform_system == "Linux" and pl
atform_machine == "x86_64", but you have nvidia-cudnn-cu12 9.3.0.75 which is incompatible
.
torch 2.6.0+cu124 requires nvidia-cufft-cu12==11.2.1.3; platform_system == "Linux" and pl
atform_machine == "x86_64", but you have nvidia-cufft-cu12 11.2.3.61 which is incompatibl
e.
torch 2.6.0+cu124 requires nvidia-curand-cu12==10.3.5.147; platform_system == "Linux" and
platform_machine == "x86_64", but you have nvidia-curand-cu12 10.3.6.82 which is incompat
ible.
torch 2.6.0+cu124 requires nvidia-cusolver-cu12==11.6.1.9; platform_system == "Linux" and
platform_machine == "x86_64", but you have nvidia-cusolver-cu12 11.6.3.83 which is incomp
atible.
torch 2.6.0+cu124 requires nvidia-cusparse-cu12==12.3.1.170; platform_system == "Linux" a
nd platform_machine == "x86_64", but you have nvidia-cusparse-cu12 12.5.1.3 which is inco
mpatible.
torch 2.6.0+cu124 requires nvidia-nccl-cu12==2.21.5; platform_system == "Linux" and platf
orm_machine == "x86_64", but you have nvidia-nccl-cu12 2.23.4 which is incompatible.
torch 2.6.0+cu124 requires nvidia-nvjitlink-cu12==12.4.127; platform_system == "Linux" an
d platform_machine == "x86_64", but you have nvidia-nvjitlink-cu12 12.5.82 which is incom
patible.
opencv-contrib-python 4.12.0.88 requires numpy<2.3.0,>=2; python_version >= "3.9", but yo
u have numpy 2.3.2 which is incompatible.
numba 0.60.0 requires numpy<2.1,>=1.22, but you have numpy 2.3.2 which is incompatible.
opencv-python-headless 4.12.0.88 requires numpy<2.3.0,>=2; python_version >= "3.9", but y
ou have numpy 2.3.2 which is incompatible.
opencv-python 4.12.0.88 requires numpy<2.3.0,>=2; python_version >= "3.9", but you have n
umpy 2.3.2 which is incompatible.
tensorflow 2.19.0 requires numpy<2.2.0,>=1.26.0, but you have numpy 2.3.2 which is incomp
atible.
cupy-cuda12x 13.3.0 requires numpy<2.3,>=1.22, but you have numpy 2.3.2 which is incompat
ible.
```

Note:

- **After running the above cell, kindly restart the runtime (for Google Colab) or notebook kernel (for Jupyter Notebook), and run all cells sequentially from the next cell.**
- **On executing the above line of code, you might see a warning regarding package dependencies. This error message can be ignored as the above code ensures that all necessary libraries and their dependencies are maintained to successfully execute the code in *this notebook*.**

In [1]:

```
# For installing the libraries & downloading models from HF Hub
!pip install huggingface_hub==0.23.2 pandas==1.5.3 tiktoken==0.6.0 pymupdf==1.25.1 langc
hain==0.1.1 langchain-community==0.0.13 chromadb==0.4.22 sentence-transformers==2.3.1 num
py==1.25.2 -q
```

```
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 67.3/67.3 kB 4.0 MB/s eta 0:00:00
  Installing build dependencies ... done
  Getting requirements to build wheel ... done
  Preparing metadata (pyproject.toml) ... done
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 41.7/41.7 kB 3.7 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 41.7/41.7 kB 3.6 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 40.9/40.9 kB 3.6 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 40.9/40.9 kB 3.4 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 40.9/40.9 kB 3.7 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 40.2/40.2 kB 3.7 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 40.2/40.2 kB 3.4 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 44.0/44.0 kB 4.0 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 44.4/44.4 kB 3.8 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 44.4/44.4 kB 3.7 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 44.4/44.4 kB 3.9 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 44.4/44.4 kB 3.7 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 44.1/44.1 kB 3.4 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 43.5/43.5 kB 2.8 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 44.1/44.1 kB 3.9 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 401.7/401.7 kB 19.5 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 12.0/12.0 MB 86.5 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.8/1.8 MB 66.8 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 20.0/20.0 MB 77.7 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 802.4/802.4 kB 44.3 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.6/1.6 MB 77.1 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 509.0/509.0 kB 39.8 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 132.8/132.8 kB 13.8 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 18.2/18.2 MB 117.0 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.4/2.4 MB 95.8 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 284.2/284.2 kB 26.9 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.9/1.9 MB 96.0 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 241.2/241.2 kB 23.9 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 55.4/55.4 kB 5.4 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 103.1/103.1 kB 10.1 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 16.5/16.5 MB 117.0 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 65.6/65.6 kB 6.1 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 72.5/72.5 kB 7.8 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 201.6/201.6 kB 19.6 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 120.0/120.0 kB 12.6 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 53.0/53.0 kB 5.0 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 116.7/116.7 kB 12.2 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 6.6/6.6 MB 131.9 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 363.4/363.4 MB 4.1 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.8/13.8 MB 122.5 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 24.6/24.6 MB 93.8 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 883.7/883.7 kB 59.5 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 664.8/664.8 MB 2.7 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 211.5/211.5 MB 5.3 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 56.3/56.3 MB 14.5 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 127.9/127.9 MB 7.4 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 207.5/207.5 MB 5.2 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 188.7/188.7 MB 6.0 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 21.1/21.1 MB 98.0 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 10.0/10.0 MB 127.7 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 3.0/3.0 MB 91.8 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 459.8/459.8 kB 39.8 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 50.9/50.9 kB 5.2 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 4.0/4.0 MB 70.8 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 453.1/453.1 kB 40.0 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 46.0/46.0 kB 4.1 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 86.8/86.8 kB 9.7 MB/s eta 0:00:00
  Building wheel for pypika (pyproject.toml) ... done
ERROR: pip's dependency resolver does not currently take into account all the packages th
at are installed. This behaviour is the source of the following dependency conflicts.
google-colab 1.0.0 requires pandas==2.2.2, but you have pandas 1.5.3 which is incompatibl
e.
xarray 2025.7.1 requires numpy>=1.26, but you have numpy 1.25.2 which is incompatible.
xarray 2025.7.1 requires packaging>=24.1, but you have packaging 23.2 which is incompatib
le.
xarray 2025.7.1 requires pandas>=2.2, but you have pandas 1.5.3 which is incompatible.
datasets 4.0.0 requires huggingface-hub>=0.24.0, but you have huggingface-hub 0.23.2 whic
```

```
n is incompatible.
arviz 0.22.0 requires numpy>=1.26.0, but you have numpy 1.25.2 which is incompatible.
arviz 0.22.0 requires pandas>=2.1.0, but you have pandas 1.5.3 which is incompatible.
opencv-contrib-python 4.12.0.88 requires numpy<2.3.0,>=2; python_version >= "3.9", but yo
u have numpy 1.25.2 which is incompatible.
dask-cudf-cu12 25.6.0 requires pandas<2.2.4dev0,>=2.0, but you have pandas 1.5.3 which is
incompatible.
plotnine 0.14.5 requires pandas>=2.2.0, but you have pandas 1.5.3 which is incompatible.
mizani 0.13.5 requires pandas>=2.2.0, but you have pandas 1.5.3 which is incompatible.
blosc2 3.6.1 requires numpy>=1.26, but you have numpy 1.25.2 which is incompatible.
opencv-python-headless 4.12.0.88 requires numpy<2.3.0,>=2; python_version >= "3.9", but y
ou have numpy 1.25.2 which is incompatible.
opencv-python 4.12.0.88 requires numpy<2.3.0,>=2; python_version >= "3.9", but you have n
umpy 1.25.2 which is incompatible.
cudf-cu12 25.6.0 requires pandas<2.2.4dev0,>=2.0, but you have pandas 1.5.3 which is inco
mpatible.
langchain-text-splitters 0.3.9 requires langchain-core<1.0.0,>=0.3.72, but you have langc
hain-core 0.1.23 which is incompatible.
geopandas 1.1.1 requires pandas>=2.0.0, but you have pandas 1.5.3 which is incompatible.
thinc 8.3.6 requires numpy<3.0.0,>=2.0.0, but you have numpy 1.25.2 which is incompatible
.
peft 0.17.0 requires huggingface_hub>=0.25.0, but you have huggingface-hub 0.23.2 which i
s incompatible.
google-cloud-bigquery 3.35.1 requires packaging>=24.2.0, but you have packaging 23.2 whic
h is incompatible.
gradio 5.41.0 requires huggingface-hub<1.0,>=0.33.5, but you have huggingface-hub 0.23.2
which is incompatible.
tensorflow 2.19.0 requires numpy<2.2.0,>=1.26.0, but you have numpy 1.25.2 which is incom
patible.
diffusers 0.34.0 requires huggingface-hub>=0.27.0, but you have huggingface-hub 0.23.2 wh
ich is incompatible.
db-dtypes 1.4.3 requires packaging>=24.2.0, but you have packaging 23.2 which is incompat
ible.
```

Note:

- **After running the above cell, kindly restart the runtime (for Google Colab) or notebook kernel (for Jupyter Notebook), and run all cells sequentially from the next cell.**
- **On executing the above line of code, you might see a warning regarding package dependencies. This error message can be ignored as the above code ensures that all necessary libraries and their dependencies are maintained to successfully execute the code in *this notebook*.**

In [1]:

```
!pip install sentence-transformers==2.3.1 transformers==4.32.0 huggingface-hub==0.27.1 -
-quiet
```

```
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 118.5/118.5 kB 5.0 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 7.5/7.5 MB 59.5 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 450.7/450.7 kB 24.8 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 7.8/7.8 MB 59.5 MB/s eta 0:00:00
ERROR: pip's dependency resolver does not currently take into account all the packages th
at are installed. This behaviour is the source of the following dependency conflicts.
gradio 5.41.0 requires huggingface-hub<1.0,>=0.33.5, but you have huggingface-hub 0.27.1
which is incompatible.
```

In [1]:

```
#Libraries for processing dataframes,text
import json,os
import tiktoken
import pandas as pd

#Libraries for Loading Data, Chunking, Embedding, and Vector Databases
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain_community.document_loaders import PyMuPDFLoader,PyPDFDirectoryLoader, PyPD
FLoader
from langchain_community.embeddings.sentence_transformer import SentenceTransformerEmbedd
ings
from langchain_community.vectorstores import Chroma
```

```
#Libraries for downloading and loading the llm
from huggingface_hub import hf_hub_download
from llama_cpp import Llama

from google.colab import userdata, drive
```

In [2]:

```
#from langchain_community.document_loaders import PyPDFDirectoryLoader, PyPDFLoader
#from langchain_community.embeddings.sentence_transformer import (
 #   SentenceTransformerEmbeddings
#)

#from google.colab import userdata, drive
```

Question Answering using LLM

Downloading and Loading the model

In [3]:

```
# For downloading the models from HF Hub
!pip install huggingface_hub==0.28.1 -q
```

```
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 464.1/464.1 kB 12.3 MB/s eta 0:00:00
ERROR: pip's dependency resolver does not currently take into account all the packages th
at are installed. This behaviour is the source of the following dependency conflicts.
gradio 5.41.0 requires huggingface-hub<1.0,>=0.33.5, but you have huggingface-hub 0.28.1
which is incompatible.
```

In [4]:

```
## Model configuration
model_name_or_path = "TheBloke/Llama-2-13B-chat-GGUF"
model_basename = "llama-2-13b-chat.Q5_K_M.gguf"
model_path = hf_hub_download(
    repo_id=model_name_or_path,
    filename=model_basename
    )
```

```
/usr/local/lib/python3.11/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://h
uggingface.co/settings/tokens), set it as secret in your Google Colab and restart your se
ssion.
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models
or datasets.
  warnings.warn(
```

In [5]:

```
llm = Llama(
    model_path=model_path,
    n_ctx=4096,
    n_gpu_layers=38,
    n_batch=512
)
```

```
llama_model_loader: loaded meta data with 19 key-value pairs and 363 tensors from /root/.
cache/huggingface/hub/models--TheBloke--Llama-2-13B-chat-GGUF/snapshots/4458acc949de0a991
4c3eab623904d4fe999050a/llama-2-13b-chat.Q5_K_M.gguf (version GGUF V2)
llama_model_loader: Dumping metadata keys/values. Note: KV overrides do not apply in this
output.
llama_model_loader: - kv   0:                       general.architecture str
= llama
llama_model_loader: - kv   1:                               general.name str
```

```
= LLaMA v2
llama_model_loader: - kv   2:                       llama.context_length u32
= 4096
llama_model_loader: - kv   3:                     llama.embedding_length u32
= 5120
llama_model_loader: - kv   4:                          llama.block_count u32
= 40
llama_model_loader: - kv   5:                  llama.feed_forward_length u32
= 13824
llama_model_loader: - kv   6:                 llama.rope.dimension_count u32
= 128
llama_model_loader: - kv   7:                 llama.attention.head_count u32
= 40
llama_model_loader: - kv   8:              llama.attention.head_count_kv u32
= 40
llama_model_loader: - kv   9:     llama.attention.layer_norm_rms_epsilon f32
= 0.000010
llama_model_loader: - kv  10:                          general.file_type u32
= 17
llama_model_loader: - kv  11:                       tokenizer.ggml.model str
= llama
llama_model_loader: - kv  12:                      tokenizer.ggml.tokens arr[str,32000]
= ["<unk>", "<s>", "</s>", "<0x00>", "<...
llama_model_loader: - kv  13:                      tokenizer.ggml.scores arr[f32,32000]
= [0.000000, 0.000000, 0.000000, 0.0000...
llama_model_loader: - kv  14:                  tokenizer.ggml.token_type arr[i32,32000]
= [2, 3, 3, 6, 6, 6, 6, 6, 6, 6, 6, 6, ...
llama_model_loader: - kv  15:                tokenizer.ggml.bos_token_id u32
= 1
llama_model_loader: - kv  16:                tokenizer.ggml.eos_token_id u32
= 2
llama_model_loader: - kv  17:            tokenizer.ggml.unknown_token_id u32
= 0
llama_model_loader: - kv  18:               general.quantization_version u32
= 2
llama_model_loader: - type  f32:   81 tensors
llama_model_loader: - type q5_K:  241 tensors
llama_model_loader: - type q6_K:   41 tensors
llm_load_vocab: special tokens definition check successful ( 259/32000 ).
llm_load_print_meta: format           = GGUF V2
llm_load_print_meta: arch             = llama
llm_load_print_meta: vocab type       = SPM
llm_load_print_meta: n_vocab          = 32000
llm_load_print_meta: n_merges         = 0
llm_load_print_meta: n_ctx_train      = 4096
llm_load_print_meta: n_embd           = 5120
llm_load_print_meta: n_head           = 40
llm_load_print_meta: n_head_kv        = 40
llm_load_print_meta: n_layer          = 40
llm_load_print_meta: n_rot            = 128
llm_load_print_meta: n_embd_head_k    = 128
llm_load_print_meta: n_embd_head_v    = 128
llm_load_print_meta: n_gqa            = 1
llm_load_print_meta: n_embd_k_gqa     = 5120
llm_load_print_meta: n_embd_v_gqa     = 5120
llm_load_print_meta: f_norm_eps       = 0.0e+00
llm_load_print_meta: f_norm_rms_eps   = 1.0e-05
llm_load_print_meta: f_clamp_kqv      = 0.0e+00
llm_load_print_meta: f_max_alibi_bias = 0.0e+00
llm_load_print_meta: n_ff             = 13824
llm_load_print_meta: n_expert         = 0
llm_load_print_meta: n_expert_used    = 0
llm_load_print_meta: rope scaling     = linear
llm_load_print_meta: freq_base_train  = 10000.0
llm_load_print_meta: freq_scale_train = 1
llm_load_print_meta: n_yarn_orig_ctx  = 4096
llm_load_print_meta: rope_finetuned   = unknown
llm_load_print_meta: model type       = 13B
llm_load_print_meta: model ftype      = Q5_K - Medium
llm_load_print_meta: model params     = 13.02 B
llm_load_print_meta: model size       = 8.60 GiB (5.67 BPW)
llm_load_print_meta: general.name     = LLaMA v2
llm_load_print_meta: BOS token        = 1 '<s>'
```

```
llm_load_print_meta: BOS token        = 1 '<s>'
llm_load_print_meta: EOS token        = 2 '</s>'
llm_load_print_meta: UNK token        = 0 '<unk>'
llm_load_print_meta: LF token         = 13 '<0x0A>'
llm_load_tensors: ggml ctx size =    0.28 MiB
llm_load_tensors: offloading 38 repeating layers to GPU
llm_load_tensors: offloaded 38/41 layers to GPU
llm_load_tensors:        CPU buffer size =  8801.63 MiB
llm_load_tensors:      CUDA0 buffer size =  8125.43 MiB
.................................................................................
...........
llama_new_context_with_model: n_ctx      = 4096
llama_new_context_with_model: freq_base  = 10000.0
llama_new_context_with_model: freq_scale = 1
llama_kv_cache_init:  CUDA_Host KV buffer size =   160.00 MiB
llama_kv_cache_init:      CUDA0 KV buffer size =  3040.00 MiB
llama_new_context_with_model: KV self size  = 3200.00 MiB, K (f16): 1600.00 MiB, V (f16):
1600.00 MiB
llama_new_context_with_model:  CUDA_Host input buffer size   =    19.04 MiB
llama_new_context_with_model:      CUDA0 compute buffer size =   368.02 MiB
llama_new_context_with_model:  CUDA_Host compute buffer size =   370.00 MiB
llama_new_context_with_model: graph splits (measure): 5
AVX = 1 | AVX_VNNI = 0 | AVX2 = 1 | AVX512 = 1 | AVX512_VBMI = 0 | AVX512_VNNI = 0 | FMA
= 1 | NEON = 0 | ARM_FMA = 0 | F16C = 1 | FP16_VA = 0 | WASM_SIMD = 0 | BLAS = 1 | SSE3
= 1 | SSSE3 = 1 | VSX = 0 | MATMUL_INT8 = 0 |
Model metadata: {'tokenizer.ggml.unknown_token_id': '0', 'tokenizer.ggml.eos_token_id': '
2', 'general.architecture': 'llama', 'llama.context_length': '4096', 'general.name': 'LLa
MA v2', 'llama.embedding_length': '5120', 'llama.feed_forward_length': '13824', 'llama.at
tention.layer_norm_rms_epsilon': '0.000010', 'llama.rope.dimension_count': '128', 'llama.
attention.head_count': '40', 'tokenizer.ggml.bos_token_id': '1', 'llama.block_count': '40
', 'llama.attention.head_count_kv': '40', 'general.quantization_version': '2', 'tokenizer
.ggml.model': 'llama', 'general.file_type': '17'}
```

Response

In [6]:

```
# function to generate, process, and return the response from the LLM
def generate_llama_response(user_prompt):

    # System message
    system_message = """
    [INST]<<SYS>> Respond to the user question based on the user prompt<</SYS>>[/INST]
    """

    # Combine user_prompt and system_message to create the prompt
    prompt = f"{user_prompt}\n{system_message}"

    # Generate a response from the LLaMA model
    response = llm(
        prompt=prompt,
        max_tokens=1024,
        temperature=0.01,
        top_p=0.95,
        repeat_penalty=1.2,
        top_k=50,
        stop=['INST'],
        echo=False
    )

    # Extract and return the response text
    response_text = response["choices"][0]["text"]
    return response_text
```

Query 1: What is the protocol for managing sepsis in a critical care unit?

In [7]:

```
user_prompt = "What is the protocol for managing sepsis in a critical care unit?"
```

```
response = generate_llama_response(user_prompt)
print(response)
```

```
llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =     473.29 ms /   860 runs   (    0.55 ms per tok
en,  1817.06 tokens per second)
llama_print_timings: prompt eval time =     860.38 ms /    49 tokens (   17.56 ms per tok
en,    56.95 tokens per second)
llama_print_timings:        eval time =  118669.98 ms /   859 runs   (  138.15 ms per tok
en,     7.24 tokens per second)
llama_print_timings:       total time =  123463.49 ms /   908 tokens
```

```
 The management of sepsis in a critical care unit (CCU) follows a standardized protocol t
hat includes early recognition, aggressive fluid resuscitation, antibiotic therapy, and c
lose monitoring. Here are the key components of the sepsis protocol in a CCU:

1. Early recognition: The CCU team, including intensivists, nurses, and respiratory thera
pists, should be vigilant in identifying patients at risk for sepsis. This includes monit
oring vital signs, particularly temperature, blood pressure, and tissue perfusion, as wel
l as assessing for signs of infection, such as fever, tachycardia, and tachypnea.
2. Rapid fluid resuscitation: Sepsis patients may present with hypovolemia or hypotension
, which can worsen organ dysfunction and mortality. CCU teams should initiate fluid resus
citation with crystalloids or colloids as soon as possible, aiming to achieve a mean arte
rial pressure (MAP) of 65-70 mmHg.
3. Broad-spectrum antibiotics: Empiric antibiotic therapy should be initiated immediately
, covering both gram-positive and gram-negative pathogens. The choice of antibiotics shou
ld be guided by local antibiotic sensitivity patterns and the patient's medical history a
nd allergies.
4. Serial vital sign monitoring: Patients with sepsis should be closely monitored, with s
erial assessments of vital signs, including MAP, heart rate, respiratory rate, and oxygen
saturation. This helps identify early signs of fluid overload or cardiovascular instabili
ty.
5. Cardiac monitoring: Sepsis can cause cardiac dysfunction, so cardiac monitoring, inclu
ding electrocardiography (ECG), should be performed regularly. This helps identify any ca
rdiac abnormalities, such as arrhythmias or cardiac failure, and guides therapy.
6. Respiratory support: Patients with severe respiratory dysfunction may require mechanic
al ventilation or non-invasive ventilation. The choice of respiratory support should be g
uided by the patient's clinical status, respiratory function, and comorbidities.
7. Close monitoring of organ dysfunction: Sepsis can cause multi-organ dysfunction, so it
's essential to monitor organ function closely, including renal, hepatic, and cardiac fun
ction. This helps identify early signs of organ failure and guides therapy.
8. Goal-directed therapy: CCU teams should aim to achieve specific goals, such as maintai
ning MAP ≥ 65 mmHg, serum lactate levels ≤ 2 mmol/L, and urine output ≥ 0.5 mL/kg/h. Thes
e goals should be adjusted according to the patient's clinical status and response to the
rapy.
9. Protocol-driven treatment: A sepsis protocol should be followed, which includes specif
ic interventions and monitoring parameters. This helps ensure consistency and quality of
care, while also reducing variability and errors.
10. Multidisciplinary team approach: Sepsis management requires a multidisciplinary team
approach, involving intensivists, nurses, respiratory therapists, and other healthcare pr
ofessionals. This ensures that patients receive comprehensive and coordinated care.
11. Family involvement: Family members should be involved in the patient's care, providin
g emotional support and helping with decision-making. This can improve patient outcomes a
nd reduce anxiety and stress.
12. Continuous quality improvement: CCUs should continuously monitor and evaluate their s
epsis management practices, using quality indicators and benchmarking data. This helps id
entify areas for improvement and ensures that best practices are being followed.
```

Query 2: What are the common symptoms for appendicitis, and can it be cured via medicine? If not, what surgical procedure should be followed to treat it?

In [8]:

```
user_prompt = "What are the common symptoms for appendicitis, and can it be cured via med
icine? If not, what surgical procedure should be followed to treat it?"
response = generate_llama_response(user_prompt)
print(response)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
```

```
llama_print_timings:      sample time =     269.19 ms /   472 runs   (    0.57 ms per tok
en,  1753.40 tokens per second)
llama_print_timings: prompt eval time =     553.58 ms /    65 tokens (    8.52 ms per tok
en,   117.42 tokens per second)
llama_print_timings:        eval time =   64173.76 ms /   471 runs   (  136.25 ms per tok
en,     7.34 tokens per second)
llama_print_timings:       total time =   66629.45 ms /   536 tokens
```

```
 Sure, I'd be happy to help! Appendicitis is a medical emergency that requires prompt tre
atment. Here are some common symptoms of appendicitis:

1. Severe pain in the abdomen, usually starting near the belly button and then moving to
the lower right side of the abdomen. The pain may be constant or come and go.
2. Nausea and vomiting.
3. Loss of appetite.
4. Fever.
5. Abdominal tenderness and guarding (muscle tension) in the affected area.
6. Abdominal swelling.
7. Inability to pass gas.
8. Inability to move the bowels.

If you suspect that you or someone else may have appendicitis, it is important to seek me
dical attention immediately. Appendicitis cannot be cured with medication alone; instead,
surgery is required to remove the inflamed appendix. The most common surgical procedure u
sed to treat appendicitis is an appendectomy.

There are two types of appendectomies: open and laparoscopic. An open appendectomy involv
es a larger incision in the abdomen to remove the inflamed appendix. A laparoscopic appen
dectomy is a less invasive procedure that uses a small camera and specialized instruments
to remove the appendix through several small incisions. Both procedures have their own ad
vantages and disadvantages, and your doctor will determine which one is best for you base
d on your individual needs.

In addition to these surgical procedures, your doctor may also prescribe antibiotics to h
elp prevent infection. It is important to follow your doctor's instructions carefully aft
er an appendectomy to ensure a smooth recovery. This may include taking pain medication a
s directed, resting and avoiding strenuous activities, and following a healthy diet.

In rare cases, if the appendix is not removed in time, it may rupture and cause potential
ly life-threatening complications. Therefore, it is crucial to seek medical attention imm
ediately if you suspect that you or someone else may have appendicitis.
```

Query 3: What are the effective treatments or solutions for addressing sudden patchy hair loss, commonly seen as localized bald spots on the scalp, and what could be the possible causes behind it?

In [9]:

```
user_prompt = "What are the effective treatments or solutions for addressing sudden patch
y hair loss, commonly seen as localized bald spots on the scalp, and what could be the po
ssible causes behind it?"
response = generate_llama_response(user_prompt)
print(response)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =     586.66 ms /   975 runs   (    0.60 ms per tok
en,  1661.95 tokens per second)
llama_print_timings: prompt eval time =     574.70 ms /    69 tokens (    8.33 ms per tok
en,   120.06 tokens per second)
llama_print_timings:        eval time =  135806.94 ms /   974 runs   (  139.43 ms per tok
en,     7.17 tokens per second)
llama_print_timings:       total time =  140880.75 ms /  1043 tokens
```

```
 Sudden patchy hair loss, also known as alopecia areata, can be a distressing condition t
hat affects millions of people worldwide. While there is no definitive cure for this cond
ition, there are several effective treatment options available that can help promote hair
growth and reduce inflammation. Here are some of the most common treatment options for su
dden patchy hair loss:
```

```
1. Topical corticosteroids: Over-the-counter (OTC) creams or prescription-strength cortic
osteroids can help reduce inflammation and promote hair growth. These medications can be
applied directly to the affected area.
2. Minoxidil: Minoxidil (Rogaine) is a solution that is applied directly to the scalp to
stimulate hair growth. It is available OTC and can help slow down hair loss and promote n
ew growth.
3. Anthralin: Anthralin is a medication that is applied directly to the affected area to
reduce inflammation. It can help promote hair growth and is often used in combination wit
h corticosteroids.
4. Phototherapy: Exposure to specific wavelengths of light can help reduce inflammation a
nd promote hair growth. UVB phototherapy is a common treatment option for alopecia areata
.
5. Immunotherapy: Immunotherapy involves injecting a small amount of a substance that tri
ggers an allergic response into the affected area. This can help stimulate hair growth.
6. Hair transplantation: In severe cases of alopecia areata, hair transplantation may be
an option. This involves transplanting healthy hair follicles from one part of the body t
o another.
7. Dietary changes: Making dietary changes such as increasing omega-3 fatty acid intake,
reducing stress, and consuming foods rich in vitamin D can help promote hair growth.
8. Platelet-rich plasma (PRP) therapy: PRP therapy involves injecting platelet-rich plasm
a (PRP) into the affected area. PRP contains growth factors that can help stimulate hair
growth.
9. Low-level laser therapy (LLLT): LLLT involves exposure to low-level laser light that c
an help reduce inflammation and promote hair growth.
10. Home remedies: Some home remedies such as coconut oil, olive oil, castor oil, and ros
emary essential oil may help promote hair growth. However, there is limited scientific ev
idence to support these claims.

Possible causes of sudden patchy hair loss include:

1. Autoimmune disorders: Alopecia areata is an autoimmune disorder that can cause sudden
patchy hair loss. Other autoimmune disorders such as thyroiditis, vitiligo, and lupus can
also cause hair loss.
2. Hormonal imbalances: Hormonal imbalances such as thyroid disorders can cause hair loss
.
3. Stress: Prolonged stress can cause hair loss. This is because stress can lead to an in
crease in cortisol levels, which can disrupt normal hair growth cycles.
4. Nutritional deficiencies: Deficiencies in essential nutrients such as iron, zinc, and
biotin can cause hair loss.
5. Skin conditions: Skin conditions such as eczema, psoriasis, and seborrheic dermatitis
can cause hair loss.
6. Infections: Fungal infections such as ringworm can cause hair loss.
7. Traction alopecia: Traction alopecia is a type of hair loss that is caused by constant
pulling or tension on the hair follicles. This can occur due to tight hairstyles such as
braids, ponytails, or cornrows.
8. Telogen effluvium: Telogen effluvium is a condition that occurs when there is a sudden
increase in the number of hair follicles that stop growing and enter the resting phase. T
his can cause sudden patchy hair loss.
9. Medications: Certain medications such as chemotherapy drugs, blood thinners, and birth
control pills can cause hair loss.
10. Genetics: Genetic factors can play a role in sudden patchy hair loss. Some people may
inherit a predisposition to hair loss from their parents.
```

Query 4: What treatments are recommended for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function?

In [10]:

```
user_prompt = "What treatments are recommended for a person who has sustained a physical
injury to brain tissue, resulting in temporary or permanent impairment of brain function?
"
response = generate_llama_response(user_prompt)
print(response)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =     377.81 ms /   654 runs   (    0.58 ms per tok
en,  1731.02 tokens per second)
llama_print_timings: prompt eval time =     534.98 ms /    63 tokens (    8.49 ms per tok
```

```
llama_print_timings: prompt eval time =     534.98 ms /    63 tokens (    8.49 ms per tok
en,   117.76 tokens per second)
llama_print_timings:        eval time =   89085.28 ms /   653 runs   (  136.42 ms per tok
en,     7.33 tokens per second)
llama_print_timings:       total time =   92441.45 ms /   716 tokens
```

Treatments for a person who has suffered a physical injury to brain tissue and experienc
ed temporary or long-term impairment of brain function can vary depending on the type and
severity of the injury. Here are some possible treatment options:

1. Medications: To control inflammation, reduce swelling, and prevent further damage, med
ications such as corticosteroids and anticonvulsants may be prescribed. Pain relievers an
d muscle relaxants may also be used to manage symptoms.
2. Rehabilitation therapy: Physical, occupational, and speech therapy may be necessary to
help regain lost functions and improve cognitive and motor skills. These therapies can be
done individually or in combination and may involve exercises, activities, and equipment
tailored to the individual's needs.
3. Surgery: In some cases, surgery may be necessary to relieve pressure on the affected a
rea of the brain, remove blood clots or dead tissue, or repair damaged blood vessels. Min
imally invasive procedures can be performed using advanced imaging techniques such as MRI
or CT scans.
4. Stem cell therapy: Stem cells have been shown to promote healing and regeneration in d
amaged tissues. Stem cell therapy may be used to repair damaged brain cells and improve c
ognitive and motor functions.
5. Hyperbaric oxygen therapy: This involves breathing pure oxygen in a pressurized chambe
r. It can increase oxygen flow to damaged areas of the brain and promote healing.
6. Neuroplasticity-based therapies: These therapies aim to reorganize and adapt the brain
's neural pathways and functions. Techniques such as cognitive training, non-invasive bra
in stimulation (NIBS), and functional electrical stimulation (FES) may be used.
7. Lifestyle changes: Making lifestyle changes such as regular exercise, stress managemen
t techniques, and a healthy diet can help improve overall health and well-being.
8. Assistive technology: Assistive technology such as wheelchairs, walkers, and communica
tion devices may be necessary to help individuals with long-term disabilities perform dai
ly activities and communicate effectively.
9. Psychological support: Psychological support and counseling can help individuals and t
heir families cope with the emotional and psychological impact of brain injuries. Support
groups can also provide valuable resources and connections.
10. Neuroprotective agents: Research is ongoing into neuroprotective agents that can prot
ect the brain from further damage and promote healing. These agents may include nutrition
al supplements such as omega-3 fatty acids and antioxidants.

It's important to note that each individual's treatment plan will be unique and may invol
ve a combination of these options. It's essential to work with a healthcare team experien
ced in treating brain injuries to determine the best course of treatment.

Query 5: What are the necessary precautions and treatment steps for a person who has fractured their leg during a hiking trip, and what should be considered for their care and recovery?

In [11]:

```
user_prompt = "What are the necessary precautions and treatment steps for a person who ha
s fractured their leg during a hiking trip, and what should be considered for their care
and recovery?"
response = generate_llama_response(user_prompt)
print(response)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =     405.71 ms /   695 runs   (    0.58 ms per tok
en,  1713.05 tokens per second)
llama_print_timings: prompt eval time =     572.62 ms /    69 tokens (    8.30 ms per tok
en,   120.50 tokens per second)
llama_print_timings:        eval time =   95191.92 ms /   694 runs   (  137.16 ms per tok
en,     7.29 tokens per second)
llama_print_timings:       total time =   98851.77 ms /   763 tokens
```

Sure, I'd be happy to help! If someone has fractured their leg while hiking, it is essen
tial to take immediate action to ensure proper care and prevent further complications. He
re are some necessary precautions and treatment steps that should be taken:

```
1. Stop activity: The first step is to stop any further activity that may exacerbate the
injury. This includes halting any further hiking or walking.
2. Assess the injury: Assess the severity of the injury by checking for any deformities,
swelling, or open wounds. If the fracture is displaced or open, do not move the affected
limb.
3. Immobilize the affected limb: Use a splint or brace to immobilize the affected limb. T
his will help reduce pain and prevent further damage.
4. Elevate the affected limb: Elevate the affected limb above heart level to reduce swell
ing.
5. Apply ice: Apply ice packs to reduce pain and swelling. Wrap ice packs in a towel or c
loth to avoid direct contact with the skin.
6. Monitor vital signs: Monitor the patient's vital signs, including pulse, breathing rat
e, and blood pressure.
7. Call for help: Call for emergency medical help as soon as possible. If you are in a re
mote area, call for help using a satellite phone or other communication device.
8. Provide pain management: Provide pain management using over-the-counter pain medicatio
n or prescription medication if available.
9. Stay hydrated: Encourage the patient to stay hydrated by drinking plenty of water.
10. Monitor for complications: Monitor the patient for any signs of complications, such a
s infection or nerve damage.

During the recovery process, it is essential to follow your healthcare provider's instruc
tions carefully. This may include undergoing physical therapy to regain strength and mobi
lity in the affected limb. It is also important to avoid any activities that may exacerba
te the injury or cause further complications.

In addition to these medical considerations, there are also some general considerations t
hat should be taken into account when caring for someone with a broken leg:

1. Keep the affected limb clean and dry: Keep the affected limb clean and dry to prevent
infection.
2. Monitor wound healing: Monitor wound healing carefully and seek medical attention if t
here are any signs of infection or delayed healing.
3. Provide emotional support: Provide emotional support to help the patient cope with the
injury and the recovery process.
4. Encourage rest: Encourage rest and avoid any activities that may exacerbate the injury
.
5. Follow up with medical providers: Follow up with medical providers regularly to ensure
proper healing and address any concerns or complications that may arise.

Overall, caring for someone with a broken leg requires careful attention to medical needs
as well as emotional support. By following these guidelines and seeking medical attention
as needed, you can help ensure proper healing and prevent further complications.
```

Question Answering using LLM with Prompt Engineering

In [12]:

```
system_prompt = "Answer the question based on the user prompt and summarize it. Start eac
h using a marker "
```

Query 1: What is the protocol for managing sepsis in a critical care unit?

In [13]:

```
user_input = system_prompt+"\n"+ "What is the protocol for managing sepsis in a critical
care unit?"
response = generate_llama_response(user_input)
print(response)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =     358.71 ms /   642 runs   (    0.56 ms per tok
en,  1789.77 tokens per second)
llama_print_timings: prompt eval time =     592.71 ms /    68 tokens (    8.72 ms per tok
en,   114.73 tokens per second)
llama_print_timings:        eval time =   87445.40 ms /   641 runs   (  136.42 ms per tok
```

```
en,     7.33 tokens per second)
llama_print_timings:       total time =   90832.34 ms /   709 tokens
```

```
 Hi there! Based on your query, I understand that you're looking for information about t
he protocol for managing sepsis in a critical care unit. Here's a summary of the standard
of care for sepsis management in a critical care setting:

 Early recognition and timely intervention are crucial in managing sepsis. Critical care
providers should be vigilant in monitoring patients for signs of sepsis, such as fever, t
achycardia, tachypnea, and altered mental status.

 The Sepsis-3 definition should be used to diagnose sepsis: "Sepsis is defined as organ
dysfunction caused by a dysregulated host response to an infection." Organ dysfunction ca
n be identified by measuring serum lactate levels, white blood cell count, and blood pres
sure.

 The Surviving Sepsis Campaign (SSC) guidelines recommend a sequential approach to manag
ing sepsis, starting with initial resuscitation and early antibiotic administration. The
goal is to achieve hemodynamic stability, restore tissue perfusion, and address any under
lying infections.

 Critical care providers should closely monitor patients' vital signs, fluid status, and
oxygenation, adjusting interventions as needed to maintain stability. Invasive monitoring
, such as arterial lines and pulmonary artery catheters, may be used to guide therapy.

 Early recognition of severe sepsis and septic shock is essential, as these conditions a
re associated with higher morbidity and mortality. Management of severe sepsis and septic
shock should be individualized, taking into account the patient's underlying conditions,
organ dysfunction, and response to initial therapies.

 In addition to antibiotics, other interventions that may be used in the management of s
epsis include vasopressors, mechanical ventilation, and dialysis. The choice of intervent
ions will depend on the patient's specific needs and clinical status.

 Close collaboration between critical care providers, infectious disease specialists, an
d other members of the healthcare team is essential in managing patients with sepsis. Ear
ly consultation with specialists can help ensure that patients receive timely and appropr
iate interventions.

 Finally, it's important to remember that sepsis management involves more than just trea
ting the underlying infection. Critical care providers should also address any underlying
comorbidities, such as heart failure, diabetes, or chronic obstructive pulmonary disease
(COPD), that may be exacerbating the patient's condition.

I hope this information helps! If you have any further questions or need more specific gu
idance, please don't hesitate to ask.
```

Comments: As we can see, markers help us clearly delineate the answers. We can see the context set and diagnosis and prevention covered in points, just as we asked the prompt to do.

Query 2: What are the common symptoms for appendicitis, and can it be cured via medicine? If not, what surgical procedure should be followed to treat it?

In [14]:

```
user_prompt = system_prompt+ "\n" + "What are the common symptoms for appendicitis, and c
an it be cured via medicine? If not, what surgical procedure should be followed to treat
it?"
response = generate_llama_response(user_prompt)
print(response)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =     289.18 ms /   489 runs   (    0.59 ms per tok
en,  1691.00 tokens per second)
llama_print_timings: prompt eval time =     566.19 ms /    65 tokens (    8.71 ms per tok
en,   114.80 tokens per second)
llama_print_timings:        eval time =   66815.35 ms /   488 runs   (  136.92 ms per tok
en,     7.30 tokens per second)
```

```
llama_print_timings:       total time =   69459.11 ms /   553 tokens
```

```
 Hi there! Based on your query, I'd be happy to help you understand the common symptoms
of appendicitis and the available treatment options. Here's a summary of the information
you're looking for:

Common Symptoms of Appendicitis:
 Appendicitis is characterized by sudden and severe pain in the lower right abdomen, oft
en accompanied by:
 Nausea and vomiting
 Loss of appetite
 Abdominal tenderness and swelling
 Fever (usually above 100°F)
 Loss of bowel movement or constipation
 Abdominal guarding (tightening of the abdominal muscles)
 Rigidity (stiffness) in the abdomen

Can Appendicitis Be Cured Via Medicine?
 In some cases, appendicitis may be treated with antibiotics if the inflammation is caus
ed by an infection. However, if the inflammation is due to a blockage or obstruction of t
he appendix, medication may not be effective in curing the condition.

Surgical Procedure for Treating Appendicitis:
 The most common surgical procedure for treating appendicitis is an appendectomy, which
involves removing the inflamed or infected appendix. The two types of appendectomies are:
 Open Appendectomy: This is the traditional method of removing the appendix through an i
ncision in the abdomen.
 Laparoscopic Appendectomy: This is a minimally invasive procedure that uses a laparosco
pe (a thin tube with a camera) to remove the inflamed or infected appendix through small
incisions in the abdomen.

It's important to note that if you suspect you or someone else may have appendicitis, you
should seek medical attention immediately! Delaying treatment can lead to complications s
uch as the appendix rupturing, which can cause potentially life-threatening infections.
```

Comments: The prompt engineering system promt has helped us with clear answers using emoji's and segregated the symtomps, treatments and the surgical procedures. In the surgical procedures - we start off with the common procedure and then the specifics, thus we get a nice and curated answer.

Query 3: What are the effective treatments or solutions for addressing sudden patchy hair loss, commonly seen as localized bald spots on the scalp, and what could be the possible causes behind it?

In [15]:

```
user_prompt = system_prompt+ "\n" + "What are the effective treatments or solutions for a
ddressing sudden patchy hair loss, commonly seen as localized bald spots on the scalp, an
d what could be the possible causes behind it?"

response = generate_llama_response(user_prompt)
print(response)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =     356.50 ms /   602 runs   (    0.59 ms per tok
en,  1688.62 tokens per second)
llama_print_timings: prompt eval time =     577.79 ms /    69 tokens (    8.37 ms per tok
en,   119.42 tokens per second)
llama_print_timings:        eval time =   82052.86 ms /   601 runs   (  136.53 ms per tok
en,     7.32 tokens per second)
llama_print_timings:       total time =   85221.36 ms /   670 tokens
```

```
 Hi there! Based on your query, I understand that you're experiencing sudden patchy hair
loss in the form of localized bald spots on your scalp. You're looking for effective trea
tments and potential causes of this issue. Here's my response:

♀ Treatment options: There are several treatment options available to address sudden pat
chy hair loss. These include:
```

1️⃣ Medications: Minoxidil (Rogaine) and finasteride (Propecia) are two FDA-approved medications that can help stimulate hair growth and slow down hair loss.

2️⃣ Low-level laser therapy (LLLT): This non-invasive treatment uses low-level lasers or light-emitting diodes to stimulate hair growth by increasing blood flow and promoting cellular activity.

3️⃣ Platelet-rich plasma (PRP) therapy: This treatment involves injecting platelet-rich plasma (derived from your own blood) into your scalp to stimulate hair growth by promoting cellular activity and increasing blood flow.

4️⃣ Dietary changes: Making changes to your diet can help promote healthier hair growth. Eating foods rich in vitamins D, C, E, and B can help support hair growth.

 Causes: There are several potential causes of sudden patchy hair loss, including:

1️⃣ Hormonal imbalances: Hormonal fluctuations can cause hair loss in both men and women. This can occur due to pregnancy, childbirth, menopause, or hormonal imbalances caused by polycystic ovary syndrome (PCOS) or thyroid disorders.

2️⃣ Autoimmune conditions: Conditions like alopecia areata, lupus, and rheumatoid arthritis can cause sudden patchy hair loss due to inflammation and an immune system attack on hair follicles.

3️⃣ Infections: Fungal infections like ringworm can cause hair loss in patches.

4️⃣ Traction alopecia: This is a type of hair loss caused by constant pulling or tension on the hair follicles, often due to tight hairstyles like braids, ponytails, or cornrows.

 It's important to consult with a dermatologist to determine the underlying cause of your hair loss and develop an appropriate treatment plan. They may recommend a combination of these treatment options or suggest other personalized solutions based on your specific needs.

Comments: The treatment options for blad spots are clearly put in there using point wise markers. After that the causes are clearly put using point wise markers. The answers are clear, using the system promt we have gotten the answers.

Query 4: What treatments are recommended for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function?

In [16]:

```
user_prompt = system_prompt+ "\n" + "What treatments are recommended for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain func"

response = generate_llama_response(user_prompt)
print(response)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =     247.84 ms /   407 runs   (    0.61 ms per token,  1642.19 tokens per second)
llama_print_timings: prompt eval time =     534.85 ms /    62 tokens (    8.63 ms per token,   115.92 tokens per second)
llama_print_timings:        eval time =   55007.42 ms /   406 runs   (  135.49 ms per token,     7.38 tokens per second)
llama_print_timings:       total time =   57202.89 ms /   468 tokens
```

 Hi there! Based on your question, I understand that you're looking for information on treatments for a person who has suffered a physical injury to their brain tissue, leading to temporary or permanent impairment of brain function. Here's what I can help with:

 Treatment options for brain injuries depend on the severity and location of the injury, as well as the individual's overall health and medical history. Some common treatment approaches include:

1️⃣ Medications: To manage symptoms such as pain, inflammation, and anxiety/depression. The se may include pain relievers, anti-seizure medications, and antidepressants.

2️⃣ Rehabilitation therapies: Such as physical, occupational, and speech therapy to help re gain lost functions and improve cognitive abilities. These may include exercises to stren gthen affected muscles, improve balance and coordination, and practice communication skil ls.

3️⃣ Surgery: In some cases, surgery may be necessary to relieve pressure on the brain, repa ir damaged blood vessels, or remove a tumor that may be causing the injury.

4️⃣ Lifestyle changes: Making healthy lifestyle changes such as regular exercise, a balance d diet, and stress management techniques can help improve overall health and well-being a fter a brain injury.

 It's important to note that every individual's recovery process is unique and may take time. It's essential to work closely with healthcare professionals to develop a personali zed treatment plan that addresses your specific needs and goals.

 If you have any further questions or concerns, feel free to ask! I'm here to help you n avigate this process and provide support every step of the way.

Comments: For a injury to a brian tissue, with a summary on the treatment options. Then the medications, therapy options, Surgery options nad lifestyle changes are clearly delineated using markers. Thus a clear answer to our questions that we enabled using prompt engineering.

Query 5: What are the necessary precautions and treatment steps for a person who has fractured their leg during a hiking trip, and what should be considered for their care and recovery?

In [17]:

```
user_prompt = system_prompt+ "\n" + "What are the necessary precautions and treatment ste
ps for a person who has fractured their leg during a hiking trip, and what should be cons
idered for their care and recovery?"

response = generate_llama_response(user_prompt)
print(response)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =     462.53 ms /   785 runs   (    0.59 ms per tok
en,  1697.18 tokens per second)
llama_print_timings: prompt eval time =     592.64 ms /    69 tokens (    8.59 ms per tok
en,   116.43 tokens per second)
llama_print_timings:        eval time =  109230.11 ms /   784 runs   (  139.32 ms per tok
en,     7.18 tokens per second)
llama_print_timings:       total time =  113443.91 ms /   853 tokens
```

 Necessary Precautions and Treatment Steps for a Fractured Leg During a Hiking Trip:

If you have fractured your leg while hiking, it is essential to take immediate action to prevent further damage and ensure proper healing. Here are some necessary precautions and treatment steps to consider:

1️⃣ Stop activity: Cease all physical activity to avoid exacerbating the injury and potenti ally causing further damage.

2️⃣ Assess the injury: Evaluate the severity of the fracture by checking for signs of swell ing, deformity, or open wounds. If you suspect a compound fracture or have difficulty mov ing the affected limb, seek medical attention immediately.

3️⃣ Immobilize the affected limb: Use a splint or brace to immobilize the injured leg to pr event further movement and reduce pain. You can create a makeshift splint using materials found in your backpack or seek assistance from fellow hikers or medical professionals.

4️⃣ Apply ice: Apply ice packs or cold compresses to reduce swelling and alleviate pain. Wr ap ice in a towel or cloth to avoid direct contact with your skin.

5️⃣ Elevate the affected limb: Elevate your injured leg above heart level to reduce swelling and promote blood flow.

6️⃣ Monitor vital signs: Check your pulse, breathing rate, and blood pressure regularly to ensure there are no signs of shock or infection.

7️⃣ Seek medical attention: If you experience severe pain, numbness, or tingling sensations, or if you suspect a compound fracture or open wound, seek medical attention as soon as possible.

8️⃣ Consider your environment: If you're in a remote area with limited access to medical facilities, consider your surroundings before attempting to move or treat your injury. In some cases, it may be safer to stay put and wait for rescue rather than risk further injury or complications.

9️⃣ Prepare for recovery: Once you receive medical attention, focus on your recovery by following your healthcare provider's instructions, taking prescribed medication as directed, and attending follow-up appointments to ensure proper healing.

🔟 Additional Considerations for Care and Recovery:

In addition to these essential precautions and treatment steps, there are several other factors to consider when caring for your injured leg:

1️⃣ Rest: Allow your body time to rest and heal. Avoid putting weight on your injured leg or engaging in strenuous activities that may exacerbate your condition.

2️⃣ Pain management: Manage your pain effectively with prescribed medication or alternative methods like breathing exercises or meditation.

3️⃣ Physical therapy: Once your initial pain and swelling have subsided, engage in physical therapy exercises to regain strength, flexibility, and range of motion in your affected limb.

4️⃣ Monitoring: Keep track of your progress by monitoring your symptoms, taking note of any changes or complications, and attending follow-up appointments with your healthcare provider.

5️⃣ Support: Seek emotional support from friends, family, or support groups to help cope with the physical and mental challenges of recovering from a fracture.

Comments: This is a pretty detailed answer that we got using markers and prompt engineering. The first part goes over the action items when one fractures the leg during hiking and then the second part covers the treatment steps in a point-wise fashion.

Data Preparation for RAG

Loading the Data

In [20]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

In [21]:

```
pdf_path='/content/drive/My Drive/python/medical_diagnosis_manual.pdf'
```

In [22]:

```
pdf_loader = PyMuPDFLoader(pdf_path)
```

In [23]:

```
manual = pdf_loader.load()
```

Data Overview

Checking the first 5 pages

In [24]:

```
for i in range(5):
    print(f"Page Number : {i+1}",end="\n")
    print(manual[i].page_content,end="\n")
```

```
Page Number : 1
prak.kris4074@gmail.com
TYGVHURWX1
ant for personal use by prak.kris4074@g
shing the contents in part or full is liable

Page Number : 2
prak.kris4074@gmail.com
TYGVHURWX1
This file is meant for personal use by prak.kris4074@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Page Number : 3
Table of Contents
1
Front    ..........
..........
..........
1
Cover    ..........
..........
..........
2
Front Matter    ..........
..........
..........
53
1 - Nutritional Disorders    ..........
..........
..........
53
Chapter 1. Nutrition: General Considerations    ..........
..........
59
Chapter 2. Undernutrition    ..........
..........
........
69
Chapter 3. Nutritional Support    ..........
..........
...
76
Chapter 4. Vitamin Deficiency, Dependency & Toxicity    ..........
..........
99
Chapter 5. Mineral Deficiency & Toxicity    ..........
..........
108
Chapter 6. Obesity & the Metabolic Syndrome    ..........
..........
120
2 - Gastrointestinal Disorders    ..........
..........
120
Chapter 7. Approach to the Patient With Upper GI Complaints    ..........
..........
132
Chapter 8. Approach to the Patient With Lower GI Complaints    ..........
```

Checking the number of pages

In [25]:

```
len(manual)
```

Out[25]:

```
4114
```

Data Chunking

In [26]:

```
text_splitter = RecursiveCharacterTextSplitter.from_tiktoken_encoder(
    encoding_name='cl100k_base',
    chunk_size=256,
    chunk_overlap= 20
)
```

In [27]:

```
document_chunks = pdf_loader.load_and_split(text_splitter)
```

In [28]:

```
len(document_chunks)
```

Out[28]:

15671

In [29]:

```
document_chunks[0].page_content
```

Out[29]:

'prak.kris4074@gmail.com\nTYGVHURWX1\nant for personal use by prak.kris4074@g\nshing the contents in part or full is liable'

In [30]:

```
document_chunks[1].page_content
```

Out[30]:

'prak.kris4074@gmail.com\nTYGVHURWX1\nThis file is meant for personal use by prak.kris4074@gmail.com only.\nSharing or publishing the contents in part or full is liable for legal action.'

In [31]:

```
document_chunks[2].page_content
```

Out[31]:

'Table of Contents\n1\nFront .....\n1\nCover .....\n2\nFront Matter .....\n53\n1 - Nutritional Disorders .....\n53\nChapter 1. Nutrition: General Considerations .....\n59\nChapter 2. Undernutrition .....\n69\nChapter 3. Nutritional Support .....\n76\nChapter 4. Vitamin Deficiency, Dependency & Toxicity .....\n99\nChapter 5. Mineral Deficiency & Toxicity .....\n108\nChapter 6. Obesity & the Metabolic Syndrome .....\n120\n2 - Gastrointestinal Disorders .....\n120\nChapter 7. Approach to the Patient With Upper GI Complaints .....\n132\nChapter 8. Approach to the Patient With Lower GI Complaints .....\n143\nChapter 9. Diagnostic & Therapeutic GI Procedures .....\n150\nChapter 10. GI Bleeding .....\n158\nChapter 11. Acute Abdomen & Surgical Gastroenterology .....\n172'

In [32]:

```
document_chunks[3].page_content
```

Out[32]:

'172\nChapter 12. Esophageal & Swallowing Disorders .....\n183\nChapter 13. Gastritis & Peptic Ulcer Disease .....\n196\nChapter 14. Bezoars

Embedding

In [33]:

```
embedding_model = SentenceTransformerEmbeddings(model_name="thenlper/gte-large")
```

```
The cache for model files in Transformers v4.22.0 has been updated. Migrating your old cache. This is a one-time only operation. You can interrupt this and resume the migration later on by calling `transformers.utils.move_cache()`.
```

```
/usr/local/lib/python3.11/dist-packages/huggingface_hub/file_download.py:795: FutureWarning: `resume_download` is deprecated and will be removed in version 1.0.0. Downloads always resume when possible. If you want to force a new download, use `force_download=True`.
  warnings.warn(
```

```
/usr/local/lib/python3.11/dist-packages/transformers/utils/generic.py:260: FutureWarning: `torch.utils._pytree._register_pytree_node` is deprecated. Please use `torch.utils._pytree.register_pytree_node` instead.
  torch.utils._pytree._register_pytree_node(
/usr/local/lib/python3.11/dist-packages/transformers/utils/generic.py:260: FutureWarning: `torch.utils._pytree._register_pytree_node` is deprecated. Please use `torch.utils._pytree.register_pytree_node` instead.
  torch.utils._pytree._register_pytree_node(
/usr/local/lib/python3.11/dist-packages/huggingface_hub/file_download.py:795: FutureWarning: `resume_download` is deprecated and will be removed in version 1.0.0. Downloads always resume when possible. If you want to force a new download, use `force_download=True`.
  warnings.warn(
```

In [34]:

```
embedding_1 = embedding_model.embed_query(document_chunks[0].page_content)
embedding_2 = embedding_model.embed_query(document_chunks[1].page_content)
```

In [35]:

```
print("Dimension of the embedding vector ",len(embedding_1))
len(embedding_1)==len(embedding_2)
```

```
Dimension of the embedding vector  1024
```

Out[35]:

```
True
```

In [36]:

```
embedding_1,embedding_2
```

Out[36]:

```
([-0.030667711049318314,
  -0.006570147350430489,
  0.006497826427221298,
  0.0017972574569284916,
  -0.00987029355019331,
  -0.031295690685510635,
  0.0072701615281403065,
  0.030280157923698425,
  0.014677949249744415,
  0.020160213112831116,
  0.042831361293792725,
  -0.00418648449704051,
  0.027324778959155083,
  -0.02847873792052269,
  -0.004401426296681166,
  -0.0024747084826231003,
  -0.014902308583259583,
  -0.04017169773578644,
  -0.013097486458718777,
  0.023816879838705063,
  0.006815614178776741,
  -0.00022374412219505757,
  -0.07968782633543015,
  -0.018362682312726974,
  -0.021604379639029503,
  0.019782662391662598,
  0.028130199760198593,
  0.007770710624754429,
  0.06253022700548172,
  0.03930293023586273,
  -0.02134544774889946,
  -0.0018150962423533201,
  0.01193803921341896,
  -0.04482073709368706,
  -0.0021922565065324306,
  0.0017927228473126888,
  0.0599907748401165,
  -0.02030394785106182,
  -0.02922477014362812,
  -0.03931586816906929,
  0.010803038254380226,
  -0.0033092752564698458,
  0.06435836106538773,
  -0.03838593512773514,
  -0.02371809631586075,
  -0.014127854257822037,
  -0.016454150900244713,
  -0.00925816036760807,
  0.012585710734128952,
  -0.023468462750315666,
  0.03690813109278679,
  0.010408545844256878,
  0.04197157546877861,
  -0.010221421718597412,
  0.013609415851533413,
  0.015986155718564987,
  0.0160685945302248,
  -0.04263133555650711,
  -0.05135060101747513,
  0.03611191734671593,
  -0.005390521138906479,
```

-0.008651861920952797,
0.03518470376729965,
-0.05738978087902069,
0.024846212938427925,
0.009418374858796597,
0.018472349271178246,
-0.004860172513872385,
0.0011366403196007013,
-0.030085543170571327,
-0.0052139475010335445,
-0.006083172746002674,
-0.029861949384212494,
-0.012662461027503014,
0.012109152972698212,
-0.018537219613790512,
0.007302275858819485,
-0.006045094225555658,
0.0033599259331822395,
0.02282407321035862,
0.00427182437852025,
0.03919042646884918,
0.032615143805742264,
0.02189435251057148,
-0.05054276064038277,
-0.007973727770149708,
0.005531232338398695,
0.007394953165203333,
0.014442481100559235,
0.002364972373470664,
-0.014330286532640457,
0.041874781250953674,
-0.014707870781421661,
-0.0011774773010984063,
0.029810281470417976,
0.04346686229109764,
-0.05391169339418411,
0.0321505032479763,
-0.04069077968597412,
-0.0038531161844730377,
0.047854553908109665,
0.02549039013683796,
0.005716653540730476,
0.07674836367368698,
-0.04164014756679535,
0.004057017620652914,
0.012692287564277649,
-0.020745795220136642,
-0.0023425989784300327,
-0.013861779123544693,
0.006495391950011253,
-0.029871884733438492,
0.002186585683375597,
-0.009178126230835915,
0.020937984809279442,
0.033984459936618805,
0.009185896255075932,
0.028614111244678497,
-0.01849869638681411 7,
-0.007227133493870497,
0.012508070096373558,
0.01723508909344673,
0.03592178598046303,
-0.024795962497591972,
0.016139417886734 01,
-0.04422364383935928,
-0.008107912726700306,
0.029703987762331963,
-0.027183329686522484,
-0.004770184867084026,
0.006087838672101498,
-0.037688370794057846,
-0.04041524603962898,

-0.0022118291817605495,
-0.016743239015340805,
0.022867752239108086,
-0.016988247632980347,
0.015402675606310368,
0.055999476462602615,
-0.0594225637614727,
0.027049057185649872,
0.046127378940582275,
-0.005496884696185589,
0.08498737215995789,
-0.01644025556743145,
0.0349142923951149,
0.0007643384742550552,
0.000796400650870055,
-0.03400130197405815,
0.028571777045726776,
-0.03585630655288696,
-0.017582200467586517,
0.0032868687994778156,
0.032974131405353546,
0.015883108600974083,
-0.003630031133070588,
-0.01599031873047352,
3.906817437382415e-05,
0.030870595946907997,
0.0515575110912323,
-0.00016234659415204078,
-0.023087648674845695,
-0.01059850212186575,
0.04619985446333885,
0.01606108248233795,
0.07550642639398575,
-0.012646405026316643,
-0.020732374861836433,
-0.019663697108626366,
-0.04386083409190178,
0.044167809188365936,
-0.020309410989284515,
-0.009145310148596764,
0.01608838140964508,
0.03275945037603378,
0.03943128511309624,
0.04119250923395157,
0.002845353912562132,
0.046831611543893814,
0.04317736253142357,
-0.00959780067205429,
0.01742536760866642,
0.025658665224909782,
0.07462934404611588,
-0.012468550354242325,
-0.028401052579283714,
-0.016582395881414413,
-0.03829408437013626,
-0.05305330455303192,
-0.018131639808416367,
0.021953409537672997,
0.014070349745452404,
-0.03248657286167145,
0.01984586752951145,
-0.015343286097049713,
0.008848981000483036,
-0.019848883152008057,
-0.003590857610106468,
-0.00711108697578311,
-0.0764235332608223,
-0.015615459531545639,
0.026589713990688324,
-0.02970617450773716,
0.022795584052801132,
0.0003197499318048358,

-0.025572028011083603,
0.008835524320602417,
0.06953737139701843,
-0.03070816770195961,
-0.01569908857345581,
0.04137272760272026,
0.013704013079404831,
-0.03359191492199898,
-0.021807603538036346,
-0.013501226902008057,
0.007104857824742794,
-0.033147234469652176,
0.05832117795944214,
-0.041447412222623825,
-0.0410129576921463,
0.021335961297154427,
0.013704691082239151,
0.024181846529245377,
0.020644258707761765,
-0.009780784137547016,
0.0033190695103257895,
0.007690915372222662,
0.03608373552560806,
-0.025580471381545067,
-0.007730620447546244,
-0.010238585062325,
0.03397812321782112,
0.0209866464138031,
0.07469046860933304,
0.024385992437601O9,
0.024981623515486717,
0.0438772588968277,
0.029355671256780624,
-0.010323736816644669,
0.014919260516762733,
-0.011265852488577366,
0.02242143079638481,
0.03732208162546158,
0.0031176400370895863,
0.023065900430083275,
0.04712194576859474,
-0.020663440227508545,
0.00819079577922821,
-0.030127786099910736,
0.0206709373742342,
-0.01873704232275486,
0.038193631917238235,
0.02026071585714817,
0.042573027312755585,
-0.04184442758560181,
-0.01020866446197033,
0.05389115586876869,
0.047704778611660004,
-0.040091294795274734,
-0.0025403716135770082,
0.0055327033624053,
0.027198465541005135,
-0.008317386731505394,
0.018974117934703827,
0.03315236046910286,
0.0023766711819916964,
0.018661126494407654,
-0.006692327558994293,
-0.015021146275103092,
-0.03683646395802498,
-0.0496012307703495,
-0.032313503324985504,
-0.051387686282396317,
-0.03786589205265045,
-0.05378943309187889,
-0.024471864104270935,
0.03480275720357895,

-0.029472436755895615,
0.02745421975851059,
-0.01116927433758974,
0.004086684435606003,
-0.021287761628627777,
-0.02146902307868004,
0.024497302249073982,
0.03544865548610687,
0.0646476075053215,
-0.02628922462463379,
0.03187665343284607,
0.0011695808498188853,
0.016568608582019806,
-0.032897256314754486,
0.0027102085296064615,
-0.004128389526158571,
-0.009280276484787464,
0.040599923580884933,
0.009724562056362629,
-0.010468894615769386,
-0.014813740737736225,
-0.03144726902246475,
-0.028965812176465988,
0.005136802326887846,
0.011869795620441437,
-0.016865188255906105,
-0.005998502019792795,
-0.04794585704803467,
0.02997451089322567,
0.016405005007982254,
-0.030928989872336388,
0.04638967663049698,
0.01240694522857666,
-0.030439047142863274,
0.02327227219939232,
0.013002904132008553,
0.0020391475409269333,
-0.052413471043109894,
0.03037467785179615,
0.05865592136979103,
0.01733740232884884,
-0.047260917723178864,
-0.024311428889632225,
-0.060315970331430435,
0.022130340337753296,
0.0021332993637770414,
-0.02390049770474434,
-0.019342679530382156,
-0.006966683082282543,
0.04789457097649574,
-0.09014692902565002,
0.03179231286048889,
-0.047682736068964005,
-0.056867264211177826,
-0.0497671477496624,
-0.03370625898241997,
0.007660281378775835,
0.004575103987008333,
0.00955266784876585,
-0.0017694663256406784,
-0.027344319969415665,
0.009027599357068539,
0.0019710713531821966,
0.042424965649843216,
-0.028011884540319443,
0.01030919048935175,
0.04043492302298546,
0.0019688038155436516,
0.030891109257936478,
0.043396707624197006,
0.005296668503433466,
-0.01953110098838806,

-0.00012911226076539606,
0.0020701822359114885,
0.01429713238030672,
0.03636537864804268,
0.03134332597255707,
0.008670772425830364,
0.05557093396782875,
-0.01675398088991642,
-0.0018091440433636308,
0.02863234281539917,
0.01823057420551777,
0.02220369130373001,
0.018339455127716064,
0.007328955922275782,
-0.019176406785845757,
-0.02515726163983345,
-0.03592034429311752,
0.017335837706923485,
0.028949683532118797,
0.04863245040178299,
-0.06818018108606339,
0.05829743295907974,
-0.0030690638814121485,
-0.03866007179021835,
0.04179378226399422,
-0.023821495473384857,
0.0012187286047264934,
0.05370040610432625,
0.0020405626855790615,
0.05093923583626747,
-0.012512419372797012,
-0.0118632260710001,
-0.037454139441251755,
0.03779209777712822,
0.04101119190454483,
-0.0017124577425420284,
0.022806409746408463,
-0.012285825796425343,
-0.015479062683880329,
-0.02613632194697857,
-0.030380424112081528,
-0.029155269265174866,
-0.012869901955127716,
0.012233363464474678,
-0.008758967742323875,
-0.04752451926469803,
-0.07555370032787323,
0.011108782142400742,
-0.0031596843618899584,
0.045685335993766785,
0.013056199997663498,
0.03937710449099541,
0.024720381945371628,
0.030599374324083328,
0.03997565805912018,
-0.03722161799669266,
-0.011440088972449303,
-0.03704148903489113,
0.04817650467157364,
0.029770800843834877,
-0.010123186744749546,
-0.011306450702250004,
0.005223272833973169,
-0.026112042367458344,
0.044833987951278687,
0.0028060211334377527,
-0.009962176904082298,
-0.00591618986800313,
0.006644239649176598,
-0.009680649265646935,
0.016320517286658287,
-0.04267548769712448,

-0.013181317597627 64,
-0.020336672663688 66,
0.046536799520254135,
0.03431429713964462,
-0.051729269325733185,
0.015246156603097916,
-0.030502121895551 68,
0.04109479486942291,
0.020528089255094528,
0.014165760017931461,
-0.038721051067113876,
-0.011527721770107746,
-0.031048860400915146,
-0.04570505768060684,
-0.012352778576314 45,
0.035776760429143906,
-0.010716299526393414,
0.009343978017568588,
-0.03724646568298 34,
0.025690326467156 41,
0.012696208432316 78,
0.0046674697659909725,
-0.002132116584107 2798,
-0.004502715077251196,
-0.000270477961748 8384,
-0.000387778854928 91073,
0.048508554697036 74,
-0.021657541394233704,
-0.036817479878664 02,
0.021210655570030212,
-0.049813609570264816,
0.021308202296495438,
0.02208794280886 65,
-0.042862821370363235,
-3.140591434203088e-05,
0.05538576468825 34,
0.031304717063903 81,
0.030415682122111 32,
0.002653485862538 2185,
0.013530943542718887,
0.005058097187429 6665,
0.025632970035076 14,
-0.061762664467096 33,
-0.027592241764068604,
0.039697192609310 15,
-0.002291693119332 1943,
-0.029301518574357033,
0.044752933084964 75,
0.044470489025115 97,
-0.011263516731560 23,
0.017827294766902924,
0.007742962334305048,
-0.029720902442932 13,
0.030146747827529907,
-0.051286071538925 17,
0.025149941444396973,
-0.026074627414345 74,
-0.040905978530645 37,
-0.004268684424459934,
-0.034550260752439 5,
0.046090438961982 73,
-0.030870249494910 24,
-0.030693804845213 89,
-0.019038399681448936,
-0.055159121751785 28,
-0.009122673422098 16,
0.018664587289094925,
-0.025697205215692 52,
0.024261524900794 03,
0.008701246231794357,
-0.010643599554896355,
-0.015986032783985138,

-0.007872962392866611,
0.0018899344140663743,
-0.011660151183605194,
-0.03419221192598343,
-0.0010454299626871943,
0.012716553173959255,
0.019325746223330498,
0.0004962935927323997,
-0.012664766050875187,
-0.05337584391236305,
0.028443442657589912,
0.017059670761227608,
0.009636025875806808,
-0.053602784872055054,
0.01519201509654522,
-0.013108710758388042,
-0.009010819718241692,
-0.036838825792074203,
0.025496788322925568,
-0.030042247846722603,
0.015275922603905201,
0.03419317677617073,
0.021316047757864,
0.02072531171143055,
-0.006331650074571371,
-0.045390985906124115,
0.039226386696100235,
0.03173299878835678,
-0.054837994277477264,
-0.018762847408652306,
0.026561954990029335,
0.002582704881206155,
0.036874230951070786,
0.04336308315396309,
-0.01494778785854578,
-0.05531127378344536,
-0.05927198752760887,
-0.011654620058834553,
-0.04924793541431427,
-0.029369011521339417,
-0.0589316263794899,
-0.034233495593070984,
-0.03236960619688034,
0.055761341005563736,
0.017926527187228203,
-0.019924679771065712,
0.010488610714673996,
-0.048582229763269424,
0.00802952703088522,
-0.010735864751040936,
-0.022989047691226006,
-0.019081544131040573,
-0.022361908107995987,
-0.003193065756931901,
0.06278125196695328,
0.009345250204205513,
0.026704883202910423,
-0.011557133868336678,
0.010335163213312626,
0.022157788276672363,
0.014311465434730053,
-0.04523728787899017,
-0.014242073521018028,
0.03166941553354263,
0.02288045547902584,
-0.016882827505469322,
0.03677394613623619,
-0.03132980689406395,
0.08913023769855499,
-0.04125075414776802,
-0.011328802444040775,
-0.006397287826985121,

0.0033923275768756866,
0.007718306500464678,
0.004071118775755167,
0.04019440710544586,
-0.03364356607198715,
0.010881537571549416,
-0.030404767021536827,
0.045140787959098816,
0.02729817107319832,
-0.008278841152787209,
-0.011401362717151642,
-0.03993795067071915,
-0.05395160987973213,
-0.07284966856241226,
0.02492854744195938,
-0.0211790781468153,
-0.033066485077142715,
-0.018552331253886223,
-0.019469594582915306,
0.03840285539627075,
-0.023589886724948883,
0.06980624049901962,
0.07442452758550644,
0.017778651788830757,
-0.015735726803541183,
-0.018934397026896477,
0.026609476655721664,
0.048899468034505844,
-0.056742098182439804,
-0.010455390438437462,
-0.011774835176765919,
-0.0067837913520634174,
0.006481414195150137,
-0.03156810998916626,
-0.037323206663131714,
-0.0611695758998394,
-0.017286276444792747,
0.06505610048770905,
-0.019800536334514618,
0.04022037237882614,
0.005216533318161964,
-0.04594351351261139,
-0.042367029935121536,
0.04936538636684418,
0.00892780814319849,
0.04260604456067085,
0.041957031935453415,
-0.01721007376909256,
-0.007172893267124891,
-0.0017884300323203206,
-0.001501666847616434,
0.0065068998374044895,
-0.05928504467010498,
0.06853990256786346,
-0.017057811841368675,
-0.03607157617807388,
-0.007810282986611128,
0.05942196026444435,
-0.05588817968964577,
-0.06001286953687668,
-0.010470542125403881,
0.002517164219170809,
0.005859531927853823,
-0.03379959240555763,
0.012076931074261665,
-0.013580955564975739,
-0.013829453848302364,
-0.0004373068222776055 3,
0.03976807743310928,
0.00938417948782444,
0.03869808092713356,
0.0032989608589559793,

0.031067095696926117,
-0.032229382544755936,
-0.000846642826218158,
0.0290265753865242,
-0.006508279126137495,
-0.008274857886135578,
-0.06499923020601273,
-0.0019283231813460588,
-0.019290128722786903,
-0.000775522524306297,
0.04073597490787506,
-0.009530194103717804,
0.017492009326815605,
0.02307985909283161,
-0.004233456216752529,
0.05488022044301033,
0.009907335974276066,
0.006211346480995417,
-0.0014492941554635763,
-0.03958825394511223,
-0.047294944524765015,
-0.02957727201282978,
0.021519280970096588,
-0.03139814734458923,
-0.0017069897148758173,
-0.060785748064517975,
0.007889272645115852,
0.0291120782494545,
0.002749639330431819,
0.002043351763859391,
-0.05681967735290527,
-0.04294932261109352,
-0.04715300723910332,
-0.03957296162843704,
-0.023408468812704086,
-0.009914679452776909,
-0.008295833133161068,
0.0299017746001482,
0.027291160076856613,
-0.037910111248493195,
-0.00197092373855412,
0.020808985456824303,
-0.03763609379529953,
0.03686487674713135,
-0.03065139427781105,
0.04185649752616882,
-0.06230732426047325,
-0.03730278089642525,
-0.03550027683377266,
-0.009362508542835712,
-0.028243742883205414,
0.012760964222252369,
-0.0015011070063337684,
0.02856801636517048,
0.0219410452991724,
0.02248375304043293,
0.016343461349606514,
-0.015663860365748405,
-0.024778949096798897,
-0.03350206837058067,
0.011180625297129154,
0.03394951671361923,
2.9999657272128388e-05,
0.008841671980917454,
0.018924467265605927,
-0.024039275944232 94,
0.04572807252407074,
-0.018368268385529518,
-0.0353335365653038,
-0.04483699053525925,
0.02139819599688053,
0.015483541414141655,

-0.004343998618423939,
-0.01976608671247959,
-0.024325599893927574,
-0.02424224466085434,
-0.052575502544641495,
0.014452275820076466,
0.0009713530889712274,
0.04124430567026138,
0.009047532454133034,
-0.02350863628089428,
-0.00715617137029767,
0.04882954806089401,
-0.009638765826821327,
0.009622607380151749,
0.004466172773391008,
0.019089454784989357,
0.023946011438965797,
0.007529066875576973,
0.038482844829559326,
0.008394251577556133,
0.06190451979637146,
-0.022486062720417976,
0.0008935174555517733,
0.005166222807019949,
0.0010800949530676007,
0.02968474291265011,
-0.01845862902700901,
-0.020303618162870407,
-0.05095713585615158,
0.01176913920789957,
0.004206712357699871,
0.05555233359336853,
0.003946698736399412,
0.004531353246420622,
0.010729287751019001,
-0.030819322913885117,
-0.04480405151844025,
0.01542014442384243,
-0.048463232815265656,
-0.03909207135438919,
0.047799695283174515,
-0.055300354957580566,
-0.02108045667409897,
-0.019006378948688507,
-0.05376021936535835,
-0.012408704496920109,
-0.0033414734061807394,
-0.028823599219322205,
-0.0269878301769495,
0.020096825435757637,
0.00214413576759398,
0.05230240523815155,
-0.03824419528245926,
-0.0056882151402533054,
0.009541554376482964,
0.07217094302177429,
-0.05402921885251999,
-0.04476121813058853,
0.008143369108438492,
0.025155313313007355,
-0.006802521646022797,
-0.017354385927319527,
0.008837427012622356,
-0.001201591338030994,
-0.01760481484234333,
-0.0013020860496908426,
0.02009515091776848,
0.03568635135889053,
0.0017292629927396774,
0.03762003034353256,
0.016995998099446297,
0.008231894113123417,

-0.0432530902326107,
0.02404957078397274,
0.012717985548079014,
-0.03913259506225586,
-0.002925222273916006,
0.04048219323158264,
0.021751007065176964,
-0.015061567537486553,
0.049303531646728516,
0.040400583297014236,
0.03673436865210533,
0.0033214467111974955,
0.015050691552460194,
-0.016843348741531372,
0.04084017127752304,
0.06328536570072174,
0.031097179278731346,
-0.03080611303448677,
0.01829361543059349,
0.043600331991910934,
-0.01637408882379532,
-0.009121639654040337,
-0.011406932957470417,
0.0058978041633963585,
-0.007223143242299557,
0.012579631991684437,
-0.02591581456363201,
-0.01889161393046379,
-0.005181158427149057,
-0.0496077835559845,
-0.003396478248760104,
-0.05299503356218338,
0.005138229113072157,
-0.03719083219766617,
-0.02052805759012699,
0.04781705141067505,
-0.042308468371629715,
0.011612134985625744,
0.018289178609848022,
-0.023705260828137398,
0.0031208419241011143,
0.03372952714562416,
0.032261356711387634,
0.0013933208538219333,
0.05576213076710701,
0.041811373084783554,
0.0003375167434569448,
0.03574563190340996,
0.0024057654663920403,
0.0008790779393166304,
-0.05680059641599655,
0.020062707364559174,
0.01804955117404461,
-0.013327925466001034,
-0.039351437240839005,
-0.003232947550714016,
-0.03815429285168648,
0.03370795398950577,
-0.032324500381946564,
0.00479769054800272,
0.050282664597034454,
-0.04866406321525574,
-0.005248159170150757,
-0.03404255211353302,
0.035817284137010574,
-0.023673495277762413,
-0.0022058768663555384,
0.02593930996954441,
-0.024253064766526222,
-0.031101573258638382,
0.06740589439868927,
-0.01812499202787876,

0.022601503878831863,
0.015241713263094425,
0.05846760794520378,
-0.021876053884625435,
0.043459728360176086,
0.06106221675872803,
-0.010913856327533722,
0.01049842219799757,
0.0010406096698716283,
-0.03195137158036232,
-0.02710483781993389,
-0.00592478783801198,
-0.05254151672124863,
-0.0028600762598216534,
0.0014663580805063248,
0.008076347410678864,
0.0057233842089772224,
0.04992511123418808,
0.020318666473031044,
-0.04395770654082298,
0.012543108314275742,
0.03251813352108002,
-0.024990687146782875,
-0.000649411987978965,
0.003185351612046361,
0.05013545975089073,
-0.0022207247093319893,
-0.022771000862121582,
-0.0050545986741781235,
-0.05946974456310272,
-0.0038308529183268547,
-0.06654298305511475,
0.0418134480714798,
-0.004801803734153509,
-0.029514247551560402,
-0.04386197403073311,
-0.049168772995471954,
-0.0201284047216177,
0.01556877139955759,
-0.027850911021232605,
-0.024432361125946045,
0.04996798560023308,
0.05487367510795593,
0.007770303636789322,
0.020771479234099388,
-0.019606124609708786,
-0.013351675122976303,
0.030479855835437775,
0.0651073306798935,
0.01271777506917715,
0.05332284793257713,
0.05055456981062889,
-0.003854509675875306,
0.022267375141382217,
-0.048184480518102646,
0.012688757851719856,
-0.025737261399626732,
-0.029389861971139908,
-0.03675638884305954,
0.02813337743282318,
-0.03799116611480713,
-0.06543595343828201,
0.016400178894400597,
0.04346107318997383,
-0.0024894848465919495,
-0.03191111236810684,
-0.06447511166334152,
-0.005914723034948111,
-0.06235651299357414,
0.02022537775337696,
-0.06231554225087166,
0.008296003565192223,

0.011064848862588406,
-0.015904244035482407,
0.013317807577550411,
-0.07357223331928253,
0.13766056299209595,
0.0648920014500618,
0.0316673181951046,
0.004913331475108862,
0.027584131807088852,
0.03950321301817894,
0.029807377606630325,
-0.0230970811098814,
-0.013439335860311985,
-0.03706496208906174,
0.043436113744974136,
0.004910472314804792,
0.023366492241621017,
0.009511450305581093,
0.026711352169513702,
0.06987334042787552,
-0.04089134559035301,
-0.008910792879760265,
0.01629733294248581,
-0.05665846914052963,
-0.056656286120414734,
0.032408419996500015,
0.014975384809076786,
0.033954206854104996,
-0.0028153988532721996,
0.018437694758176804,
0.03579263389110565,
-0.04541632533073425,
-0.018786178901791573,
-0.017928844317793846,
0.021638330072164536,
-0.03204715624451637,
0.06660794466733932,
-0.03573809564113617,
-0.04322410002350807,
0.03505402058362961,
-0.039481598883867264,
-0.028406716883182526,
0.017145806923508644,
0.004230057820677757,
-0.02357427030801773,
0.009404762648046017,
0.018138285726308823,
-0.03879043832421303,
0.025511058047413826,
0.01400088332593441,
-0.030397431924939156,
0.014513217844069004,
0.0289479810744524,
-0.047237880527973175,
0.036724600940942764,
-0.03396030515432358,
0.018058769404888153,
-0.03165655583143234,
-0.04131685569882393,
0.015903932973742485,
0.017299076542258263,
-0.02911374531686306,
-0.009772587567567825,
-0.006180111318826675,
0.01329275593161583,
0.010824867524206638,
-0.005506141576915979,
-0.0022321012802422047,
-0.058725617825984955,
0.013104230165481567,
0.00833550002425909,
0.013880939222872257,

```
 -0.002985205966979265,
 -0.00845495704561472,
 -0.014790680259466171,
 ...],
[-0.017914261668920517,
 0.00304871192201972,
 0.006507609970867634,
 0.019544662907719612,
 -0.020989812910556793,
 -0.04169988632202148 4,
 0.007329227402806282,
 0.031477030366659164,
 0.005683635827153921,
 0.023424800485372543,
 0.041020240634679794,
 0.006950050126761198,
 0.034151341766119,
 -0.027339722961187363,
 -0.010062103159725666,
 0.005643375217914581,
 -0.0024361193645745516,
 -0.046067994087934494,
 0.0023712667170912027,
 0.02469686232507229,
 0.0015312108444049954,
 0.0030021334532648325,
 -0.08710447698831558,
 -0.019362283870577812,
 -0.015620095655322075,
 0.023459166288375854,
 0.03036116436123848,
 -0.012919639237225056,
 0.05649659410119057,
 0.04108012840151787,
 -0.02338954247534275,
 -0.007986357435584068,
 -0.005621970631182194,
 -0.04613392427563667,
 -0.0019731016363948584,
 -0.0062515526078641415,
 0.06557554751634598,
 -0.021752143278717995,
 -0.024240735918283463,
 -0.04556846246123314,
 -0.005010239779949188,
 -0.005858223885297775,
 0.04692354053258896,
 -0.03626240789890289,
 -0.023889735341072083,
 -0.02334217168390751,
 -0.02656758762896061,
 -0.0027194328140467405,
 0.014004017226397991,
 -0.018242843449115753,
 0.03282788023352623,
 0.006183944642543793,
 0.03615782782435417,
 -0.005511885974556208,
 0.006025954615324736,
 0.02549619972705841,
 0.023681972175836563,
 -0.03201200067996979,
 -0.054847653955221176,
 0.04686058685183525,
 0.005378264002501965,
 -0.010251943953335285,
 0.03427581116557121,
 -0.05347805097699165,
 0.03332686796784401,
 0.01384037546813488,
 0.02082720585167408,
 0.0037972067948430777,
```

-0.00424007885158062,
-0.034828074276447296,
-0.030420197173953056,
0.003932840656489134,
-0.03736238181591034,
-0.008273404091596603,
0.023209786042571068,
-0.00386534514836967,
0.0033510702196508646,
-0.0010338358115404844,
0.007772072218358517,
0.02684636041522026,
0.016997138038277626,
0.05583178251981735,
0.034429971128702164,
0.03404710814356804,
-0.0458589643239975,
-0.009453828446567059,
0.006647884380072355,
0.012889825738966465,
0.01852736435830593,
0.005158138461410999,
-0.007800152990967035,
0.04743959382176399,
-0.026355845853686333,
0.0010960274375975132,
0.04233444109559059,
0.03346965089440346,
-0.05253659933805466,
0.04421866685152054,
-0.03363523259758949,
-0.00693622836843133,
0.047799140214920044,
0.028268324211239815,
0.010310882702469826,
0.060346879065036774,
-0.044188883155584335,
0.010727807879447937,
0.013843314722180367,
-0.017543276771903038,
-0.003487106878310442,
-0.008906777016818523,
0.012410136871039867,
-0.03514539822936058,
0.02076079323887825,
-0.00801443587988615,
0.013347771018743515,
0.040264759212732315,
-0.0013449209509417415,
0.023569410666823387,
-0.019289419054985046,
-0.023633945733308792,
0.00888710655272007,
0.020827915519475937,
0.056396663188934326,
-0.02635318413376808,
0.01813187077641487,
-0.04588325321674347,
-0.016065062955021858,
0.037304770201444626,
-0.020110592246055603,
-0.012341700494289398,
0.0036583170294761658,
-0.030459992587566376,
-0.026230880990624428,
0.008543359115719795,
-0.021253403276205063,
0.011055677197873592,
-0.008997385390102863,
0.008990862406790257,
0.04525196552276611,
-0.0754673108458519,

0.02776999957859516,
0.05697902292013168,
-0.011932807974517345,
0.09457311779260635,
-0.031055454164743423,
0.03938145563006401,
0.014308806508779526,
-0.005713818594813347,
-0.04016181454062462,
0.011517181992530823,
-0.0297502800822258,
-0.015172549523413181,
0.005877826828509569,
0.026842422783374786,
0.017834153026342392,
0.009572119452059269,
-0.02030988037586212,
-0.007843186147511005,
0.02881968766450882,
0.0431872233748436,
-0.0003308426821604371,
-0.006959668826311827,
-0.006399684585630894,
0.038913607597351074,
-0.0009887145133689046,
0.06662430614233017,
-0.020533209666609764,
-0.019286563619971275,
-0.006699288729578257,
-0.04094313085079193,
0.039918895810842514,
-0.012995222583413124,
-0.013689303770661354,
0.021061887964606285,
0.03038196824491024,
0.012554364278912544,
0.03346347436308861,
-0.0008418228244408965,
0.05377287417650223,
0.04676776006817818,
-0.0072335535660386086,
0.02782675065100193,
0.024647962301969528,
0.0859767273068428,
-0.015117555856704712,
-0.013684188015758991,
-0.015929749235510826,
-0.03463093563914299,
-0.05379323661327362,
-0.013423512689769268,
-0.0039920988492667675,
0.031762659549713135,
-0.013439697213470936,
0.01682845503091812,
-0.0013219279935583472,
0.00629634503275156,
-0.02751769870519638,
-0.008465331047773361,
-0.01844962127506733,
-0.07183581590652466,
-0.012001380324363708,
0.0363936610519886,
-0.032664038240909576,
0.035423118621110916,
0.005803989712148905,
-0.016365613788366318,
0.002480384660884738,
0.07806365191936493,
-0.02391202561557293,
-0.018740296363830566,
0.056445397436618805,
0.0012721286620944738,

-0.04191784933209419,
-0.02565430849790573,
-0.01435265876352787,
-0.004482163116335869,
-0.017384132370352745,
0.05600491166114807,
-0.03283475711941719,
-0.029767274856567383,
0.025108974426984787,
0.018379991874098778,
0.03117060661315918,
0.023189008235931396,
-0.022748615592718124,
0.007539950776845217,
0.020586544647812843,
0.024779533967375755,
-0.026531187817454338,
-0.012315343134105206,
-0.011001559905707836,
0.0395776703953743,
0.01113832462579012,
0.07021842151880264,
0.035432640463113785,
0.029091814532876015,
0.05081140995025635,
0.038390208035707474,
-0.018353048712015152,
0.010319305583834648,
-0.031434688717126846,
0.01717357151210308,
0.023202860727906227,
0.025845246389508247,
0.017087383195757866,
0.03888199105858803,
-0.026856621727347374,
-0.0003992761776316911,
-0.013944042846560478,
0.011866314336657524,
-0.02486376464366913,
0.03764715790748596,
0.024629881605505943,
0.039895765483379364,
-0.040343593806028366,
-0.008695229887962341,
0.05128856748342514,
0.04489883780479431,
-0.03695620596408844,
0.0014718747697770596,
-0.010265662334859371,
0.03351273760199547,
-0.014827573671936989,
0.0037557603791356087,
0.026126524433493614,
0.007479908876121044,
0.008936196565628052,
0.0069572655484080315,
-0.004520800430327654,
-0.04910369217395782 5,
-0.05236328765749931,
-0.04049495980143547,
-0.06319818645715714,
-0.036631666123867035,
-0.04530130326747894,
-0.015995066612958908,
0.04833296313881874,
-0.04272451624274254,
0.019843067973852158,
-0.003813542891293764,
0.009733827784657478,
-0.015799155458807945,
-0.012499972246587276,
0.02208753302693367,

0.020814143121242523,
0.05431470274925232,
-0.015544181689620018,
0.026555757969617844,
0.0030081155709922314,
0.025484124198555946,
-0.03299370035529137,
-0.0033858371898531914,
-0.0015315412310883403,
0.008092988282442093,
0.04479844495654106,
-0.006286058574914932,
-0.007288158405572176,
-0.020724846050143242,
-0.020468851551413536,
-0.02419385500252247,
-0.006061166524887085,
0.009439540095627308,
-0.021100036799907684,
-0.013561089523136616,
-0.05365919694304466,
0.028238371014595032,
0.018083030357956886,
-0.03038579784333706,
0.042688000947237015,
0.02023269422352314,
-0.02160995453596115,
0.016122465953230858,
0.02664795145392418,
0.0010803736513480544,
-0.04211399704217911,
0.03922225534915924,
0.04860673472285271,
0.004455172456800938,
-0.038521695882081985,
-0.028054866939783096,
-0.05737524852156639,
0.021585706621408463,
0.002307151909917593,
-0.029244957491755486,
-0.006147994659841061,
0.01421384047716856,
0.03189637511968613,
-0.08238915354013443,
0.03174296021461487,
-0.060788638888978958,
-0.06796364486217499,
-0.045465823262929916,
-0.04177284240722656,
0.02401539869606495,
0.010982207953929901,
0.012064374051988125,
0.006157605908811092,
-0.026161732152104378,
0.007473476231098175,
0.011932427994906902,
0.037855494767427444,
-0.027988579124212265,
0.008271020837128162,
0.028922362253069878,
-0.0052170599810779095,
0.029964879155158997,
0.0501517690718174,
0.00721321813762188,
-0.013732262887060642,
0.0003036967827938497,
0.006706665735691786,
-0.0007131994934752584,
0.03792450949549675,
0.03975202888250351,
0.010256902314722538,
0.05464963987469673,

-0.026340804994106293,
0.00830459501594305,
0.013479577377438545,
0.0014656621497124434,
0.028400348499417305,
0.011253241449594498,
0.008050665259361267,
-0.018933841958642006,
-0.005629838909953833,
-0.04845760017633438,
0.011965997517108917,
0.0298940297216177,
0.05142587795853615,
-0.06194167956709862,
0.07403832674026489,
0.008163606747984886,
-0.06097736582159996,
0.04407711327075958,
-0.019513331353664398,
0.00723069766536355,
0.06264545023441315,
-0.011730329133570194,
0.04671846330165863,
-0.02767462469637394,
-0.015064587816596031,
-0.047893621027469635,
0.030647626146674156,
0.0488833487033844,
0.00798844639211893,
0.036888863891363144,
-0.010166793130338192,
-0.01682460680603981,
-0.019841736182570457,
-0.03143346309661865,
-0.02063911408185959,
-0.012685807421803474,
0.009288381785154343,
-0.014213950373232365,
-0.060424309223890305,
-0.06781770288944244,
0.017078625038266182,
0.013179211877286434,
0.03748795762658119,
0.011764910072088242,
0.04756525531411171,
0.014044146053493023,
0.04960833862423897,
0.03606041893362999,
-0.03488771617412567,
-1.888576844066847e-05,
-0.040138669312000275,
0.05682002753019333,
0.025343164801597595,
-0.0021652942523360252,
-0.005605677142739296,
-0.021609067916870117,
-0.02535623125731945,
0.04662661254405975,
0.007091720122843981,
-0.007591933477669954,
-0.008274879306554794,
0.007282701320946217,
0.0013663509162142873,
0.007661245297640562,
-0.03606151044368744,
-0.02161683514714241,
-0.028862951323390007,
0.04994882643222809,
0.04378928989171982,
-0.03692840784788132,
0.010449039749801159,
-0.03225337713956833,

0.029208721593022346,
0.011555220000445843,
0.0020056446082890034,
-0.05193581059575081,
-0.027537347748875618,
-0.03277117758989334,
-0.04231000691652298,
-0.017091797664761543,
0.030154703184962273,
0.011035341769456863,
0.007363723590970039,
-0.042158227413892746,
0.019418099895119667,
0.012063435278832912,
0.012567305937409401,
-0.012138241901993752,
-0.01761360466480255,
0.0006192137370817363,
-0.005135165061801672,
0.032171979546546936,
-0.022377140820026398,
-0.04193858802318573,
0.023055555298924446,
-0.06472378969192505,
0.008622318506240845,
0.011549812741577625,
-0.04252485930919647,
-0.006474762689322233,
0.04771754518151283,
0.033782586455345154,
0.03346170112490654,
1.3744304851570632e-05,
0.01403473038226366,
0.012006786651909351,
0.008588142693042755,
-0.04881145805120468,
-0.04551799222826958,
0.049570366740226746,
-0.002306038746610284,
-0.03321518748998642,
0.04059775173664093,
0.04881759360432625,
-0.014340993948280811,
0.014783844351768494,
0.01321497093886137,
-0.031192218884825706,
0.03384074196219444,
-0.062335871160030365,
0.023008445277810097,
-0.02075658179819584,
-0.030945809558033943,
-0.011385940946638584,
-0.03605683520436287,
0.04338502138853073,
-0.022411759942770004,
-0.017851170152425766,
-0.014493344351649284,
-0.06123151630163193,
-0.012533972971141338,
0.003156151156872511,
-0.02314917929470539,
0.0364871546626091,
-0.005225738976150751,
-0.007192740682512522,
-0.02413523569703102,
-0.02101905643939972,
-0.016299840062856674,
-0.007704850751906633,
-0.03256744518876076,
0.009505124762654305,
0.005887425038963556,
0.022566357627511024,

-0.01014514360576868,
0.0019205498974770308,
-0.0448288694024086,
0.03280662000179291,
0.001604246674105525,
0.005350739695131779,
-0.04766170307993889,
0.016209792345762253,
-0.009942601434886456,
-0.007568604778498411,
-0.038959912955760956,
0.013559980317950249,
-0.03389827162027359,
0.013480331748723984,
0.009463200345635414,
0.021137120202183723,
0.02836349792778492,
0.001272119814530015,
-0.03835471719503403,
0.04215202480554581,
0.03935258090496063,
-0.047304682433605194,
-0.03253700211644173,
0.006080201826989651,
0.005398996639996767,
0.030568599700927734,
0.034718405455350876,
-0.016007931903004646,
-0.0543600432574749,
-0.055149439722299576,
-0.006646972149610519,
-0.052978549152612686,
-0.015926247462630272,
-0.06006723269820213,
-0.029997356235980988,
-0.040872130542993546,
0.045750685036182404,
0.014848655089735985,
-0.017325641587376595,
0.016355857253074646,
-0.040224604308605194,
-0.0005063688731752336,
-0.0202487800270319,
-0.014700442552566528,
-0.015299816615879536,
-0.016094248741865158,
-0.004398390185087919,
0.059960540384054184,
0.000354461109964177,
0.019004633650183678,
-0.0004224211152177304,
0.004421890247613192,
0.026386022567749023,
0.015814878046512604,
-0.02596943825483322,
-0.023569729179143906,
0.02018645592033863,
0.01619723066687584,
-0.026995830237865448,
0.0197391789406538,
-0.036026496440172195,
0.08667510747909546,
-0.03895043581724167,
-0.003150335280224681,
-0.0006268357974477112,
-0.007358158007264137,
0.004954582545906305,
0.0010175466304644942,
0.039932116866111755,
-0.04068211466607399,
0.011888531036674976,
-0.02187281660735607,

0.028120122849941254,
0.027761688455939293,
-0.002485380740836262 7,
-0.0069459727965295315,
-0.03556620702147484,
-0.045845936983823776,
-0.07437250018119812,
0.025691375136375427,
-0.020275864750146866,
-0.038272321224212646,
-0.027648787945508957,
-0.010340142995119095,
0.04396086186170578,
-0.021332237869501114,
0.06734032928943634,
0.07090521603822708,
0.021472468972206116,
-0.01219962164759636,
-0.033079277724027634,
0.029865825548768044,
0.050084441900253296,
-0.05018686130642891,
-0.008996373042464256,
-0.012472518719732761,
-0.0200300645083189,
0.001344028627499938,
-0.023250041529536247,
-0.03847624361515045,
-0.06462156027555466,
-0.012182907201349735,
0.054789949208498,
-0.01676413230597973,
0.04683743044734001,
0.0017094470094889402,
-0.03814107924699783,
-0.04096877574920654,
0.04724431410431862,
0.006903103087097406,
0.020426832139492035,
0.05174718797206879,
-0.033540137112140656,
-0.0030849275644868612,
0.003771175630390644,
0.0016039644833654165,
0.022373339161276817,
-0.05426246300339699,
0.06756619364023209,
-0.019821420311927795,
-0.04241780936717987,
0.007301196455955505,
0.05942472070455551,
-0.060298822820186615,
-0.04865741357207298,
-0.02588616870343685,
0.008247070014476776,
0.017711637541651726,
-0.02864292822778225,
0.01802721992135048,
-0.018040653318166733,
-0.018790699541568756,
0.004661642014980316,
0.03619244322180748,
0.015899736434221268,
0.04498397931456566,
-0.00029516275390051305,
0.034542620182037354,
-0.036387961357831955,
0.016038814559578896,
0.027433807030320168,
-0.00841466523706913,
-0.015346063300967216,
-0.05856669694185257,

0.007415336091071367,
-0.015029856003820896,
0.00506911426782608,
0.043886493891477585,
-0.012395059689879417,
0.020645104348659515,
0.01770724542438984,
-0.0005400059162639081,
0.05124860629439354,
0.017856866121292114,
0.011706111021339893,
-0.006361138075590134,
-0.02412148378789425,
-0.04786495119333267,
-0.025074811652302742,
0.02363721653819084,
-0.026312746107578278,
0.01683548279106617,
-0.053629230707883835,
0.005438227206468582,
0.020332688465714455,
0.013943626545369625,
-0.0018061015289276838,
-0.056627802550792694,
-0.044084884226322174,
-0.03754692152142525,
-0.029501279816031456,
-0.014161752536892891,
0.010737446136772633,
-0.02090030163526535,
0.024747107177972794,
0.021723831072449684,
-0.03605707362294197,
-0.0035073941107839346,
0.017388874664902687,
-0.03468378260731697,
0.03854851797223091,
-0.03745099529623985,
0.035325996577739716,
-0.05381456017494202,
-0.03922128677368164,
-0.04169561341404915,
-0.006623526103794575,
-0.006305626127868891,
0.008522234857082367,
0.006312082521617413,
0.022120054811239243,
0.011374300345778465,
0.020698081701993942,
0.010538767091929913,
0.0034768034238368273,
-0.02409934625029564,
-0.03259539604187012,
-0.0027123659383505583,
0.03531107306480408,
-0.004582555498927832,
0.022206898778676987,
0.01334422081708908,
-0.014683335088193417,
0.04448984935879707,
-0.014210709370672703,
-0.033326227217912674,
-0.045081838965415955,
0.02542310394346714,
0.015418151393532753,
0.01785695180296898,
-0.01887572929263115,
-0.024188106879591942,
-0.0244063138961792,
-0.05529989302158356,
0.01809380203485489,
0.01902196556329727,

0.033365219831466675,
0.00864905584603548,
-0.012560417875647545,
-0.0015228373231366277,
0.04236896336078644,
-0.001818720018491149,
0.0052851843647658825,
-0.0020362497307360172,
0.027417823672294617,
0.03257900848984718,
0.003067895071581006,
0.018977822735905647,
0.008265440352261066,
0.048998892307281494,
-0.023425797000527382,
0.006014690268784761,
-0.0007991245365701616,
0.0062791393138468266,
0.018786398693919182,
-0.03043617121875286,
-0.025561237707734108,
-0.051115792244672775,
-0.0001585001591593027,
0.0033809335436671972,
0.057674504816532135,
-0.0004743505851365626,
-0.006230047903954983,
0.0059714894741773605,
-0.024875016883015633,
-0.05456719920039177,
0.01076346356421709,
-0.04700569808483124,
-0.05041594058275223,
0.04278821870684624,
-0.05694510415196419,
-0.01621295139193535,
-0.03496873751282692,
-0.03881807252764702,
0.006838295608758926,
0.005601443350315094,
-0.028015131130814552,
-0.038213133811950684,
0.011774818412959576,
0.0018304813420400023,
0.04871012270450592,
-0.046484459191560745,
0.0034897427540272474,
0.009881420992314816,
0.05603133887052536,
-0.06413774937391281,
-0.04087129607796669,
0.007689925841987133,
0.031939491629600525,
0.004865460097789764,
-0.029666081070899963,
-0.01159835234284401,
-0.0005865807179361582,
-0.009628059342503548,
-0.011487101204693317,
0.013632762245833874,
0.026217274367809296,
0.006773520726710558,
0.029168283566832542,
0.011627924628555775,
0.001117678708396852,
-0.045578427612781525,
0.01987822726368904,
0.02308647707104683,
-0.03444456681609154,
-0.009917085990309715,
0.055709268897771835,
0.012441837228834629,

-0.03180450573563576,
0.05989253893494606,
0.0418490469455719,
0.03762369230389595,
-0.005156244616955519,
0.01976749487221241,
-0.005286272149533033,
0.04440935328602791,
0.060523148626089096,
0.017002156004309654,
-0.02041194960474968,
0.017761075869202614,
0.03207273781299591,
-0.010809822008013725,
-0.011496719904243946,
0.004344920627772808,
0.01006230153143406,
-0.016177957877516747,
0.0239210594445467,
-0.015657994896173477,
-0.02106749452650547,
0.011109866201877594,
-0.042161546647548676,
0.004652564879506826,
-0.03554088994860649,
0.001491223112680018,
-0.03389892354607582,
-0.026683252304792404,
0.0589214526116848,
-0.03311343863606453,
-0.0036830962635576725,
0.009762012399733067,
-0.014763608574867249,
-0.007393624167889357,
0.03376411274075508,
0.03143079951405525,
0.02064036764204502,
0.03684137761592865,
0.03352590277791023,
0.006104095838963985 4,
0.028083300217986107,
0.008615816943347454,
0.009936526417732239,
-0.05589742586016655,
0.02917511574923992,
0.02620202861726284,
-0.01899959333240986,
-0.04229825362563133,
-0.009969344362616539,
-0.03444446623325348,
0.046007078140974045,
-0.03975262865424156,
-0.0009956435533240438,
0.038314640522003174,
-0.037987641990184784,
0.0024918627459555864,
-0.024704547598958015,
0.03667309507727623,
-0.020025592297315598,
-0.0004199163231533021,
0.030570074915885925,
-0.02940654568374157,
-0.02068157307803631,
0.06542303413152695,
-0.01456325501203537,
0.024932658299803734,
0.014145219698548317,
0.060192205011844635,
-0.027788406237959862,
0.05262495204806328,
0.04923805594444275,
-0.01643659546971321,

0.01380761805921793,
-0.013305594213306904,
-0.024035504087805748,
-0.041139524430036545,
-0.026414524763822556,
-0.04884824529290199,
-0.0184493288397789,
-0.005608795676380396,
0.0016003857599571347,
-0.012020536698400974,
0.0369306355714798,
0.012228217907249928,
-0.04233897477388382,
0.023119814693927765,
0.032706864178180695,
-0.023594314232468605,
0.006882196292281151,
0.00136578269302845,
0.045898228883743286,
-0.005770384334027767,
-0.020614219829440117,
-0.0013555531622841954,
-0.06522198766469955,
-0.010091133415699005,
-0.05912945047020912,
0.048351746052503586,
-0.007993977516889572,
-0.03430180251598358,
-0.03927387669682503,
-0.04874245822429657,
-0.012235647067427635,
0.005973318126052618,
-0.01652912050485611,
-0.017548855394124985,
0.051815927028656006,
0.058984607458114624,
0.016158482059836388,
0.025666985660791397,
-0.023362861946225166,
0.004865453578531742,
0.024162445217370987,
0.07491252571344376,
0.005695951171219349,
0.06425133347511292,
0.04079429805278778,
-0.019505934789776802,
0.02032923512160778,
-0.04160463064908981,
0.017095327377319336,
-0.014999252744019032,
-0.030814042314887047,
-0.043706219643354416,
0.03216080367565155,
-0.037647657096385956,
-0.07005618512630463,
-0.005246192682534456,
0.038605980575084686,
-0.0005431511672213674,
-0.016670528799295425,
-0.0715782567858696,
-0.009734678082168102,
-0.0553276352584362,
0.016863277181982994,
-0.05510079860687256,
0.008675960823893547,
0.015411518514156342,
-0.014562396332621574,
0.02016495354473591,
-0.06837598979473114,
0.15704329311847687,
0.07474680989980698,
0.04023011401295662,

```
0.0025917384773492813,
0.020019184798002243,
0.036506544798612595,
0.028911560773849487,
-0.011259187944233418,
-0.007455748040229082,
-0.0452619306743145,
0.04291554167866707,
0.005930607672780752,
0.026562901213765144,
0.00633339723572135,
0.01730709709227085,
0.06864196062088013,
-0.0557808093726635,
-0.021153215318918228,
0.01303285825997591,
-0.06339044123888016,
-0.050293076783418655,
0.03506246209144592,
0.0008694143034517765,
0.027818942442536354,
-0.009973193518817425,
0.023770928382873535,
0.02690751664340496,
-0.04354410618543625,
-0.011127310805022717,
-0.01680518314242363,
0.01999378390610218,
-0.040259342640638 35,
0.05425499752163887,
-0.0452224425971508,
-0.031164200976490974,
0.03912131115794182,
-0.028186267241835594,
-0.022074155509471893,
0.011169012635946274,
0.006523371208459139,
-0.023059342056512833,
0.0031686776783317327,
0.003101367736235261,
-0.04496088624000549,
0.015044784173369408,
0.027157824486494064,
-0.02592008374631405,
0.015029286034405231,
0.03744533285498619,
-0.051962710916996,
0.03291945904493332,
-0.03948870301246643,
0.022560590878129005,
-0.0341622456908226,
-0.04234945401549339,
0.008062916807830334,
0.01978762447834015,
-0.03258198872208595,
-0.011498958803713322,
-0.007651034742593765,
0.016515394672751427,
0.0039027261082082987,
-0.0064824833534657955,
-0.0024058741983026266,
-0.06411287188529968,
0.0187620222256851196,
0.0025729082990437746,
0.014636128209531307,
-0.021104471758008003,
-0.02224799431860447,
-0.018325969576835632,
...])
```

Vector Database

Vector Database

In [37]:

```
out_dir = 'medical_db'

if not os.path.exists(out_dir):
  os.makedirs(out_dir)
```

In [38]:

```
vectorstore = Chroma.from_documents(
    document_chunks,
    embedding_model,
    persist_directory=out_dir
)
```

```
ERROR:chromadb.telemetry.product.posthog:Failed to send telemetry event ClientStartEvent:
capture() takes 1 positional argument but 3 were given
ERROR:chromadb.telemetry.product.posthog:Failed to send telemetry event ClientCreateColle
ctionEvent: capture() takes 1 positional argument but 3 were given
```

In [39]:

```
vectorstore = Chroma(persist_directory=out_dir,embedding_function=embedding_model)
```

```
ERROR:chromadb.telemetry.product.posthog:Failed to send telemetry event ClientStartEvent:
capture() takes 1 positional argument but 3 were given
ERROR:chromadb.telemetry.product.posthog:Failed to send telemetry event ClientCreateColle
ctionEvent: capture() takes 1 positional argument but 3 were given
```

In [40]:

```
vectorstore.embeddings
```

Out[40]:

```
HuggingFaceEmbeddings(client=SentenceTransformer(
  (0): Transformer({'max_seq_length': 512, 'do_lower_case': False}) with Transformer mode
l: BertModel
  (1): Pooling({'word_embedding_dimension': 1024, 'pooling_mode_cls_token': False, 'pooli
ng_mode_mean_tokens': True, 'pooling_mode_max_tokens': False, 'pooling_mode_mean_sqrt_len
_tokens': False, 'pooling_mode_weightedmean_tokens': False, 'pooling_mode_lasttoken': Fal
se})
  (2): Normalize()
), model_name='thenlper/gte-large', cache_folder=None, model_kwargs={}, encode_kwargs={},
multi_process=False)
```

Retriever

In [42]:

```
retriever = vectorstore.as_retriever(
    search_type='similarity',
    search_kwargs={'k': 3}
)
```

System and User Prompt Template

In [43]:

```
qna_system_message = """
You are an assistant whose work is to give answers to questions with repect to a context.
User input will have the context required by you to answer user questions.

This context will begin with the token: ###Context.
The context contains references to specific portions of a document relevant to the user q
uery.
```

```
User questions will begin with the token: ###Question.

Strictly answer only using the information provided in the ###Context.
Do not mention anything about the information in ###Context or the question in ###Questio
n in your final answer.

If the answer to ###Question cannot be derived from the ###Context, just respond by sayin
g "I don't know".

Remember that the answer to ###Question might not always be directly present in the infor
mation provided in the ###Context.
the answer can be indirectly derived from the information in ###Context.

"""
```

In [44]:

```
qna_user_message_template = """
Conider the following ###Context and ###Question
###Context
{context}

###Question
{question}
"""
```

Response Function

In [67]:

```
def generate_rag_response(user_input,k=3,max_tokens=512,temperature=0,top_p=0.95,top_k=5
0):
    global qna_system_message,qna_user_message_template
    # Retrieve relevant document chunks
    relevant_document_chunks = retriever.get_relevant_documents(query=user_input,k=k)
    context_list = [d.page_content for d in relevant_document_chunks]

    # Combine document chunks into a single context
    context_for_query = ". ".join(context_list)

    user_message = qna_user_message_template.replace('{context}', context_for_query)
    user_message = user_message.replace('{question}', user_input)

   # prompt = qna_system_message + '\n' + user_message
   # Combine user_prompt and system_message to create the prompt
    prompt = f"""[INST]{qna_system_message}\n
                {'user'}: {qna_user_message_template.format(context=context_for_query, q
uestion=user_input)}
                [/INST]"""

    # Generate the response
    try:
        response = llm(
                  prompt=prompt,
                  max_tokens=max_tokens,
                  temperature=temperature,
                  top_p=top_p,
                  top_k=top_k
                  )

        # Extract and print the model's response
        response = response['choices'][0]['text'].strip()
    except Exception as e:
        response = f'Sorry, I encountered the following error: \n {e}'

    return response
```

In [45]:

Question Answering using RAG

Query 1: What is the protocol for managing sepsis in a critical care unit?

In [68]:

```
user_input = "What is the protocol for managing sepsis in a critical care unit?"
generate_rag_response(user_input,top_k=50)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =      96.01 ms /   171 runs   (    0.56 ms per tok
en,  1780.99 tokens per second)
llama_print_timings: prompt eval time =       0.00 ms /     1 tokens (    0.00 ms per tok
en,      inf tokens per second)
llama_print_timings:        eval time =   25322.80 ms /   171 runs   (  148.09 ms per tok
en,     6.75 tokens per second)
llama_print_timings:       total time =   26073.72 ms /   172 tokens
```

Out[68]:

```
'Based on the information provided in the ###Context, the protocol for managing sepsis in
a critical care unit includes the following elements:\n\n1. Fluid resuscitation with 0.9%
normal saline\n2. Oxygen therapy\n3. Broad-spectrum antibiotics, modified by culture resu
lts\n4. Drainage of abscesses and excision of necrotic tissue\n5. Normalization of blood
glucose levels\n6. Replacement-dose corticosteroids\n\nIt is important to note that the s
pecific protocol for managing sepsis may vary depending on the individual patient and the
specific needs of the critical care unit. However, the elements listed above are generall
y included in the management of sepsis in a critical care setting.'
```

Comments: As we can see the answers are specific to the document and dont have the LLM involved anymore. We can see that the answers have been taken from our document (vector DB).

Query 2: What are the common symptoms for appendicitis, and can it be cured via medicine? If not, what surgical procedure should be followed to treat it?

In [87]:

```
user_input = "What are the common symptoms for appendicitis, and can it be cured via medi
cine? If not, what surgical procedure should be followed to treat it?"
generate_rag_response(user_input,top_k=50)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =     182.76 ms /   322 runs   (    0.57 ms per tok
en,  1761.87 tokens per second)
llama_print_timings: prompt eval time =    3472.22 ms /  1168 tokens (    2.97 ms per tok
en,   336.38 tokens per second)
llama_print_timings:        eval time =   47380.91 ms /   321 runs   (  147.60 ms per tok
en,     6.77 tokens per second)
llama_print_timings:       total time =   52385.82 ms /  1489 tokens
```

Out[87]:

```
"Sure, I'd be happy to help! Here's my answer based on the provided context:\n\n###Questi
on: What are the common symptoms for appendicitis, and can it be cured via medicine? If n
ot, what surgical procedure should be followed to treat it?\n\n###Context:\n\nThe provide
d context mentions that appendicitis is acute inflammation of the vermiform appendix, typ
ically resulting in abdominal pain, anorexia, and abdominal tenderness. The diagnosis is
clinical, often supplemented by CT or ultrasound. Treatment is surgical removal.\n\nAnswe
r:\n\nThe common symptoms for appendicitis include abdominal pain, anorexia, and abdomina
l tenderness. However, it cannot be cured via medicine. Surgical procedure should be foll
owed to treat it, specifically open or laparoscopic appendectomy. If the appendix is perf
orated, antibiotics should be continued until the patient's temperature and WBC count hav
e normalized or continued for a fixed course, according to the surgeon's preference. If s
```

```
urgery is impossible, antibiotics may be given, although they are not curative and only i
mprove survival rates. In cases where a large inflammatory mass is found involving the ap
pendix, terminal ileum, and cecum, resection of the entire mass and ileocolostomy are pre
ferable."
```

Comments: The RAG pattern has helped us with more specific answers this time around, which are not observed in the LLM or with prompt engineering. The medical manual has provided very specific information. For e.g., the answer is that appendicites CANNOT be cured via medicine and surgery is the best option. This is not what the LLM gave us.

Query 3: What are the effective treatments or solutions for addressing sudden patchy hair loss, commonly seen as localized bald spots on the scalp, and what could be the possible causes behind it?

In [88]:

```
user_input = "What are the effective treatments or solutions for addressing sudden patchy
hair loss, commonly seen as localized bald spots on the scalp, and what could be the poss
ible causes behind it?"
generate_rag_response(user_input,top_k=50)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =     287.49 ms /   512 runs   (    0.56 ms per tok
en,  1780.96 tokens per second)
llama_print_timings: prompt eval time =    2564.32 ms /   890 tokens (    2.88 ms per tok
en,   347.07 tokens per second)
llama_print_timings:        eval time =   75431.90 ms /   511 runs   (  147.62 ms per tok
en,     6.77 tokens per second)
llama_print_timings:       total time =   80505.91 ms /  1401 tokens
```

Out[88]:

```
"Based on the provided context, there are several effective treatments for sudden patchy
hair loss, commonly seen as localized bald spots on the scalp. These include:\n\n1. Topic
al corticosteroids: These medications can help reduce inflammation and slow down hair los
s. They can be applied directly to the affected area and are available in different stren
gths.\n2. Minoxidil: This medication is applied topically and can help stimulate hair gro
wth and slow down hair loss. It is available over-the-counter and can be used in combinat
ion with corticosteroids.\n3. Anthralin: This medication is applied topically and can hel
p reduce inflammation and promote hair growth. It is often used in combination with corti
costeroids.\n4. Immunotherapy: This treatment involves using small amounts of a substance
that causes an allergic reaction to stimulate the immune system and promote hair growth.
It can be applied topically or taken orally.\n5. Oral antifungals: These medications can
be used to treat fungal infections that may be contributing to hair loss.\n\nPossible cau
ses of sudden patchy hair loss include:\n\n1. Alopecia areata: This is a autoimmune condi
tion that causes hair loss in patches. It can affect anyone, but it is more common in you
ng adults and children.\n2. Tinea capitis (scalp ringworm): This is a fungal infection th
at can cause hair loss and scaling on the scalp. It is more common in children and can be
contagious.\n3. Traction alopecia: This is hair loss caused by constant pulling or tensio
n on the hair, such as from tight hairstyles or hair extensions.\n4. Trichotillomania: Th
is is a condition where people compulsively pull out their own hair, leading to hair loss
.\n5. Scarring alopecia: This is hair loss caused by inflammation and scarring on the sca
lp, often seen in conditions such as central centrifugal scarring alopecia, dissecting ce
llulitis of the scalp, and acne keloidalis nuchae.\n\nIt's important to note that hair lo
ss can be temporary or permanent, depending on the underlying cause. A thorough evaluatio
n and diagnosis by a health"
```

Comments: The RAG model has given us a better answer here, rather than the generic answer given by the LLM. The answer for a patchy hair loss is to use steroids and/or immunotherapy that stimulates hair growth.

Query 4: What treatments are recommended for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function?

In [89]:

```
user_input = "What treatments are recommended for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function?"
generate_rag_response(user_input,top_k=50)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =     143.53 ms /   261 runs   (    0.55 ms per token,  1818.40 tokens per second)
llama_print_timings: prompt eval time =    2687.29 ms /   918 tokens (    2.93 ms per token,   341.61 tokens per second)
llama_print_timings:        eval time =   38299.68 ms /   260 runs   (  147.31 ms per token,     6.79 tokens per second)
llama_print_timings:       total time =   42178.73 ms /  1178 tokens
```

Out[89]:

"Based on the provided context, I can answer your question as follows:\n\nFor a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function, the recommended treatments include:\n\n1. Supportive care to prevent systemic complications due to immobilization, such as pneumonia, UTI, and thromboembolic disease.\n2. Good nutrition to support recovery.\n3. Prevention of pressure ulcers.\n4. Rehabilitation therapy to address cognitive dysfunction, muscle weakness, spasticity, incoordination, and ataxia.\n5. Early intervention by rehabilitation specialists to maximize functional recovery.\n6. Cognitive therapy for patients with severe cognitive dysfunction, which may be begun immediately after injury and continued for months or years.\n\nIt is important to note that the effectiveness of these treatments may vary depending on the level and extent of the injury, as well as the patient's age and overall health. Additionally, improvement may occur late, and some patients may remain severely disabled or die within a few years after the initial injury."

Comments: The answer we can see is more specific to the injury. It covers patients with severe cognitive dysfunction, diabled patients etc. Thus using the LLM the answers are more generic and the answers in the medical journal is much better.

Query 5: What are the necessary precautions and treatment steps for a person who has fractured their leg during a hiking trip, and what should be considered for their care and recovery?

In [90]:

```
user_input = "What are the necessary precautions and treatment steps for a person who has fractured their leg during a hiking trip, and what should be considered for their care and recovery?"
generate_rag_response(user_input,top_k=50)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =     251.59 ms /   440 runs   (    0.57 ms per token,  1748.87 tokens per second)
llama_print_timings: prompt eval time =    2691.68 ms /   923 tokens (    2.92 ms per token,   342.91 tokens per second)
llama_print_timings:        eval time =   65398.12 ms /   439 runs   (  148.97 ms per token,     6.71 tokens per second)
llama_print_timings:       total time =   70268.97 ms /  1362 tokens
```

Out[90]:

'Based on the provided context, here is the answer to the question:\n\nFor a person who has fractured their leg during a hiking trip, necessary precautions and treatment steps include:\n\n1. Immobilization: The leg should be immobilized using a splint or cast to prevent further movement and promote healing.\n2. Pain management: Analgesics should be administered to manage pain and discomfort.\n3. Elevation: The affected leg should be elevated above heart level for the first 48 hours to reduce swelling and promote blood flow.\n4. Wound care: The wound should be cleaned and dressed with sterile compresses to prevent infection and promote healing.\n5. Monitoring for infection: The patient should be monitored for signs of infection, such as fever or an odor emanating from the cast.\n6. Early mobilization: Resumption of active motion within the first few days or weeks can minimize contractures and muscle atrophy, thus accelerating functional recovery.\n\nConsiderations for

```
care and recovery include:\n\n1. Prolonged immobilization: Prolonged immobilization (more
than 3-4 weeks) can cause stiffness, contractures, and muscle atrophy, particularly in th
e elderly.\n2. Good hygiene: Good hygiene is important to prevent infection and promote h
ealing.\n3. Monitoring for complications: The patient should be monitored for complicatio
ns such as deep venous thrombosis and urinary tract infections.\n4. Early medical attenti
on: If an odor emanates from within the cast or if a fever develops, medical attention sh
ould be sought immediately.\n\nIt is important to note that the answer to this question c
annot be directly derived from the provided context, and therefore, I cannot provide a di
rect answer. However, the necessary precautions and treatment steps can be inferred from
the information provided in the context.'
```

Comments: The RAG model usage helps with more specific answers, like for e.g., how pain management is done. Also, for the last part, when an odor emanates instead of halucinating, the RAG model clearly says that it does not have information.

Fine-tuning

Query 1: What is the protocol for managing sepsis in a critical care unit?

In [91]:

```
user_input = "What is the protocol for managing sepsis in a critical care unit?"
generate_rag_response(user_input,temperature=0.5)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =     113.04 ms /   201 runs   (    0.56 ms per tok
en,  1778.18 tokens per second)
llama_print_timings: prompt eval time =    2867.77 ms /   909 tokens (    3.15 ms per tok
en,   316.97 tokens per second)
llama_print_timings:        eval time =   29285.70 ms /   200 runs   (  146.43 ms per tok
en,     6.83 tokens per second)
llama_print_timings:       total time =   33073.22 ms /  1109 tokens
```

Out[91]:

```
'Based on the provided context, I can answer your question as follows:\n\nThe protocol fo
r managing sepsis in a critical care unit includes the following components:\n\n1. Fluid
resuscitation with 0.9% normal saline\n2. Oxygen therapy\n3. Broad-spectrum antibiotics,
modified by culture results\n4. Drainage of abscesses and excision of necrotic tissue\n5.
Normalization of blood glucose levels\n6. Replacement-dose corticosteroids.\n\nIt is impo
rtant to note that early aggressive therapy, within 6 hours of suspected diagnosis, is cr
ucial for improving outcomes in patients with septic shock. Additionally, monitoring for
signs of sepsis, such as abnormal vital signs, altered mental status, and the presence of
infection, is essential for timely intervention and management.'
```

Query 2: What are the common symptoms for appendicitis, and can it be cured via medicine? If not, what surgical procedure should be followed to treat it?

In [92]:

```
user_input_2 = "What are the common symptoms for appendicitis, and can it be cured via me
dicine? If not, what surgical procedure should be followed to treat it?"
generate_rag_response(user_input_2,temperature=0.5)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =     186.62 ms /   333 runs   (    0.56 ms per tok
en,  1784.39 tokens per second)
llama_print_timings: prompt eval time =    2931.64 ms /   962 tokens (    3.05 ms per tok
en,   328.14 tokens per second)
llama_print_timings:        eval time =   48968.01 ms /   332 runs   (  147.49 ms per tok
en,     6.78 tokens per second)
llama_print_timings:       total time =   53463.09 ms /  1294 tokens
```

Out[92]:

```
"Sure, I'd be happy to help! Here's my answer based on the provided context:\n\n###Questi
on: What are the common symptoms for appendicitis, and can it be cured via medicine? If n
ot, what surgical procedure should be followed to treat it?\n\n###Context:\n\nThe provide
d context mentions that appendicitis is acute inflammation of the vermiform appendix, typ
ically resulting in abdominal pain, anorexia, and abdominal tenderness. Diagnosis is clin
ical, often supplemented by CT or ultrasound. Treatment is surgical removal.\n\nAnswer:\n
\nThe common symptoms for appendicitis include abdominal pain, anorexia, and abdominal te
nderness. However, it cannot be cured via medicine, and surgical procedure is necessary t
o treat it. The surgical procedure for appendicitis is appendectomy, which can be either
open or laparoscopic. If the appendix is perforated, antibiotics should be continued unti
l the patient's temperature and WBC count have normalized or continued for a fixed course
, according to the surgeon's preference. If surgery is impossible, antibiotics can marked
ly improve the survival rate, but they are not curative. In cases where a large inflammat
ory mass is found involving the appendix, terminal ileum, and cecum, resection of the ent
ire mass and ileocolostomy are preferable."
```

Query 3: What are the effective treatments or solutions for addressing sudden patchy hair loss, commonly seen as localized bald spots on the scalp, and what could be the possible causes behind it?

In [93]:

```
user_input_3 = "What are the effective treatments or solutions for addressing sudden patc
hy hair loss, commonly seen as localized bald spots on the scalp, and what could be the p
ossible causes behind it?"
generate_rag_response(user_input_3,temperature=0.5)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =     288.41 ms /   512 runs   (    0.56 ms per tok
en,  1775.28 tokens per second)
llama_print_timings: prompt eval time =    2652.13 ms /   890 tokens (    2.98 ms per tok
en,   335.58 tokens per second)
llama_print_timings:        eval time =   75750.92 ms /   511 runs   (  148.24 ms per tok
en,     6.75 tokens per second)
llama_print_timings:       total time =   80947.92 ms /  1401 tokens
```

Out[93]:

```
'Based on the provided context, there are several effective treatments for sudden patchy
hair loss, commonly seen as localized bald spots on the scalp. These include:\n\n1. Topic
al corticosteroids: These medications can help reduce inflammation and slow down hair los
s. They can be applied directly to the affected area and can be purchased over-the-counte
r or prescribed by a dermatologist.\n2. Minoxidil: This medication is applied topically t
o the scalp and can help stimulate hair growth and slow down hair loss. It is available o
ver-the-counter and can be used in combination with corticosteroids.\n3. Anthralin: This
medication is applied topically to the scalp and can help reduce inflammation and promote
hair growth. It is often used in combination with corticosteroids.\n4. Immunotherapy: Thi
s treatment involves using small amounts of the substance that causes the allergic reacti
on to desensitize the body and promote hair growth. It can be applied topically or taken
orally.\n5. Oral antifungals: These medications can be used to treat fungal infections su
ch as tinea capitis, which can cause patchy hair loss.\n\nPossible causes of sudden patch
y hair loss include:\n\n1. Alopecia areata: This is a autoimmune condition that causes su
dden patchy hair loss. It can affect any hairy area of the body and can be treated with t
opical or oral medications.\n2. Tinea capitis: This is a fungal infection that can cause
patchy hair loss, especially in children. It can be treated with topical or oral antifung
als.\n3. Traction alopecia: This is a type of hair loss caused by physical traction or st
ress to the scalp, such as from tight hairstyles or hair extensions. Treatment involves e
liminating the physical traction or stress.\n4. Trichotillomania: This is a condition whe
re the patient has an irresistible urge to pull out their own hair, leading to patchy hai
r loss. Treatment involves behavior modification and medication.\n5. Scarring alopecia: T
his is a type of hair loss caused by inflammation and scarring of the scalp, such as from
central centrif'
```

Query 4: What treatments are recommended for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function?

In [94]:

```
user_input_4 = "What treatments are recommended for a person who has sustained a physical
```

```
injury to brain tissue, resulting in temporary or permanent impairment of brain function?
"
generate_rag_response(user_input_4,temperature=0.5)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =     145.81 ms /   255 runs   (    0.57 ms per tok
en,  1748.82 tokens per second)
llama_print_timings: prompt eval time =    2754.65 ms /   918 tokens (    3.00 ms per tok
en,   333.25 tokens per second)
llama_print_timings:        eval time =   37215.94 ms /   254 runs   (  146.52 ms per tok
en,     6.83 tokens per second)
llama_print_timings:       total time =   41136.74 ms /  1172 tokens
```

Out[94]:

"Based on the provided context, I can answer your question as follows:\n\nFor a person wh
o has sustained a physical injury to brain tissue, resulting in temporary or permanent im
pairment of brain function, the recommended treatments include:\n\n1. Supportive care to
prevent systemic complications due to immobilization, such as pneumonia, UTI, and thrombo
embolic disease.\n2. Good nutrition to support recovery.\n3. Prevention of pressure ulcer
s.\n4. Rehabilitation therapy to address cognitive dysfunction, muscle weakness, spastici
ty, incoordination, and ataxia.\n5. Cognitive therapy for patients with severe cognitive
dysfunction, which may be begun immediately after injury and continued for months or year
s.\n\nIt is important to note that early intervention by rehabilitation specialists is cr
ucial for maximal functional recovery, and treatment should be tailored to the individual
's specific abnormalities and needs. Additionally, patients with severe cognitive dysfunc
tion may require extensive cognitive therapy, and those in a persistent vegetative state
may have a poor prognosis."

Query 5: What are the necessary precautions and treatment steps for a person who has fractured their leg during a hiking trip, and what should be considered for their care and recovery?

In [95]:

```
user_input_5 = "What are the necessary precautions and treatment steps for a person who h
as fractured their leg during a hiking trip, and what should be considered for their care
and recovery?"
generate_rag_response(user_input_5,temperature=0.5)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =     259.31 ms /   454 runs   (    0.57 ms per tok
en,  1750.83 tokens per second)
llama_print_timings: prompt eval time =    2878.97 ms /   923 tokens (    3.12 ms per tok
en,   320.60 tokens per second)
llama_print_timings:        eval time =   66684.51 ms /   453 runs   (  147.21 ms per tok
en,     6.79 tokens per second)
llama_print_timings:       total time =   71747.90 ms /  1376 tokens
```

Out[95]:

'Based on the provided context, here is the answer to the question:\n\nFor a person who h
as fractured their leg during a hiking trip, necessary precautions and treatment steps in
clude:\n\n1. Immobilization: The leg should be immobilized using a splint or cast to prev
ent further damage and promote healing.\n2. Pain management: Analgesics should be adminis
tered to manage pain and discomfort.\n3. Elevation: The affected leg should be elevated a
bove heart level for the first 48 hours to reduce swelling and promote blood flow.\n4. Cl
eanliness: Wound care should be meticulous, and dressings should be applied daily to keep
the wound clean and dry.\n5. Monitoring: Patients should be monitored for signs of infect
ion, such as fever or an odor emanating from within the cast.\n6. Early mobilization: Res
umption of active motion within the first few days or weeks can minimize contractures and
muscle atrophy, thus accelerating functional recovery.\n\nConsiderations for care and rec
overy include:\n\n1. Prolonged immobilization: Prolonged immobilization (more than 3 to 4
weeks) of a joint can cause stiffness, contractures, and muscle atrophy, particularly in
the elderly.\n2. Good hygiene: Good hygiene is important to prevent infection and promote
healing.\n3. Monitoring for complications: Patients should be monitored for signs of comp
lications such as deep venous thrombosis and urinary tract infections.\n4. Early medical
attention: Medical attention should be sought at once if an odor emanates from within the
cast or if a fever develops, which may indicate infection.\n\nIt is important to note tha

```
t the answer to the question cannot be directly derived from the provided context, and th
erefore, I cannot provide a direct answer to the question. However, the necessary precaut
ions and treatment steps can be derived indirectly from the provided context.'
```

Comments: Fine tuning the LLM's using temperature, changes the answers a little and adds some randomness to it. Although, much of the answers remain the same.

Output Evaluation

Let us now use the LLM-as-a-judge method to check the quality of the RAG system on two parameters - retrieval and generation. We illustrate this evaluation based on the answeres generated to the question from the previous section.

- **We are using the same Mistral model for evaluation, so basically here the llm is rating itself on how well he has performed in the task.**

In [96]:

```
groundedness_rater_system_message  = """

You will be presented a ###Question, ###Context used by the AI system and AI generated ##
#Answer.

Your task is to judge the extent to which the ###Answer is derived from ###Context.

Rate it 1 - if The ###Answer is not derived from the ###Context at all
Rate it 2 - if The ###Answer is derived from the ###Context only to a limited extent
Rate it 3 - if The ###Answer is derived from ###Context to a good extent
Rate it 4 - if The ###Answer is derived from ###Context mostly
Rate it 5 - if The ###Answer is is derived from ###Context completely

Please note: Make sure you give a single overall rating in the range of 1 to 5 along with
an overall explanation.

"""
```

In [97]:

```
relevance_rater_system_message = """

You will be presented with a ###Question, the ###Context used by the AI system to generat
e a response, and the AI-generated ###Answer.

Your task is to judge the extent to which the ###Answer is relevant to the ###Question, c
onsidering whether it directly addresses the key aspects of the ###Question based on the
provided ###Context.

Rate the relevance as follows:
- Rate 1 - The ###Answer is not relevant to the ###Question at all.
- Rate 2 - The ###Answer is only slightly relevant to the **###Question**, missing key as
pects.
- Rate 3 - The ###Answer is moderately relevant, addressing some parts of the **###Questi
on** but leaving out important details.
- Rate 4 - The ###Answer is mostly relevant, covering key aspects but with minor gaps.
- Rate 5 - The ###Answer is fully relevant, directly answering all important aspects of t
he **###Question** with appropriate details from the **###Context**.

Note: Provide a single overall rating in the range of 1 to 5, along with a brief explanat
ion of why you assigned that score.
"""
```

In [98]:

```
user_message_template = """
###Question
```

```
{question}

###Context
{context}

###Answer
{answer}
"""
```

In [99]:

```
def generate_ground_relevance_response(user_input,k=3,max_tokens=128,temperature=0,top_p=
0.95,top_k=50):
    global qna_system_message,qna_user_message_template
    # Retrieve relevant document chunks
    relevant_document_chunks = retriever.get_relevant_documents(query=user_input,k=3)
    context_list = [d.page_content for d in relevant_document_chunks]
    context_for_query = ". ".join(context_list)

    # Combine user_prompt and system_message to create the prompt
    prompt = f"""[INST]{qna_system_message}\n
                {'user'}: {qna_user_message_template.format(context=context_for_query, q
uestion=user_input)}
                [/INST]"""

    response = llm(
            prompt=prompt,
            max_tokens=max_tokens,
            temperature=temperature,
            top_p=top_p,
            top_k=top_k,
            stop=['INST'],
            )

    answer =  response["choices"][0]["text"]

    # Combine user_prompt and system_message to create the prompt
    groundedness_prompt = f"""[INST]{groundedness_rater_system_message}\n
                {'user'}: {user_message_template.format(context=context_for_query, quest
ion=user_input, answer=answer)}
                [/INST]"""

    # Combine user_prompt and system_message to create the prompt
    relevance_prompt = f"""[INST]{relevance_rater_system_message}\n
                {'user'}: {user_message_template.format(context=context_for_query, quest
ion=user_input, answer=answer)}
                [/INST]"""

    response_1 = llm(
            prompt=groundedness_prompt,
            max_tokens=max_tokens,
            temperature=temperature,
            top_p=top_p,
            top_k=top_k,
            stop=['INST'],
            )

    response_2 = llm(
            prompt=relevance_prompt,
            max_tokens=max_tokens,
            temperature=temperature,
            top_p=top_p,
            top_k=top_k,
            stop=['INST'],
            )

    return response_1['choices'][0]['text'],response_2['choices'][0]['text']
```

Query 1: What is the protocol for managing sepsis in a critical care unit?

In [100]:

```
ground,rel = generate_ground_relevance_response(user_input="What is the protocol for mana
ging sepsis in a critical care unit?",max_tokens=200)

print(ground,end="\n\n")
print(rel)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =      95.21 ms /   171 runs   (    0.56 ms per tok
en,  1796.07 tokens per second)
llama_print_timings: prompt eval time =    2713.03 ms /   909 tokens (    2.98 ms per tok
en,   335.05 tokens per second)
llama_print_timings:        eval time =   24783.46 ms /   170 runs   (  145.79 ms per tok
en,     6.86 tokens per second)
llama_print_timings:       total time =   28258.14 ms /  1079 tokens
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =      75.27 ms /   137 runs   (    0.55 ms per tok
en,  1820.02 tokens per second)
llama_print_timings: prompt eval time =    3798.29 ms /  1263 tokens (    3.01 ms per tok
en,   332.52 tokens per second)
llama_print_timings:        eval time =   20243.96 ms /   136 runs   (  148.85 ms per tok
en,     6.72 tokens per second)
llama_print_timings:       total time =   24694.04 ms /  1399 tokens
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =      79.67 ms /   149 runs   (    0.53 ms per tok
en,  1870.29 tokens per second)
llama_print_timings: prompt eval time =    4100.04 ms /  1344 tokens (    3.05 ms per tok
en,   327.80 tokens per second)
llama_print_timings:        eval time =   22064.00 ms /   148 runs   (  149.08 ms per tok
en,     6.71 tokens per second)
llama_print_timings:       total time =   26854.01 ms /  1492 tokens
```

Based on the information provided in the context, I would rate the answer as a 5, as it is completely derived from the context. The answer provides a comprehensive list of the e lements of the protocol for managing sepsis in a critical care unit, and each element is supported by specific information from the context. The answer also acknowledges that the specific protocol may vary depending on the individual patient and the specific needs of the critical care unit, which demonstrates an understanding of the context. Overall, the answer is well-derived from the context and provides a complete and accurate summary of t he protocol for managing sepsis in a critical care unit.

Based on the provided context and answer, I would rate the relevance of the answer as a 5, fully relevant. The answer directly addresses all key aspects of the question, includi ng the elements of fluid resuscitation, oxygen therapy, broad-spectrum antibiotics, drain age of abscesses, normalization of blood glucose levels, and replacement-dose corticoster oids. The answer also acknowledges that the specific protocol for managing sepsis may var y depending on the individual patient and the specific needs of the critical care unit, w hich demonstrates an understanding of the context provided. Overall, the answer is fully relevant and directly addresses all key aspects of the question.

Query 2: What are the common symptoms for appendicitis, and can it be cured via medicine? If not, what surgical procedure should be followed to treat it?

In [101]:

```
ground,rel = generate_ground_relevance_response(user_input="What are the effective treatm
ents or solutions for addressing sudden patchy hair loss, commonly seen as localized bald
spots on the scalp, and what could be the possible causes behind it?",max_tokens=50, k=2)

print(ground,end="\n\n")
print(rel)
```

```
Llama.generate: prefix-match hit
```

```
llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =      28.95 ms /    50 runs   (    0.58 ms per tok
en,  1726.88 tokens per second)
llama_print_timings: prompt eval time =    3503.60 ms /  1096 tokens (    3.20 ms per tok
en,   312.82 tokens per second)
llama_print_timings:        eval time =    7274.11 ms /    49 runs   (  148.45 ms per tok
en,     6.74 tokens per second)
llama_print_timings:       total time =   11017.35 ms /  1145 tokens
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =      26.91 ms /    50 runs   (    0.54 ms per tok
en,  1857.91 tokens per second)
llama_print_timings: prompt eval time =    3499.19 ms /  1124 tokens (    3.11 ms per tok
en,   321.22 tokens per second)
llama_print_timings:        eval time =    7102.79 ms /    49 runs   (  144.95 ms per tok
en,     6.90 tokens per second)
llama_print_timings:       total time =   10826.78 ms /  1173 tokens
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =      26.58 ms /    50 runs   (    0.53 ms per tok
en,  1881.40 tokens per second)
llama_print_timings: prompt eval time =    3833.93 ms /  1205 tokens (    3.18 ms per tok
en,   314.30 tokens per second)
llama_print_timings:        eval time =    6825.78 ms /    49 runs   (  139.30 ms per tok
en,     7.18 tokens per second)
llama_print_timings:       total time =   10873.78 ms /  1254 tokens
```

Based on the provided context, the answer provided is rated as 4, as it is mostly deriv
ed from the context. The answer lists several effective treatments for sudden patchy hair
loss, including topical corticosteroids, which

Based on the provided context and answer, I would rate the relevance of the answer as 4
out of 5. The answer provides several effective treatments for sudden patchy hair loss, i
ncluding topical corticosteroids, which

Query 3: What are the effective treatments or solutions for addressing sudden patchy hair loss, commonly seen as localized bald spots on the scalp, and what could be the possible causes behind it?

In [102]:

```
ground,rel = generate_ground_relevance_response(user_input="What are the effective treatm
ents or solutions for addressing sudden patchy hair loss, commonly seen as localized bald
spots on the scalp, and what could be the possible causes behind it?",max_tokens=50, k=2)

print(ground,end="\n\n")
print(rel)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =      25.87 ms /    50 runs   (    0.52 ms per tok
en,  1933.11 tokens per second)
llama_print_timings: prompt eval time =    3676.33 ms /  1096 tokens (    3.35 ms per tok
en,   298.12 tokens per second)
llama_print_timings:        eval time =    6738.47 ms /    49 runs   (  137.52 ms per tok
en,     7.27 tokens per second)
llama_print_timings:       total time =   10614.42 ms /  1145 tokens
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =      27.05 ms /    50 runs   (    0.54 ms per tok
en,  1848.36 tokens per second)
llama_print_timings: prompt eval time =    3570.81 ms /  1124 tokens (    3.18 ms per tok
en,   314.77 tokens per second)
llama_print_timings:        eval time =    7087.16 ms /    49 runs   (  144.64 ms per tok
en,     6.91 tokens per second)
llama_print_timings:       total time =   10888.05 ms /  1173 tokens
Llama.generate: prefix-match hit
```

```
llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =      27.73 ms /    50 runs   (    0.55 ms per tok
en,  1802.97 tokens per second)
llama_print_timings: prompt eval time =    3736.12 ms /  1205 tokens (    3.10 ms per tok
en,   322.53 tokens per second)
llama_print_timings:        eval time =    7402.98 ms /    49 runs   (  151.08 ms per tok
en,     6.62 tokens per second)
llama_print_timings:       total time =   11377.49 ms /  1254 tokens
```

Based on the provided context, the answer provided is rated as 4, as it is mostly deriv ed from the context. The answer lists several effective treatments for sudden patchy hair loss, including topical corticosteroids, which

Based on the provided context and answer, I would rate the relevance of the answer as 4 out of 5. The answer provides several effective treatments for sudden patchy hair loss, i ncluding topical corticosteroids, which

Query 4: What treatments are recommended for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function?

In [103]:

```
ground,rel = generate_ground_relevance_response(user_input="What treatments are recommend
ed for a person who has sustained a physical injury to brain tissue, resulting in tempora
ry or permanent impairment of brain function?",max_tokens=50,k=2)

print(ground,end="\n\n")
print(rel)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =      29.39 ms /    50 runs   (    0.59 ms per tok
en,  1701.03 tokens per second)
llama_print_timings: prompt eval time =    3530.63 ms /  1124 tokens (    3.14 ms per tok
en,   318.36 tokens per second)
llama_print_timings:        eval time =    7300.66 ms /    49 runs   (  148.99 ms per tok
en,     6.71 tokens per second)
llama_print_timings:       total time =   11066.60 ms /  1173 tokens
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =      26.74 ms /    50 runs   (    0.53 ms per tok
en,  1869.86 tokens per second)
llama_print_timings: prompt eval time =    3571.72 ms /  1152 tokens (    3.10 ms per tok
en,   322.53 tokens per second)
llama_print_timings:        eval time =    7047.52 ms /    49 runs   (  143.83 ms per tok
en,     6.95 tokens per second)
llama_print_timings:       total time =   10840.77 ms /  1201 tokens
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =      26.62 ms /    50 runs   (    0.53 ms per tok
en,  1877.93 tokens per second)
llama_print_timings: prompt eval time =    3937.25 ms /  1233 tokens (    3.19 ms per tok
en,   313.16 tokens per second)
llama_print_timings:        eval time =    6806.62 ms /    49 runs   (  138.91 ms per tok
en,     7.20 tokens per second)
llama_print_timings:       total time =   10952.45 ms /  1282 tokens
```

Based on the provided context, I would rate the answer as a 4 out of 5 in terms of how well it is derived from the context. The answer provides a comprehensive list of recommen ded treatments for a person with a brain injury

Based on the provided context, I would rate the relevance of the answer as 4 out of 5. The answer provides a comprehensive list of recommended treatments for a person with a ph ysical injury to brain tissue, including supportive

Query 5: What are the necessary precautions and treatment steps for a person who has fractured their leg during a hiking trip, and what should be considered for their care and recovery?

In [104]:

```
ground,rel = generate_ground_relevance_response(user_input="What are the necessary precau
tions and treatment steps for a person who has fractured their leg during a hiking trip,
and what should be considered for their care and recovery?",max_tokens=50,k=2)

print(ground,end="\n\n")
print(rel)
```

```
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =      26.47 ms /    50 runs   (    0.53 ms per tok
en,  1889.00 tokens per second)
llama_print_timings: prompt eval time =    3798.96 ms /  1129 tokens (    3.36 ms per tok
en,   297.19 tokens per second)
llama_print_timings:        eval time =    6762.15 ms /    49 runs   (  138.00 ms per tok
en,     7.25 tokens per second)
llama_print_timings:       total time =   10767.09 ms /  1178 tokens
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =      26.47 ms /    50 runs   (    0.53 ms per tok
en,  1888.86 tokens per second)
llama_print_timings: prompt eval time =    3724.89 ms /  1157 tokens (    3.22 ms per tok
en,   310.61 tokens per second)
llama_print_timings:        eval time =    7081.90 ms /    49 runs   (  144.53 ms per tok
en,     6.92 tokens per second)
llama_print_timings:       total time =   11026.52 ms /  1206 tokens
Llama.generate: prefix-match hit

llama_print_timings:        load time =     860.55 ms
llama_print_timings:      sample time =      29.16 ms /    50 runs   (    0.58 ms per tok
en,  1714.97 tokens per second)
llama_print_timings: prompt eval time =    3798.33 ms /  1238 tokens (    3.07 ms per tok
en,   325.93 tokens per second)
llama_print_timings:        eval time =    7496.20 ms /    49 runs   (  152.98 ms per tok
en,     6.54 tokens per second)
llama_print_timings:       total time =   11533.31 ms /  1287 tokens
```

```
  Based on the provided context, I would rate the answer as a 4 out of 5 in terms of how
well it is derived from the context. The answer provides a comprehensive list of necessar
y precautions and treatment steps for a person

  Based on the provided context and answer, I would rate the relevance as 4 out of 5. The
answer directly addresses all key aspects of the question, including immobilization, woun
d care, and potential complications. The answer
```

Actionable Insights and Business Recommendations

Power Ahead

The ground ratings are 4 (on 4 out of the 5 questions and then 5 on the 1st question) which is very good and thereby using the LLM as a judge has worked very well. Thus the RAG system worked well with the answers with regards to specificity and context.

Using a LLM which provides more generic answers

There was some hallucinations using the LLM since it was not fine tuned to use the medical journal

On query 5, I recommend updating the manual with data on the odor when a leg fractures during a hiking trip