```
import numpy as np
print(np.__version__)
```

```
1.25.2
```

```
import pandas as pd
print(pd.__version__)
```

```
1.5.3
```

```
!pip uninstall -y numpy pandas
!pip install -q numpy==1.25.2 pandas==1.5.3 --force-reinstall --no-
cache-dir
```

```
Found existing installation: numpy 2.0.2
Uninstalling numpy-2.0.2:
  Successfully uninstalled numpy-2.0.2
Found existing installation: pandas 2.2.2
Uninstalling pandas-2.2.2:
  Successfully uninstalled pandas-2.2.2
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 18.2/18.2 MB 137.5 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 12.0/12.0 MB 342.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 229.9/229.9 kB 353.5 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 509.2/509.2 kB 381.0 MB/s eta
0:00:00
ERROR: pip's dependency resolver does not currently take into account
all the packages that 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 incompatible.
mizani 0.13.3 requires pandas>=2.2.0, but you have pandas 1.5.3 which
is incompatible.
dask-expr 1.1.21 requires pandas>=2, but you have pandas 1.5.3 which
is incompatible.
cudf-cu12 25.2.1 requires pandas<2.2.4dev0,>=2.0, but you have pandas
1.5.3 which is incompatible.
dask-cudf-cu12 25.2.2 requires pandas<2.2.4dev0,>=2.0, but you have
pandas 1.5.3 which is incompatible.
tensorflow 2.18.0 requires numpy<2.1.0,>=1.26.0, but you have numpy
1.25.2 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.
xarray 2025.3.1 requires pandas>=2.1, 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.
```

```
blosc2 3.3.2 requires numpy>=1.26, but you have numpy 1.25.2 which is
incompatible.

# For installing the libraries & downloading models from HF Hub
!pip install -q huggingface_hub==0.23.2 \
             tiktoken==0.6.0 \
             pymupdf==1.25.1 \
             langchain==0.1.1 \
             langchain-community==0.0.13 \
             chromadb==0.4.22 \
             sentence-transformers==2.3.1

━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 67.3/67.3 kB 7.7 MB/s eta
0:00:00
ents to build wheel ... etadata (pyproject.toml) ...
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 44.0/44.0 kB 3.1 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 44.4/44.4 kB 4.4 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 44.4/44.4 kB 3.9 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 44.4/44.4 kB 4.4 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 44.4/44.4 kB 4.5 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 44.1/44.1 kB 4.1 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 43.5/43.5 kB 4.1 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 44.1/44.1 kB 4.0 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 401.7/401.7 kB 30.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.8/1.8 MB 75.1 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 20.0/20.0 MB 72.1 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 802.4/802.4 kB 45.3 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.6/1.6 MB 72.9 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 509.0/509.0 kB 32.8 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 132.8/132.8 kB 14.8 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.4/2.4 MB 81.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 284.2/284.2 kB 28.9 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 95.2/95.2 kB 9.1 MB/s eta
```

```
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.0/2.0 MB 74.6 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 241.2/241.2 kB 23.7 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 55.4/55.4 kB 5.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 101.6/101.6 kB 11.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 16.4/16.4 MB 77.1 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 55.9/55.9 kB 5.9 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 194.9/194.9 kB 15.4 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 65.8/65.8 kB 6.0 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 118.9/118.9 kB 14.4 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 53.0/53.0 kB 6.0 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 92.0/92.0 kB 9.8 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 6.0/6.0 MB 129.3 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 363.4/363.4 MB 5.9 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.8/13.8 MB 108.8 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 24.6/24.6 MB 80.9 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 883.7/883.7 kB 52.5 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 664.8/664.8 MB 1.6 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 211.5/211.5 MB 6.0 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 56.3/56.3 MB 13.4 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 127.9/127.9 MB 7.5 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 207.5/207.5 MB 5.5 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 21.1/21.1 MB 86.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 10.0/10.0 MB 123.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 3.0/3.0 MB 111.5 MB/s eta
0:00:00
```

```
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 62.5/62.5 kB 6.7 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 459.8/459.8 kB 36.4 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 50.9/50.9 kB 5.6 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 72.0/72.0 kB 7.9 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 4.0/4.0 MB 114.5 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 454.8/454.8 kB 39.6 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 46.0/46.0 kB 4.9 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 86.8/86.8 kB 10.0 MB/s eta
0:00:00
l) ... ERROR: pip's dependency resolver does not currently take into
account all the packages that are installed. This behaviour is the
source of the following dependency conflicts.
langchain-text-splitters 0.3.8 requires langchain-core<1.0.0,>=0.3.51,
but you have langchain-core 0.1.23 which is incompatible.
peft 0.15.2 requires huggingface_hub>=0.25.0, but you have
huggingface-hub 0.23.2 which is incompatible.
cudf-cu12 25.2.1 requires pandas<2.2.4dev0,>=2.0, but you have pandas
1.5.3 which is incompatible.
dask-cudf-cu12 25.2.2 requires pandas<2.2.4dev0,>=2.0, but you have
pandas 1.5.3 which is incompatible.
tensorflow 2.18.0 requires numpy<2.1.0,>=1.26.0, but you have numpy
1.25.2 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.
diffusers 0.33.1 requires huggingface-hub>=0.27.0, but you have
huggingface-hub 0.23.2 which is incompatible.
xarray 2025.3.1 requires pandas>=2.1, but you have pandas 1.5.3 which
is incompatible.
google-cloud-bigquery 3.31.0 requires packaging>=24.2.0, but you have
packaging 23.2 which is incompatible.
plotnine 0.14.5 requires pandas>=2.2.0, but you have pandas 1.5.3
which is incompatible.

# Installation for GPU llama-cpp-python
!CMAKE_ARGS="-DLLAMA_CUBLAS=on" FORCE_CMAKE=1 pip install llama-cpp-
python==0.1.85  --force-reinstall --upgrade --no-cache-dir -q

# Installation for CPU llama-cpp-python
# uncomment and run the following code in case GPU is not being used
# !CMAKE_ARGS="-DLLAMA_CUBLAS=off" FORCE_CMAKE=1 pip install llama-
cpp-python==0.1.85 --force-reinstall --no-cache-dir -q
```

```
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.8/1.8 MB 174.0 MB/s eta
0:00:00
ents to build wheel ... etadata (pyproject.toml) ...
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 62.0/62.0 kB 106.8 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 45.5/45.5 kB 276.3 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 16.4/16.4 MB 319.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 45.8/45.8 kB 275.5 MB/s eta
0:00:00
a-cpp-python (pyproject.toml) ... ERROR: pip's dependency resolver
does not currently take into account all the packages that are
installed. This behaviour is the source of the following dependency
conflicts.
langchain-community 0.0.13 requires numpy<2,>=1, but you have numpy
2.2.5 which is incompatible.
langchain 0.1.1 requires numpy<2,>=1, but you have numpy 2.2.5 which
is incompatible.
google-colab 1.0.0 requires pandas==2.2.2, but you have pandas 1.5.3
which is incompatible.
mizani 0.13.3 requires pandas>=2.2.0, but you have pandas 1.5.3 which
is incompatible.
dask-expr 1.1.21 requires pandas>=2, but you have pandas 1.5.3 which
is incompatible.
peft 0.15.2 requires huggingface_hub>=0.25.0, but you have
huggingface-hub 0.23.2 which is incompatible.
cudf-cu12 25.2.1 requires pandas<2.2.4dev0,>=2.0, but you have pandas
1.5.3 which is incompatible.
numba 0.60.0 requires numpy<2.1,>=1.22, but you have numpy 2.2.5 which
is incompatible.
dask-cudf-cu12 25.2.2 requires pandas<2.2.4dev0,>=2.0, but you have
pandas 1.5.3 which is incompatible.
tensorflow 2.18.0 requires numpy<2.1.0,>=1.26.0, but you have numpy
2.2.5 which is incompatible.
diffusers 0.33.1 requires huggingface-hub>=0.27.0, but you have
huggingface-hub 0.23.2 which is incompatible.
xarray 2025.3.1 requires pandas>=2.1, 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.
```

```python
#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
```

```
from langchain_community.embeddings.sentence_transformer import
SentenceTransformerEmbeddings
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

import warnings
warnings.filterwarnings('ignore')
```

Data Preparation

Loading the data

```
# mounting the drive

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

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

```
# specify the path for the pdf file

pdf_file = "/content/drive/My
Drive/colab_notebooks/RAG_medical_assistant/medical_diagnosis_manual.p
df"

# Initialize the PDF loader using PyMuPDF

pdf_loader = PyMuPDFLoader(pdf_file);

# Load content from the PDF file

text = pdf_loader.load()
```

Data Overview

```
# Print the content of the third page (index 2)

print(text[2].page_content,end="\n")
```

```
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 .....
.....
143
Chapter 9. Diagnostic & Therapeutic GI
Procedures .....
.....
150
Chapter 10. GI
Bleeding .....
.....
158
Chapter 11. Acute Abdomen & Surgical
Gastroenterology .....
.....
172
Chapter 12. Esophageal & Swallowing
Disorders .....
.....
183
Chapter 13. Gastritis & Peptic Ulcer
Disease .....
.....
196
Chapter 14. Bezoars & Foreign
Bodies .....
.....
199
Chapter 15.
Pancreatitis .....
.....
206
Chapter 16.
Gastroenteritis .....
.....
213
Chapter 17. Malabsorption
Syndromes .....
.....
225
Chapter 18. Irritable Bowel
Syndrome .....
.....
229

Chapter 19. Inflammatory Bowel Disease 241
Chapter 20. Diverticular Disease 246
Chapter 21. Anorectal Disorders 254
Chapter 22. Tumors of the GI Tract 275
3 - Hepatic & Biliary Disorders 275
Chapter 23. Approach to the Patient With Liver Disease 294
Chapter 24. Testing for Hepatic & Biliary Disorders 305
Chapter 25. Drugs & the Liver 308
Chapter 26. Alcoholic Liver Disease 314
Chapter 27. Fibrosis & Cirrhosis 322
Chapter 28. Hepatitis .....

.....
333
Chapter 29. Vascular Disorders of the Liver .....
341
Chapter 30. Liver Masses & Granulomas .....
348
Chapter 31. Gallbladder & Bile Duct Disorders .....
362
4 - Musculoskeletal & Connective Tissue Disorders .....
362
Chapter 32. Approach to the Patient With Joint Disease .....
373
Chapter 33. Autoimmune Rheumatic Disorders .....
391
Chapter 34. Vasculitis .....
416
Chapter 35. Joint Disorders .....
435
Chapter 36. Crystal-Induced Arthritides .....
443
Chapter 37. Osteoporosis .....
448
Chapter 38. Paget's Disease of Bone .....

.....
451
Chapter 39.
Osteonecrosis 455
Chapter 40. Infections of Joints &
Bones 463
Chapter 41. Bursa, Muscle & Tendon
Disorders 470
Chapter 42. Neck & Back
Pain 481
Chapter 43. Hand
Disorders .....

Question Answering using LLM

```
# Download Llama-2-13B-Chat (quantized)

model_path = hf_hub_download(
    repo_id = "TheBloke/Llama-2-13B-chat-GGUF",
    filename = "llama-2-13b-chat.Q5_K_M.gguf"
)

{"model_id":"2583a2e12b384b11b831e9eda69b1e64","version_major":2,"version_minor":0}

# Initialize the LLaMA language model with specified configuration

llm = Llama(
    model_path=model_path,
    n_threads=2,          # CPU threads for computation
```

```
    n_batch=512,          # Tokens processed in parallel
    n_gpu_layers=42,      # Layers offloaded to GPU
    n_ctx=2048,           # Context window size
)
```

```
AVX = 1 | AVX2 = 1 | AVX512 = 0 | 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 |
```

```
# 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'],
    )

    # 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?

```
query_1 = "What is the protocol for managing sepsis in a critical care
unit?"
response = generate_llama_response(query_1)
print(response)
```

```
 The management of sepsis in a critical care unit (CCU) follows a
standardized protocol that includes early recognition, aggressive
fluid resuscitation, antibiotics, and close monitoring. Here are the
key components of the protocol:
```

1. Early Recognition: The CCU team should be vigilant in identifying patients at risk for sepsis, using criteria such as SIRS (Systemic Inflammatory Response Syndrome) or QSOFA (Quick Sequential Organ Failure Assessment).
2. Aggressive Fluid Resuscitation: The goal is to achieve a mean arterial pressure (MAP) of 65-70 mmHg, and to maintain a central venous oxygen saturation (ScvO2) of >70%. Crystalloids are the first choice for fluid resuscitation, but vasopressors may be needed if fluid responsiveness is uncertain.
3. Antibiotics: Broad-spectrum antibiotics should be administered as soon as possible, ideally within the first hour of recognition of sepsis. The selection of antibiotics should cover common causes of sepsis, such as Gram-positive and Gram-negative bacteria, as well as fungi.
4. Close Monitoring: Patients with sepsis require frequent monitoring of vital signs, laboratory values (e.g., white blood cell count, lactate levels), and organ function (e.g., respiratory and cardiac status). This information is used to adjust therapy and guide decision-making.
5. Septic Shock: If sepsis progresses to septic shock, vasopressors should be initiated promptly to maintain mean arterial pressure ≥65 mmHg. The choice of vasopressor may depend on the specific clinical situation and patient factors (e.g., cardiovascular stability, renal function).
6. Source Control: In addition to addressing systemic inflammation, source control measures should be taken to eliminate or reduce the source of infection. This may involve surgical intervention, such as drainage of an abscess or debridement of infected tissue.
7. Multi-Disciplinary Care: The management of sepsis requires a multi-disciplinary approach, involving intensivists, critical care nurses, respiratory therapists, and other healthcare professionals. Regular rounds should be conducted to ensure that patients receive timely and appropriate interventions.
8. Protocolized Care: A sepsis protocol can help standardize care and guide decision-making. The protocol should include clear criteria for initiating and escalating therapy, as well as monitoring and communication strategies.
9. Family Engagement: Families should be informed about the diagnosis of sepsis and the ongoing management plan. This includes regular updates on the patient's condition, treatment goals, and any changes in their status.
10. Continuous Quality Improvement: Regular review of sepsis cases can help identify opportunities for improvement in care processes and outcomes. This may involve analyzing data from quality metrics, conducting root cause analysis of adverse events, and implementing change initiatives to enhance the quality of care provided.

By following this protocolized approach, critical care units can

```
provide evidence-based management of sepsis that is timely, effective,
and patient-centered.
```

- The model has answered the question fairly with numerical points.

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?

```
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?"
response = generate_llama_response(query_2)
print(response)

Llama.generate: prefix-match hit

 Sure, I'd be happy to help! Appendicitis is a medical emergency that
requires immediate attention. The common symptoms of appendicitis
include:

1. Severe pain in the abdomen, usually starting near the belly button
and then moving to the lower right side of the abdomen.
2. Nausea and vomiting.
3. Loss of appetite.
4. Fever.
5. Abdominal tenderness and guarding (muscle tension) when the abdomen
is touched.
6. Abdominal swelling.
7. Diarrhea or constipation.

If you suspect that you or someone else may have appendicitis, it's
important to seek medical attention immediately. Appendicitis can be
treated with antibiotics, but surgery is usually necessary to remove
the inflamed appendix. The surgical procedure most commonly used to
treat appendicitis is a laparoscopic appendectomy, which involves
making several small incisions in the abdomen and using a camera and
specialized instruments to remove the inflamed appendix.

In some cases, an open appendectomy may be necessary if the appendix
has already ruptured or if there are other complications present. This
procedure involves a larger incision in the abdomen to allow for
better visualization of the affected area.

It's important to note that while antibiotics can help treat any
underlying infection, they will not cure appendicitis on their own.
Surgery is still necessary to remove the inflamed appendix and prevent
further complications. If you suspect that you or someone else may
```

```
have appendicitis, it's important to seek medical attention
immediately to receive proper diagnosis and treatment.
```

- The model has done a good job answering all the three sub questions in the original question but has not gone into the details.

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?

```
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?"
response = generate_llama_response(query_3)
print(response)
```

```
Llama.generate: prefix-match hit

 Sudden patchy hair loss, also known as alopecia areata, can be a
distressing condition that affects millions of people worldwide. While
there is no definitive cure for this condition, there are several
effective treatments and solutions that can help promote hair growth
and reduce the size of bald spots.

     One of the most common treatment options for alopecia areata is
topical corticosteroids, which can be applied directly to the affected
area to suppress the immune system and promote hair growth. These
medications are available in various strengths, and your dermatologist
may recommend a stronger formula if the initial application does not
produce desired results.

     Another effective treatment for patchy hair loss is oral
medication such as minoxidil (Rogaine) or finasteride (Propecia).
Minoxidil can help stimulate hair growth and slow down hair loss,
while finasteride works by blocking the production of
dihydrotestosterone (DHT), a hormone that contributes to balding.

     In addition to medications, there are also several alternative
therapies that may be helpful in treating alopecia areata. For
example, some studies suggest that low-level laser therapy (LLLT) can
promote hair growth and reduce inflammation. Similarly, platelet-rich
plasma (PRP) therapy involves injecting platelets rich in growth
factors into the affected area to stimulate hair growth.

     While these treatments can be effective, it is essential to
understand that they may not work for everyone. Moreover, some people
may experience side effects from certain medications or therapies.
Therefore, it is crucial to consult a dermatologist to determine the
best course of treatment based on your individual needs and medical
```

```
history.

     As for possible causes behind sudden patchy hair loss, there are several factors that could contribute to this condition. These include genetics, autoimmune disorders, hormonal imbalances, stress, and certain medications. In some cases, alopecia areata may be triggered by a viral or bacterial infection, so identifying and treating any underlying infections can also help address the hair loss.

     Overall, while there is no cure for alopecia areata, there are several effective treatments and solutions that can help promote hair growth and reduce the size of bald spots. If you are experiencing sudden patchy hair loss, it is essential to consult a dermatologist to determine the underlying cause and develop an appropriate treatment plan.
```

- The model has answered our question using different paragraphs wherever necessary and the answer seems to be correct.

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?

```
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?"
response = generate_llama_response(query_4)
print(response)

Llama.generate: prefix-match hit

 Treatments for a physical injury to brain tissue depend on the severity and location of the injury, as well as the specific symptoms and impairments experienced by the individual. Here are some common treatments that may be recommended:

1. Medications: Various medications may be prescribed to manage symptoms such as pain, inflammation, seizures, or mood changes. These may include analgesics, anti-inflammatory drugs, anticonvulsants, and mood stabilizers.
2. Rehabilitation therapy: Physical, occupational, and speech therapies can help individuals regain lost function and improve their quality of life. These therapies may include exercises to strengthen affected muscles, activities to improve cognitive and motor skills, and strategies to compensate for any persistent impairments.
3. Surgery: In some cases, surgical intervention may be necessary to relieve pressure on the brain, repair damaged blood vessels, or remove a tumor or other lesion that is causing injury.
4. Stem cell therapy: This is a promising area of research for
```

```
treating brain injuries. Stem cells have the ability to differentiate
into different types of cells and can potentially help replace damaged
tissue in the brain.
5. Neuroplasticity-based interventions: Techniques such as cognitive
training, non-invasive brain stimulation (e.g., transcranial magnetic
stimulation), and functional electrical stimulation may be used to
promote neuroplasticity and improve brain function.
6. Lifestyle modifications: Making lifestyle changes such as regular
exercise, a healthy diet, stress management techniques, and adequate
sleep can help support overall brain health and recovery from injury.
7. Assistive technology: In some cases, assistive devices such as
wheelchairs, walkers, or communication aids may be necessary to help
individuals with permanent impairments lead independent lives.
8. Psychological interventions: Cognitive-behavioral therapy (CBT) and
other forms of talk therapy can help individuals cope with the
emotional and psychological effects of brain injury, such as
depression, anxiety, or personality changes.
9. Alternative therapies: Some people may find relief from alternative
therapies such as acupuncture, massage, or music therapy. These
treatments can help reduce stress, promote relaxation, and improve
overall well-being.

It's important to note that the most effective treatment plan for a
brain injury will depend on the individual case and may involve a
combination of these options. A healthcare professional should be
consulted to determine the best course of treatment based on the
specific needs and circumstances of the patient.
```

- Though the model has done a good job answering the question, some points like '6. Lifestyle modifications' don't directly adress the question being asked.

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?

```
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?"
response = generate_llama_response(query_5)
print(response)

Llama.generate: prefix-match hit

 Sure, I'd be happy to help! If someone has fractured their leg during
a hiking trip, there are several necessary precautions and treatment
steps that should be taken to ensure proper care and recovery. Here
are some considerations:

1. Stop activity: The first step is to stop any further physical
```

```
activity to prevent further damage or complications. This includes
stopping the hike and not putting weight on the affected leg.
2. Assess the injury: Evaluate the severity of the fracture by
assessing for signs such as deformity, swelling, pain, and difficulty
moving the limb. This will help determine the appropriate course of
treatment.
3. Immobilize the leg: Use a splint or brace to immobilize the
affected leg and prevent any further movement. This can be done using
materials found in the wilderness, such as sticks and bandanas, or by
using a pre-made splint if one is available.
4. Manage pain: Administer pain medication, such as ibuprofen or
acetaminophen, to help manage discomfort. However, be sure to follow
the recommended dosage instructions and only use medications that are
appropriate for the individual's medical history and allergies.
5. Elevate the leg: Elevate the affected leg above heart level to
reduce swelling and improve blood flow. This can be done by propping
the leg up on a backpack or other available object.
6. Seek medical attention: If possible, seek medical attention as soon
as possible. A healthcare professional will be able to assess the
severity of the fracture and provide appropriate treatment options,
such as setting the bone or applying a cast.
7. Monitor for complications: Keep an eye out for any signs of
complications, such as infection, nerve damage, or poor blood flow.
These can be identified by redness, swelling, increased pain, or
difficulty moving the limb. If any of these symptoms are present, seek
medical attention immediately.
8. Provide rest and recovery: Once the initial injury has been
treated, provide plenty of rest and allow the individual to recover.
This may involve taking a break from hiking for several weeks or
months, depending on the severity of the fracture.
9. Follow up with medical care: After seeking initial medical
attention, it is important to follow up with a healthcare professional
for further evaluation and treatment. This can include physical
therapy to help regain strength and mobility in the affected limb.
10. Consider prevention measures: Finally, consider taking steps to
prevent similar injuries from occurring in the future. This may
involve wearing appropriate footwear and protective gear during hiking
activities, as well as practicing safe hiking techniques such as
staying on designated trails and avoiding slippery or uneven terrain.

Overall, proper care and recovery for a fractured leg sustained during
a hiking trip requires prompt medical attention, immobilization of the
affected limb, pain management, elevation of the leg, monitoring for
complications, rest and recovery, follow-up medical care, and
prevention measures.
```

- The model seems to have answered the question correctly but has added prevention measures in point 10 which was not asked for.

Question Answering using LLM with Prompt Engineering

```
# function to generate, process, and return the response from the LLM
def generate_llama_response(user_prompt, max_tokens, temperature,
top_p, top_k, repeat_penalty):

    # 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=max_tokens,
        temperature=temperature,
        top_p=top_p,
        top_k=top_k,
        repeat_penalty=repeat_penalty,
        stop=['INST'],
        echo=False
    )

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

query_1 = '''
Answer the question without exceeding the maximum number of tokens and
without skipping any required steps. Also first provide some general
information regarding the disease.

What is the protocol for managing sepsis in a critical care unit?"
'''

response = generate_llama_response(query_1, 512, 0.5, 0.95, 50, 1.2)
print(response)
```

```
Llama.generate: prefix-match hit

 Sepsis is a life-threatening medical condition that arises when the
body's response to an infection becomes uncontrolled and causes
widespread inflammation. It can lead to severe organ dysfunction,
including respiratory failure, cardiac arrest, and death. The
```

```
management of sepsis is a complex and time-sensitive process that
requires prompt recognition, early intervention, and careful
monitoring in a critical care unit (CCU).

Here is the protocol for managing sepsis in a CCU:

1. Early Recognition and Activation of Sepsis Protocol: The first step
is to recognize the signs and symptoms of sepsis early and activate
the sepsis protocol immediately. This includes measuring vital signs,
particularly temperature, blood pressure, tachycardia, and tachypnea.
2. Rapid Administration of Antibiotics: Once sepsis is suspected,
antibiotics should be administered as soon as possible, ideally within
the first hour of recognition. The choice of antibiotic depends on the
likely source of infection and the patient's allergies and medical
history.
3. Fluid Resuscitation: Sepsis patients may have hypovolemia or
hypoperfusion due to vasodilation, so fluid resuscitation is critical
to maintain blood pressure and perfusion. Crystalloid fluids such as
lactated Ringer's solution or normal saline should be used initially,
and vasopressors may be added if necessary.
4. Oxygen Therapy: Sepsis patients often have hypoxia due to poor
oxygenation, so providing supplemental oxygen is essential. The oxygen
therapy should be titrated based on the patient's blood gases and
clinical response.
5. Monitoring of Vital Signs and Laboratory Parameters: Close
monitoring of vital signs (temperature, blood pressure, tachycardia,
and tachypnea) and laboratory parameters (white blood cell count,
serum lactate, and hematocrit) is crucial to assess the patient's
clinical status and response to treatment.
6. Management of Organ Dysfunction: Sepsis can cause severe organ
dysfunction, such as respiratory failure, cardiac arrest, and acute
kidney
```

- Though we had asked the model to answer the question without exceeding the maximum number of tokens, it has failed to do so.
- As asked in the prompt, the model has given some general information regarding the disease.

```
query_2 = '''
Answer the following question without any disclaimers(including pre-
response messages), greetings, or congratulatory messages. Use bullet
points wherever necessary.

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?
'''
```

```
response = generate_llama_response(query_2, 1024, 0.2, 0.9, 50, 1.2)
print(response)
```

```
Llama.generate: prefix-match hit

 Sure! Here's my response:

Common symptoms of appendicitis include:

• Severe pain in the lower right abdomen that begins suddenly and
worsens over time

• Nausea and vomiting

• Loss of appetite

• Fever

• Abdominal tenderness and swelling

• Inability to pass gas or have a bowel movement

Medications cannot cure appendicitis. Surgery is necessary to remove
the inflamed appendix. The surgical procedure most commonly used to
treat appendicitis is an appendectomy, which involves removing the
inflamed appendix through an incision in the abdomen. In some cases, a
laparoscopic appendectomy may be performed instead, which uses smaller
incisions and a camera to guide the surgery.

It's important to seek medical attention immediately if you suspect
that you or someone else has appendicitis, as delaying treatment can
lead to complications such as the appendix rupturing and spreading
infection throughout the abdominal cavity.
```

- The model has included pre-response message (Sure! Here's my response:) though we had asked not to do so.
- Bullet points are used and are separated by two new lines.

```
query_3 = '''
Don't give one-line big response for the following questions. Break
the answer into points as necessary. Highlight important
words/phrases. Make a clear distinction between the answers
if multiple questions are packed into one.

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?"

'''
```

```
response = generate_llama_response(query_3, 1024, 0.1, 0.9, 40, 1)
print(response)
```

```
Llama.generate: prefix-match hit

 Sure, I'd be happy to help! Here's my response to the user question
based on the user prompt:

Effective Treatments and Solutions for Sudden Patchy Hair Loss:

1. Identify the underlying cause: It's essential to determine the
cause of sudden patchy hair loss to develop an appropriate treatment
plan. Possible causes include hormonal imbalances, autoimmune
conditions, fungal infections, and nutritional deficiencies.
2. Medications: Minoxidil (Rogaine) and finasteride (Propecia) are two
FDA-approved medications for treating hair loss. Minoxidil can help
stimulate hair growth and slow down hair loss, while finasteride works
by blocking the production of dihydrotestosterone (DHT), a hormone
that contributes to hair loss.
3. Low-level laser therapy (LLLT): LLLT uses low-level lasers or
light-emitting diodes to stimulate hair growth. It's thought that the
light energy increases blood flow to the scalp, promoting hair growth.
4. Platelet-rich plasma (PRP) therapy: PRP therapy involves injecting
platelet-rich plasma (PRP) into the scalp. PRP contains growth factors
that promote hair growth and can help stimulate the hair follicles.
5. Diet and nutrition: Eating a balanced diet rich in vitamins and
minerals, particularly those that promote hair growth such as iron,
zinc, and biotin, can help address sudden patchy hair loss.

Possible Causes of Sudden Patchy Hair Loss:

1. Hormonal imbalances: Hormonal changes, such as those that occur
during menopause or pregnancy, can cause sudden patchy hair loss.
2. Autoimmune conditions: Conditions like alopecia areata, lupus, and
rheumatoid arthritis can cause sudden hair loss due to inflammation
and an immune system attack on the hair follicles.
3. Fungal infections: Fungal infections like ringworm can cause patchy
hair loss due to inflammation and damage to the hair follicles.
4. Nutritional deficiencies: Deficiencies in essential nutrients like
iron, zinc, and biotin can contribute to sudden patchy hair loss.
5. Stress and anxiety: Prolonged stress and anxiety can cause hair
loss due to the body's natural response to stress, including an
increase in the production of stress hormones like cortisol.

I hope this information helps! If you're experiencing sudden patchy
hair loss, it's essential to consult with a dermatologist to determine
the underlying cause and develop an appropriate treatment plan.
```

- The answer is not nicely formatted compared to the answer without prompt.
- The answers for different sub-questions are also clearly separated

```
query_4 = '''
Answer the below question using heading and sub-headings to give a
structured output. If points are to be made, give them one below the
other.
Do not use more than 30 words in one line - go to the next line if it
exceeds 30 words in a line.

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(query_4, 1024, 0.05, 0.8, 35, 1.3)
print(response)
```

```
Llama.generate: prefix-match hit

 Sure! Here's my response:

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

Treatment options may vary depending on the severity and location of
the injury. However, here are some common recommendations that
healthcare providers might suggest:

Rehabilitation Therapy - Rehabilitation therapies such as physical,
occupational, speech, language, cognitive-behavioral, or psychotherapy
may be recommended to help regain lost functions and improve daily
living skills. The goal is to maximize the patient's potential for
recovery and minimize any long-term disability.

Medications - Medication can play a crucial role in managing symptoms
such as pain, anxiety, depression, or seizures that may result from
brain injury. Common medications include analgesics (pain relievers),
antidepressants, anti-anxiety drugs, and mood stabilizers.

Surgery - In some cases where the damage is severe or localized,
surgical intervention might be necessary to remove a blood clot,
repair damaged tissue, or alleviate pressure on affected areas of the
brain.

Lifestyle Changes - Lifestyle modifications such as regular exercise,
proper nutrition and hydration, stress management techniques like
meditation/yoga, adequate sleep hygiene, and avoiding harmful
substances can help support recovery from a brain injury.

Assistive Devices - Assistive devices may be recommended to aid daily
living activities such as walking aids (canes or walkers),
```

```
communication tools like speech-generating devices for those with
language impairments, memory aids in the form of calendars/reminders,
and cognitive training software.

Cognitive Rehabilitation - Cognitive rehab programs focus on improving
problem-solving skills, attention span management, decision making
abilities, processing speed enhancement, etc., which can be impacted
due to brain injury. These interventions may involve computerized
exercises or paper/pencil tasks designed specifically for the
patient's needs and goals.

Support Groups - Joining support groups either online or in-person
provides an opportunity to connect with others who have experienced
similar challenges, share experiences, gain insight into different
coping strategies, receive emotional support from peers/family
members, etc., all of which can contribute positively towards the
healing process.

It's important for patients and their families or caregivers to work
closely with healthcare providers throughout this journey so that they
may tailor treatment plans according to individual needs while
monitoring progress along the way!
```

- The output for this prompt is not structured in a nice way. Thought the model has used new-lines to separate points, the format is not good.

```
query_5 = '''

Tell me what you are thinking to thoroughly answer the below question
and then answer it.

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?"

'''

response = generate_llama_response(query_5, 1024, 0.01, 0.8, 30, 1.15)
print(response)

Llama.generate: prefix-match hit

 Sure, I'd be happy to help! If someone has fractured their leg during
a hiking trip, there are several necessary precautions and treatment
steps that should be taken to ensure proper care and recovery. Here
are some considerations:

1. Stop activity: The first step is to stop any further activity or
weight-bearing on the affected leg to prevent further damage or
complications. This may involve calling for help, signaling for
assistance, or using a whistle to alert others of the situation.
```

2. Assess the injury: Next, assess the severity of the fracture by checking for signs of swelling, deformity, or open wounds. If the fracture is displaced (meaning the bone is no longer in its normal position), do not attempt to realign it. Instead, immobilize the leg using a splint or brace to keep it stable.
3. Immobilize the leg: Use a splint or brace to immobilize the affected leg and prevent any further movement. This can be done using materials found in nature, such as sticks and leaves, or by creating a makeshift splint using clothing or other available items.
4. Elevate the leg: Elevate the fractured leg above the level of the heart to reduce swelling and promote blood flow. This can be done by propping the leg up against a backpack or rock, for example.
5. Apply ice: Apply ice to the affected area to reduce pain and swelling. If ice is not available, use cold water or a cold compress instead. Wrap the ice or cold compress in a cloth to avoid direct contact with the skin.
6. Monitor vital signs: Check the person's vital signs, such as their pulse, breathing rate, and blood pressure, regularly to ensure they are stable. If they are experiencing any difficulty breathing, have a rapid heartbeat, or are feeling faint, call for emergency medical assistance immediately.
7. Prepare for evacuation: Once the person is stabilized, prepare them for evacuation by carrying them on a stretcher or using a backboard to immobilize their spine and neck. If possible, use a radio or other communication device to call for help and alert authorities of your location.
8. Seek medical attention: As soon as possible, seek medical attention at a hospital or urgent care center with an orthopedic specialist. The fracture will need to be properly set and immobilized using a cast or brace, and any other injuries or complications will need to be addressed.
9. Follow-up care: After the initial treatment and evacuation, follow up with medical professionals to ensure proper healing and recovery. This may involve regular check-ins, X-rays or other imaging tests, and physical therapy exercises to restore strength and mobility to the affected leg.

In addition to these precautions and treatment steps, there are several other considerations for the care and recovery of someone who has fractured their leg during a hiking trip:

1. Keep the leg elevated: Elevate the fractured leg above the level of the heart as much as possible to reduce swelling and promote blood flow.
2. Monitor for infection: Check for signs of infection, such as redness, swelling, or pus, regularly and seek medical attention if any are present.
3. Manage pain: Use pain medication as directed by medical professionals to manage discomfort and promote healing.

```
4. Restrict activities: Avoid putting weight on the affected leg and
restrict activities that may exacerbate the injury or slow down the
recovery process.
5. Follow a rehabilitation plan: Work with physical therapists or
other medical professionals to develop a rehabilitation plan that
addresses strength, mobility, and range of motion exercises to restore
full function to the affected leg.

Overall, proper care and treatment for a fractured leg during a hiking
trip requires prompt assessment, immobilization, elevation, ice
application, vital sign monitoring, preparation for evacuation, and
seeking medical attention as soon as possible. Additionally, follow-up
care and rehabilitation are crucial to ensure proper healing and
recovery.
```

- The model has not shown what it is thinking as asked.
- The answer has still improved compared to the not prompted answer.

```
query_6 = '''

Answer the below question accurately but in three different tones:
neutral, cautious and funny.

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?"

'''

response = generate_llama_response(query_6, 2048, 0.01, 0.8, 30, 1.15)
print(response)
```

```
Llama.generate: prefix-match hit


   Sure! Here are three different tones to answer the user's question:


1. Neutral Tone:

If a person fractures their leg during a hiking trip, it is essential
to seek immediate medical attention. The necessary precautions and
treatment steps include immobilizing the affected limb, managing pain,
and ensuring proper wound care. Additionally, consideration should be
given to the individual's overall health, age, and any other medical
conditions they may have. It is crucial to follow a healthcare
professional's advice throughout the recovery process to ensure proper
healing and minimize complications.

2. Cautious Tone:
```

```
If you or someone you know has fractured their leg during a hiking trip, it is imperative to exercise extreme caution when providing care and treatment. Firstly, immobilize the affected limb using a splint or cast to prevent further damage. Secondly, manage pain through medication or alternative methods, such as breathing exercises or meditation. Thirdly, ensure proper wound care by cleaning and dressing the injury regularly. Furthermore, consider the individual's age, health status, and any pre-existing medical conditions before administering treatment. Seek professional medical help immediately if symptoms worsen or do not improve within a reasonable time frame.

3. Funny Tone:

Ouch! Fractured your leg during a hiking trip? Well, at least you got a cool story out of it! But seriously, folks, let's get down to business. First things first, immobilize that busted limb with a splint or cast. Don't worry; it'll be like wearing a fancy club on your leg. Next up, manage that pain like a pro with some meds or some deep breathing exercises. And hey, if you're feeling extra adventurous, try out some wound care techniques - just kidding (or am I?). Anyways, take it easy and don't overdo it, especially if you've got any pre-existing medical conditions. Oh, and if things go south, don't worry; your healthcare professional will be there to save the day like a superhero!
```

- As was asked in the prompt, the model has answered the question in three different tones.

Data Preparation for RAG

Data Overview

Checking the first 5 pages

```python
# print the text of the first five pages

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

```
Page Number : 1
subhash66cs@gmail.com
A5R9PMSOBN
ant for personal use by subhash66cs@g
shing the contents in part or full is liable

Page Number : 2
subhash66cs@gmail.com
```

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 143
Chapter 9. Diagnostic & Therapeutic GI Procedures 150
Chapter 10. GI Bleeding 158
Chapter 11. Acute Abdomen & Surgical Gastroenterology 172
Chapter 12. Esophageal & Swallowing Disorders 183
Chapter 13. Gastritis & Peptic Ulcer Disease 196
Chapter 14. Bezoars & Foreign Bodies 199
Chapter 15. Pancreatitis 206
Chapter 16. Gastroenteritis .....

.....
.....
213
Chapter 17. Malabsorption
Syndromes .....
.....
225
Chapter 18. Irritable Bowel
Syndrome .....
.....
229
Chapter 19. Inflammatory Bowel
Disease .....
.....
241
Chapter 20. Diverticular
Disease .....
.....
246
Chapter 21. Anorectal
Disorders .....
.....
254
Chapter 22. Tumors of the GI
Tract .....
.....
275
3 - Hepatic & Biliary
Disorders .....
.....
275
Chapter 23. Approach to the Patient With Liver
Disease .....
.....
294
Chapter 24. Testing for Hepatic & Biliary
Disorders .....
.....
305
Chapter 25. Drugs & the
Liver .....
.....
308
Chapter 26. Alcoholic Liver

Disease 314
Chapter 27. Fibrosis & Cirrhosis 322
Chapter 28. Hepatitis 333
Chapter 29. Vascular Disorders of the Liver 341
Chapter 30. Liver Masses & Granulomas 348
Chapter 31. Gallbladder & Bile Duct Disorders 362
4 - Musculoskeletal & Connective Tissue Disorders 362
Chapter 32. Approach to the Patient With Joint Disease 373
Chapter 33. Autoimmune Rheumatic Disorders 391
Chapter 34. Vasculitis 416
Chapter 35. Joint Disorders 435
Chapter 36. Crystal-Induced

Arthritides 443
Chapter 37. Osteoporosis 448
Chapter 38. Paget's Disease of Bone 451
Chapter 39. Osteonecrosis 455
Chapter 40. Infections of Joints & Bones 463
Chapter 41. Bursa, Muscle & Tendon Disorders 470
Chapter 42. Neck & Back Pain 481
Chapter 43. Hand Disorders 491
Chapter 44. Foot & Ankle Disorders 502
Chapter 45. Tumors of Bones &

Joints 510
5 - Ear, Nose, Throat & Dental Disorders 510
Chapter 46. Approach to the Patient With Ear Problems 523
Chapter 47. Hearing Loss 535
Chapter 48. Inner Ear Disorders 542
Chapter 49. Middle Ear & Tympanic Membrane Disorders 550
Chapter 50. External Ear Disorders 554
Chapter 51. Approach to the Patient With Nasal & Pharyngeal Symptoms 567
Chapter 52. Oral & Pharyngeal Disorders 578
Chapter 53. Nose & Paranasal Sinus Disorders 584
Chapter 54. Laryngeal Disorders 590
Chapter 55. Tumors of the Head & Neck .....

600
Chapter 56. Approach to Dental & Oral Symptoms 619
Chapter 57. Common Dental Disorders 629
Chapter 58. Dental Emergencies 635
Chapter 59. Temporomandibular Disorders 641
6 - Eye Disorders 641
Chapter 60. Approach to the Ophthalmologic Patient 669
Chapter 61. Refractive Error 674
Chapter 62. Eyelid & Lacrimal Disorders 680
Chapter 63. Conjunctival & Scleral Disorders 690
Chapter 64. Corneal Disorders 703
Chapter 65. Glaucoma .....

710
Chapter 66.
Cataract 713
Chapter 67.
Uveitis 719
Chapter 68. Retinal
Disorders 731
Chapter 69. Optic Nerve
Disorders 737
Chapter 70. Orbital
Diseases 742
7 - Dermatologic
Disorders 742
Chapter 71. Approach to the Dermatologic
Patient 755
Chapter 72. Principles of Topical Dermatologic
Therapy 760
Chapter 73. Acne & Related
Disorders 766
Chapter 74. Bullous
Diseases 771
Chapter 75. Cornification

Disorders 775
Chapter 76. Dermatitis 786
Chapter 77. Reactions to Sunlight 791
Chapter 78. Psoriasis & Scaling Diseases 799
Chapter 79. Hypersensitivity & Inflammatory Disorders 808
Chapter 80. Sweating Disorders 811
Chapter 81. Bacterial Skin Infections 822
Chapter 82. Fungal Skin Infections 831
Chapter 83. Parasitic Skin Infections 836
Chapter 84. Viral Skin Diseases 841
Chapter 85. Pigmentation Disorders .....

.....
846
Chapter 86. Hair
Disorders ....
855
Chapter 87. Nail
Disorders ....
861
Chapter 88. Pressure
Ulcers ....
867
Chapter 89. Benign
Tumors ....
874
Chapter 90. Cancers of the
Skin ....
882
8 - Endocrine & Metabolic
Disorders ....
882
Chapter 91. Principles of
Endocrinology ....
887
Chapter 92. Pituitary
Disorders ....
901
Chapter 93. Thyroid
Disorders ....

action.

Page Number : 5
921
Chapter 94. Adrenal
Disorders ....
936
Chapter 95. Polyglandular Deficiency
Syndromes ....
939
Chapter 96.
Porphyrias ....
949
Chapter 97. Fluid & Electrolyte
Metabolism ....
987
Chapter 98. Acid-Base Regulation &
Disorders ....
1001
Chapter 99. Diabetes Mellitus & Disorders of Carbohydrate Metabolism
.....
1024
Chapter 100. Lipid
Disorders ....
1034
Chapter 101.
Amyloidosis ....
1037
Chapter 102. Carcinoid
Tumors ....
1040
Chapter 103. Multiple Endocrine Neoplasia
Syndromes ....
1046
9 - Hematology &

Oncology 1046
Chapter 104. Approach to the Patient With Anemia 1050
Chapter 105. Anemias Caused by Deficient Erythropoiesis 1061
Chapter 106. Anemias Caused by Hemolysis 1078
Chapter 107. Neutropenia & Lymphocytopenia 1086
Chapter 108. Thrombocytopenia & Platelet Dysfunction 1097
Chapter 109. Hemostasis 1104
Chapter 110. Thrombotic Disorders 1107
Chapter 111. Coagulation Disorders 1113
Chapter 112. Bleeding Due to Abnormal Blood Vessels 1116
Chapter 113. Spleen Disorders 1120
Chapter 114. Eosinophilic

Disorders 1126
Chapter 115. Histiocytic Syndromes 1131
Chapter 116. Myeloproliferative Disorders 1141
Chapter 117. Leukemias 1154
Chapter 118. Lymphomas 1164
Chapter 119. Plasma Cell Disorders 1172
Chapter 120. Iron Overload 1177
Chapter 121. Transfusion Medicine 1186
Chapter 122. Overview of Cancer 1198
Chapter 123. Tumor Immunology 1204
Chapter 124. Principles of Cancer

Therapy 1215
10 - Immunology; Allergic Disorders 1215
Chapter 125. Biology of the Immune System 1227
Chapter 126. Immunodeficiency Disorders 1243
Chapter 127. Allergic & Other Hypersensitivity Disorders 1263
Chapter 128. Transplantation 1281
11 - Infectious Diseases 1281
Chapter 129. Biology of Infectious Disease 1300
Chapter 130. Laboratory Diagnosis of Infectious Disease 1306
Chapter 131. Immunization 1313
Chapter 132. Bacteria & Antibacterial Drugs 1353
Chapter 133. Gram-Positive Cocci .....

.
1366
Chapter 134. Gram-Positive
Bacilli ..
..
1376
Chapter 135. Gram-Negative
Bacilli ..
1405
Chapter 136.
Spirochetes ..
1413
Chapter 137.
Neisseriaceae ..
1419
Chapter 138. Chlamydia &
Mycoplasmas ..
1421
Chapter 139. Rickettsiae & Related
Organisms ..
1431
Chapter 140. Anaerobic
Bacteria ..
1450
Chapter 141.
Mycobacteria ..
1470
Chapter 142.
Fungi ..
1493
Chapter 143. Approach to Parasitic
Infections ..
1496
Chapter 144. Nematodes

```
(Roundworms)    ..........................................................
..........................................................................
subhash66cs@gmail.com
A5R9PMSOBN
This file is meant for personal use by subhash66cs@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal
action.
```

Checking the number of pages

```
len(text)
```

```
4114
```

Data Chunking

```
# Initialize a recursive text splitter using the tiktoken encoder

text_splitter = RecursiveCharacterTextSplitter.from_tiktoken_encoder(
    encoding_name='cl100k_base',
    chunk_size=512,
    chunk_overlap= 20
)

# Load PDF content and splits it into smaller chunks using the
provided text splitter.

document_chunks = pdf_loader.load_and_split(text_splitter)

len(document_chunks)
```

```
8469
```

```
document_chunks[45].page_content
```

```
{"type":"string"}
```

```
document_chunks[46].page_content
```

```
{"type":"string"}
```

Embedding

```
# Initialize a Sentence Transformer embedding model for generating
document embeddings

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

```
/usr/local/lib/python3.11/dist-packages/huggingface_hub/utils/
_token.py:89: 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://huggingface.co/settings/tokens), set it as
secret in your Google Colab and restart your session.
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(
```


```
{"model_id":"83438c1140ed4ed8bff76887565a2706","version_major":2,"vers
ion_minor":0}

{"model_id":"b0ebca9c30cc42d8920f651f62365cc8","version_major":2,"vers
ion_minor":0}

{"model_id":"d9e5bb6c73a344e889464de12d9253a2","version_major":2,"vers
ion_minor":0}

{"model_id":"125e3618ae0a4cfe9332c32628a3c6a9","version_major":2,"vers
ion_minor":0}

{"model_id":"8c5fcfab31e4442cbbd2e0d48e4fa859","version_major":2,"vers
ion_minor":0}

{"model_id":"4a466c51173144969b0b677a2a53b305","version_major":2,"vers
ion_minor":0}

{"model_id":"badf984478cf4ee1b33f588f61af8d66","version_major":2,"vers
ion_minor":0}

{"model_id":"36a81cb5207449208d9361fed397fcb6","version_major":2,"vers
ion_minor":0}

{"model_id":"99b9b37dba184de9a0a8d410f43de50c","version_major":2,"vers
ion_minor":0}

{"model_id":"28ed138bebd44591800aa71c8fa0df38","version_major":2,"vers
ion_minor":0}
```


```
embedding_3 =
embedding_model.embed_query(document_chunks[3].page_content)
print("Dimension of the embedding vector ",len(embedding_3))

Dimension of the embedding vector  1024
```

Vector Database

```
# Define the output directory for the database and creates it if it
doesn't exist
```

```
out_dir = 'medical_assistance_db'

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

# Create a Chroma vector store from document chunks, embeddings, and
saves it to the specified directory

vectorstore = Chroma.from_documents(
    documents=document_chunks,                    # Text chunks
    embedding=embedding_model,                    # Embedding function
    collection_name="medical_articles",
    persist_directory=out_dir                     # Where to save
)
```

Retriever

```
# Create a retriever from the vector store to perform similarity-based
document retrieval

retriever = vectorstore.as_retriever(
    search_type='similarity',
    search_kwargs={'k': 2}
)

# Check the chunks from the retriever for a sample

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

relevant_document_chunks_1 =
retriever.get_relevant_documents(user_input)
relevant_document_chunks_1
```

```
[Document(page_content="16 - Critical Care Medicine\nChapter 222.
Approach to the Critically Ill Patient\nIntroduction\nCritical care
medicine specializes in caring for the most seriously ill patients.
These patients are best\ntreated in an ICU staffed by experienced
personnel. Some hospitals maintain separate units for special\
npopulations (eg, cardiac, surgical, neurologic, pediatric, or
neonatal patients). ICUs have a high\nnurse:patient ratio to provide
the necessary high intensity of service, including treatment and
monitoring\nof physiologic parameters.\nSupportive care for the ICU
patient includes provision of adequate nutrition (see p. 21) and
prevention of\ninfection, stress ulcers and gastritis (see p. 131),
and pulmonary embolism (see p. 1920). Because 15 to\n25% of patients
admitted to ICUs die there, physicians should know how to minimize
suffering and help\ndying patients maintain dignity (see p. 3480).\
nPatient Monitoring and Testing\nSome monitoring is manual (ie, by
direct observation and physical examination) and intermittent, with
```

the\nfrequency depending on the patient's illness. This monitoring usually includes measurement of vital signs\n(temperature, BP, pulse, and respiration rate), quantification of all fluid intake and output, and often daily\nweight. BP may be recorded by an automated sphygmomanometer; a transcutaneous sensor for pulse\noximetry is used as well.\nOther monitoring is ongoing and continuous, provided by complex devices that require special training\nand experience to operate. Most such devices generate an alarm if certain physiologic parameters are\nexceeded. Every ICU should strictly follow protocols for investigating alarms.\nBlood Tests\nAlthough frequent blood draws can destroy veins, cause pain, and lead to anemia, ICU patients typically\nhave routine daily blood tests to help detect problems early. Generally, patients need a daily set of\nelectrolytes and a CBC. Patients with arrhythmias should also have Mg, phosphate, and Ca levels\nmeasured. Patients receiving TPN need weekly liver enzymes and coagulation profiles. Other tests (eg,\nblood culture for fever, CBC after a bleeding episode) are done as needed.\nPoint-of-care testing uses miniaturized, highly automated devices to do certain blood tests at the patient's\nbedside or unit (particularly ICU, emergency department, and operating room). Commonly available tests",
metadata={'author': '', 'creationDate': 'D:20120615054440Z', 'creator': 'Atop CHM to PDF Converter', 'file_path': '/content/drive/My Drive/colab_notebooks/RAG_medical_assistant/medical_diagnosis_manual.pdf', 'format': 'PDF 1.7', 'keywords': '', 'modDate': 'D:20250429105108Z', 'page': 2400, 'producer': 'pdf-lib (https://github.com/Hopding/pdf-lib)', 'source': '/content/drive/My Drive/colab_notebooks/RAG_medical_assistant/medical_diagnosis_manual.pdf', 'subject': '', 'title': 'The Merck Manual of Diagnosis & Therapy, 19th Edition', 'total_pages': 4114, 'trapped': ''}),
 Document(page_content="Parenteral antibiotics should be given after specimens of blood, body fluids, and wound sites have been\ntaken for Gram stain and culture. Very prompt empiric therapy, started immediately after suspecting\nsepsis, is essential and may be lifesaving. Antibiotic selection requires an educated guess based on the\nsuspected source, clinical setting, knowledge or suspicion of causative organisms and of sensitivity\npatterns common to that specific inpatient unit, and previous culture results.\nOne regimen for septic shock of unknown cause is gentamicin or tobramycin 5.1 mg/kg IV once/day plus\na 3rd-generation cephalosporin (cefotaxime 2 g q 6 to 8 h or ceftriaxone 2 g once/day or, if Pseudomonas\nis suspected, ceftazidime 2 g IV q 8 h). Alternatively, ceftazidime plus a fluoroquinolone (eg, ciprofloxacin)\nmay be used. Monotherapy with maximal therapeutic doses of ceftazidime (2 g IV q 8 h) or imipenem (1 g\nIV q 6 h) may be effective but is not recommended.\nVancomycin must be added if resistant staphylococci or enterococci are suspected. If there is an\nabdominal source, a drug effective against anaerobes (eg, metronidazole) should be included. When\nculture and sensitivity results are available, the antibiotic regimen is changed accordingly.

```
Antibiotics are\ncontinued for at least 5 days after shock resolves
and evidence of infection subsides.\nAbscesses must be drained, and
necrotic tissues (eg, infarcted bowel, gangrenous gall-bladder,\
nabscessed uterus) must be surgically excised. The patient's condition
will continue to deteriorate despite\nantibiotic therapy unless septic
foci are eliminated.\nNormalization of blood glucose improves outcome
in critically ill patients, even those not known to be\ndiabetic. A
continuous IV insulin infusion (crystalline zinc 1 to 4 U/h) is
titrated to maintain glucose\nbetween 80 to 110 mg/dL (4.4 to 6.1
mmol/L). This approach necessitates frequent (eg, q 1 to 4 h)\nglucose
measurement.", metadata={'author': '', 'creationDate':
'D:20120615054440Z', 'creator': 'Atop CHM to PDF Converter',
'file_path': '/content/drive/My
Drive/colab_notebooks/RAG_medical_assistant/medical_diagnosis_manual.p
df', 'format': 'PDF 1.7', 'keywords': '', 'modDate':
'D:20250429105108Z', 'page': 2456, 'producer': 'pdf-lib
(https://github.com/Hopding/pdf-lib)', 'source': '/content/drive/My
Drive/colab_notebooks/RAG_medical_assistant/medical_diagnosis_manual.p
df', 'subject': '', 'title': 'The Merck Manual of Diagnosis & Therapy,
19th Edition', 'total_pages': 4114, 'trapped': ''})]

# relevant_document_chunks_2 =
retriever_2.get_relevant_documents(user_input)
# relevant_document_chunks_2
```

System and User Prompt Template¶

```
qna_system_message = """
<s>[INST] <<SYS>> You are a medical assistant whose work is to review
the report and provide the appropriate answers from the 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 query.

User question will begin with the token: ###Question.

Please answer only using the context provided in the input. Do not
mention anything about the context in your final answer.


If the answer is not found in the context, respond "I don't know".
<</SYS>>
"""

qna_user_message_template = """
###Context
Here are some documents that are relevant to the question mentioned
below.
```

```
{context}

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

Response Function

```
def generate_rag_response(user_input, k, max_tokens, temperature,
top_p, top_k):

    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 = f"""[INST]{qna_system_message}\n
                {'user'}:
{qna_user_message_template.format(context=context_for_query,
question=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
```

Question Answering using RAG

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

```
query_1 = "What is the protocol for managing sepsis in a critical care unit?"

response = generate_rag_response(query_1, 4, 1024, 0.1, 0.9, 40)
print(response)
```

```
Llama.generate: prefix-match hit

###Context

Based on the provided context, the protocol for managing sepsis in a critical care unit includes the following:

1. Early recognition of sepsis, including prompt identification of the source of infection and suspected causative organisms.
2. Prompt empiric therapy with broad-spectrum antibiotics, including gentamicin or tobramycin and a 3rd-generation cephalosporin, or alternatively, ceftazidime plus a fluoroquinolone.
3. Addition of vancomycin if resistant staphylococci or enterococci are suspected.
4. Inclusion of a drug effective against anaerobes (eg, metronidazole) if there is an abdominal source.
5. Continuation of antibiotics for at least 5 days after shock resolves and evidence of infection subsides.
6. Surgical drainage of abscesses and excision of necrotic tissues (eg, infarcted bowel, gangrenous gall-bladder, abscessed uterus) if present.
7. Maintenance of blood glucose levels between 80 to 110 mg/dL (4.4 to 6.1 mmol/L) with a continuous IV insulin infusion titrated to the patient's needs.

Please note that these recommendations are based on the provided context and may not be applicable in all situations or units. It is important to consult with experienced personnel and follow established protocols for managing sepsis in critical care units.
```

- Now we'll check the RAG response with temperature value of 0.3

```
query_1 = "What is the protocol for managing sepsis in a critical care unit?"

response = generate_rag_response(query_1, 4, 1024, 0.3, 0.9, 40)
print(response)
```

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

```
###Context

Based on the provided context, the protocol for managing sepsis in a
critical care unit includes the following:

1. Early recognition of sepsis, including prompt identification of the
source of infection and suspected causative organisms.
2. Prompt empiric therapy with broad-spectrum antibiotics, including
gentamicin or tobramycin and a 3rd-generation cephalosporin, or
alternatively, ceftazidime plus a fluoroquinolone.
3. Monitoring of vital signs, including temperature, blood pressure,
pulse, and respiration rate, as well as quantification of all fluid
intake and output.
4. Measurement of electrolytes and a complete blood count (CBC) daily,
with additional tests such as blood culture for fever and coagulation
profiles as needed.
5. Use of point-of-care testing for certain blood tests at the
patient's bedside or unit.
6. Administration of parenteral antibiotics after specimens of blood,
body fluids, and wound sites have been taken for Gram stain and
culture.
7. Surgical drainage of abscesses and excision of necrotic tissues as
needed.
8. Maintenance of normal blood glucose levels (80-110 mg/dL) through
continuous IV insulin infusion, with frequent glucose measurements.

I don't know if there is any additional information in the context
that could help answer the question.
```

- With a higher value of temperature, the model is being creative in using the words.

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?

```
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?"

response = generate_rag_response(query_2, 3, 1024, 0.3, 0.95, 50)
print(response)

Llama.generate: prefix-match hit

###Context
The document discusses the treatment of acute appendicitis, including
the use of antibiotics and surgical procedures.

###Question
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?

Answer:
The common symptoms of appendicitis include abdominal pain, anorexia,
and abdominal tenderness. Appendicitis is not curable via medicine,
and surgical intervention is necessary. The recommended surgical
procedure for treating appendicitis is appendectomy, which can be
either open or laparoscopic. If the appendix is perforated,
antibiotics should be continued until the patient's temperature and
WBC count have normalized or for a fixed course. In cases where
surgery is impossible, antibiotics can improve survival rates but are
not curative.
```

- Reduce top_p to 0.8 and top_k to 40

```
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?"

response = generate_rag_response(query_2, 3, 1024, 0.3, 0.8, 40)
print(response)

Llama.generate: prefix-match hit

###Context
• Surgical removal
• IV fluids and antibiotics
Treatment of acute appendicitis is open or laparoscopic appendectomy;
because treatment delay increases mortality, a negative appendectomy
rate of 15% is considered acceptable. The surgeon can usually remove
the appendix even if perforated. Occasionally, the appendix is
difficult to locate: In these cases, it usually lies behind the cecum
or the ileum and mesentery of the right colon. A contraindication to
appendectomy is inflammatory bowel disease involving the cecum.
However, in cases of terminal ileitis and a normal cecum, the appendix
should be removed.
Appendectomy should be preceded by IV antibiotics. Third-generation
cephalosporins are preferred. For nonperforated appendicitis, no
further antibiotics are required. If the appendix is perforated,
antibiotics should be continued until 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—
although not curative—markedly improve the survival rate. When a large
inflammatory mass is found involving the appendix, terminal ileum, and
cecum, resection of the entire mass and ileocolostomy are preferable.
In late cases in which a pericolic abscess has already formed, the
abscess is drained either by an ultrasound-guided percutaneous
catheter or by open operation (with appendectomy to follow at a later
date).
```

```
###Question
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?

Answer:
The common symptoms of appendicitis include abdominal pain, anorexia, and abdominal tenderness. Unfortunately, appendicitis cannot be cured via medicine, and surgical intervention is necessary. The recommended surgical procedure for treating appendicitis is open or laparoscopic appendectomy.
```

- With the decreased value of top_p and top_k, the model is not being elaborate in answering.

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?

```
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?"

response = generate_rag_response(query_3, 4, 1024, 0.01, 0.95, 60)
print(response)

###Answer
Based on the provided context, the effective treatments or solutions for sudden patchy hair loss, commonly seen as localized bald spots on the scalp, include:

1. Topical corticosteroids: These are effective in treating alopecia areata and can stimulate hair growth. Efficacy is usually evident within 6 to 8 months of treatment.
2. Intralesional corticosteroids: These may be useful for more severe cases of alopecia areata.
3. Topical minoxidil: This medication can help stimulate hair growth and is effective in treating male-pattern baldness.
4. Oral antifungals: These may be used to treat tinea capitis, a fungal infection that can cause patchy hair loss.
5. Behavior modification: This may be helpful in treating trichotillomania, a condition characterized by the compulsive pulling out of one's own hair.
6. Medications such as clomipramine or SSRIs (selective serotonin reuptake inhibitors): These may be effective in treating trichotillomania and other psychological conditions that can contribute to hair loss.
```

```
Possible causes of sudden patchy hair loss include:

1. Alopecia areata: This is an autoimmune disorder that can cause
sudden, patchy hair loss.
2. Telogen effluvium: This is a condition in which there is a sudden
increase in the number of hair follicles that stop growing and enter
the resting phase, leading to excessive hair shedding.
3. Traction alopecia: This is a type of hair loss caused by constant
pulling or tension on the hair shaft, often seen in people who wear
tight hairstyles such as braids or ponytails.
4. Tinea capitis: This is a fungal infection that can cause patchy
hair loss.
5. Trichotillomania: This is a psychological condition characterized
by the compulsive pulling out of one's own hair.

It is important to note that these treatments and solutions may not be
appropriate or effective for all cases of sudden patchy hair loss, and
a thorough evaluation by a healthcare professional is necessary to
determine the underlying cause and develop an appropriate treatment
plan.
```

```
len(response)
```

- Using k value of 8

```
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?"

response = generate_rag_response(query_3, 8, 1024, 0.01, 0.95, 60)
print(response)
```

```
Llama.generate: prefix-match hit

###Answer
Based on the provided context, the effective treatments or solutions
for sudden patchy hair loss, commonly seen as localized bald spots on
the scalp, include:

1. Topical corticosteroids: These are effective in treating alopecia
areata and can stimulate hair growth. Efficacy is usually evident
within 6 to 8 months of treatment.
2. Intralesional corticosteroids: These may be useful for more severe
cases of alopecia areata.
3. Topical minoxidil: This medication can help stimulate hair growth
and is effective in treating male-pattern baldness.
4. Oral antifungals: These may be used to treat tinea capitis, a
fungal infection that can cause patchy hair loss.
5. Behavior modification: This may be helpful in treating
trichotillomania, a condition characterized by the compulsive pulling
out of one's own hair.
```

```
6. Medications such as clomipramine or SSRIs (selective serotonin
reuptake inhibitors): These may be effective in treating
trichotillomania and other psychological conditions that can
contribute to hair loss.

Possible causes of sudden patchy hair loss include:

1. Alopecia areata: This is an autoimmune disorder that can cause
sudden, patchy hair loss.
2. Telogen effluvium: This is a condition in which there is a sudden
increase in the number of hair follicles that stop growing and enter
the resting phase, leading to excessive hair shedding.
3. Traction alopecia: This is a type of hair loss caused by constant
pulling or tension on the hair shaft, such as from tight hairstyles or
hair extensions.
4. Tinea capitis: This is a fungal infection that can cause patchy
hair loss.
5. Trichotillomania: This is a psychological condition characterized
by the compulsive pulling out of one's own hair.

It is important to note that these treatments and solutions may not be
appropriate or effective for all cases of sudden patchy hair loss, and
a thorough evaluation and diagnosis by a healthcare professional is
necessary to determine the underlying cause and appropriate treatment.
```

- Increasing the k value from 4 to 8 has not changed the model's response.

```
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?"

response = generate_rag_response(query_4, 2, 1024, 0, 0.8, 40)
print(response)

Llama.generate: prefix-match hit

Based on the context provided, the recommended treatments for a person
who has sustained a physical injury to brain tissue, resulting in
temporary or permanent impairment of brain function, are:

1. Supportive care to prevent systemic complications due to
immobilization, such as pneumonia, UTI, and thromboembolic disease.
2. Good nutrition to ensure adequate energy and nutrients for healing.
3. Prevention of pressure ulcers to maintain skin integrity.

These treatments are based on the information provided in Chapter 324
of The Merck Manual of Diagnosis & Therapy, 19th Edition, specifically
in the section titled "Treatment."
```

- Let's increase the k value to 6

```
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?"

response = generate_rag_response(query_4, 6, 1024, 0, 0.8, 40)
print(response)
```

```
Llama.generate: prefix-match hit

Based on the context provided, the recommended treatments for a person
who has sustained a physical injury to brain tissue, resulting in
temporary or permanent impairment of brain function, are:

1. Supportive care to prevent systemic complications due to
immobilization, such as pneumonia, UTI, and thromboembolic disease.
2. Good nutrition to ensure adequate energy and nutrients for healing.
3. Prevention of pressure ulcers to maintain skin integrity.

These treatments are based on the information provided in Chapter 324
of The Merck Manual of Diagnosis & Therapy, 19th Edition, specifically
in the section titled "Treatment."
```

- Increasing the k value from 2 to 6 has not changed the model's answer.

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?

```
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?"

response = generate_rag_response(query_5, 3, 1024, 0, 0.7, 35)
```

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

- Check with increasing k, temperature, top_p and top_k

```
khjquery_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?"

response = generate_rag_response(query_5, 5, 1024, 0.09, 0.95, 50)
print(response)
```

```
Llama.generate: prefix-match hit

###Context
The patient has fractured their leg during a hiking trip.

###Question
```

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?

Based on the provided context, here are the necessary precautions and treatment steps for a person who has fractured their leg during a hiking trip:

1. Immobilization: The injured limb should be immobilized to prevent further injury and facilitate healing. This can be achieved using a splint or cast, depending on the severity of the fracture.
2. Elevation: The injured limb should be elevated above the heart to reduce swelling and minimize the risk of compartment syndrome.
3. Pain management: Analgesics should be administered to manage pain and discomfort.
4. Warmth application: Periodic application of warmth, such as a heating pad, may relieve pain and speed healing.
5. Inspection and hygiene: The cast or splint should be inspected daily for any signs of skin irritation or infection, and the patient should be given written instructions on how to care for their cast or splint.
6. Follow-up medical care: Regular follow-up appointments with a healthcare provider are essential to monitor the healing process and address any complications that may arise.
7. Early mobilization: Resumption of active motion within the first few days or weeks can help minimize contractures and muscle atrophy, thus accelerating functional recovery.

It is important to consider the following factors when caring for a patient with a fractured leg:

1. Age: The risk of complications such as compartment syndrome, stiffness, and muscle atrophy may be higher in older patients.
2. Type of fracture: Different types of fractures require different treatment approaches. For example, vertebral compression fractures may require bed rest and immobilization, while metatarsal stress fractures may require early mobilization.
3. Co-morbidities: Patients with co-morbidities such as diabetes or heart disease may be at higher risk for complications and may require more intensive monitoring and treatment.
4. Environmental factors: The patient's living environment and access to medical care can impact their recovery and outcome. For example, patients living in remote areas may have limited access to medical resources, which can affect their ability to receive timely and appropriate treatment.

Output Evaluation

Defining the Evaluation Prompts

```
groundedness_rater_system_message = """
You are tasked with rating AI generated answers to questions posed by
users.
You will be presented a question, context used by the AI system to
generate the answer and an AI generated answer to the question.
In the input, the question will begin with ###Question, the context
will begin with ###Context while the AI generated answer will begin
with ###Answer.

Evaluation criteria:
The task is to judge the extent to which the metric is followed by the
answer.
1 - The metric is not followed at all
2 - The metric is followed only to a limited extent
3 - The metric is followed to a good extent
4 - The metric is followed mostly
5 - The metric is followed completely

Metric:
The answer should be derived only from the information presented in
the context

Instructions:
1. First write down the steps that are needed to evaluate the answer
as per the metric.
2. Give a step-by-step explanation if the answer adheres to the metric
considering the question and context as the input.
3. Next, evaluate the extent to which the metric is followed.
4. Use the previous information to rate the answer using the evaluaton
criteria and assign a score.
"""

relevance_rater_system_message = """
You are tasked with rating AI generated answers to questions posed by
users.
You will be presented a question, context used by the AI system to
generate the answer and an AI generated answer to the question.
In the input, the question will begin with ###Question, the context
will begin with ###Context while the AI generated answer will begin
with ###Answer.

Evaluation criteria:
The task is to judge the extent to which the metric is followed by the
answer.
1 - The metric is not followed at all
2 - The metric is followed only to a limited extent
```

```
3 - The metric is followed to a good extent
4 - The metric is followed mostly
5 - The metric is followed completely

Metric:
Relevance measures how well the answer addresses the main aspects of
the question, based on the context.
Consider whether all and only the important aspects are contained in
the answer when evaluating relevance.

Instructions:
1. First write down the steps that are needed to evaluate the context
as per the metric.
2. Give a step-by-step explanation if the context adheres to the
metric considering the question as the input.
3. Next, evaluate the extent to which the metric is followed.
4. Use the previous information to rate the context using the
evaluaton criteria and assign a score.
"""

user_message_template = """
###Question
{question}

###Context
{context}

###Answer
{answer}
"""

def generate_ground_relevance_response(user_input, k, max_tokens,
temperature, top_p, top_k):
    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,
question=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,
question=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,
question=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']
```

Let's check the output of the RAG for the queries using the same parameters used for each.

```
query_1 = "What is the protocol for managing sepsis in a critical care
unit?"
```

```
ground, rel = generate_ground_relevance_response(query_1, 4, 1024,
0.1, 0.9, 40)

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

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

  Sure, I'd be happy to help! Here's my evaluation of the answer based
on the provided context:

1. The metric is followed completely. The answer provides a detailed
and specific protocol for managing sepsis in a critical care unit,
derived only from the information presented in the context.
2. The answer adheres to the metric because it provides a step-by-step
approach to managing sepsis, including early recognition, prompt
empiric therapy, addition of vancomycin if necessary, and surgical
intervention as needed. The answer also mentions the importance of
maintaining blood glucose levels.
3. The answer is derived directly from the information

  Sure, I'd be happy to help! Here's my evaluation of the context and
the answer provided:

1. Relevance: The answer addresses all the main aspects of the
question, including early recognition of sepsis, prompt empiric
therapy, addition of vancomycin if resistant staphylococci or
enterococci are suspected, inclusion of a drug effective against
anaerobes if there is an abdominal source of infection, continuation
of antibiotics for at least 5 days after shock
```

```
query_1 = "What is the protocol for managing sepsis in a critical care
unit?"
ground, rel = generate_ground_relevance_response(query_1, 4, 1024,
0.3, 0.9, 40)

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

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

  Sure, I can help you evaluate the answer based on the provided
context and criteria. Here are the steps to follow:

Step 1: Evaluate if the answer is derived only from the information
presented in the context.
```

```
The answer provides a detailed protocol for managing sepsis in a
critical care unit, which is derived directly from the information
presented in the context. The answer mentions specific antibiotics and
their dosages, surgical interventions, and blood glucose management,
all of which are based on the information provided in the context.
Therefore, the answer meets this criterion fully.

Step 2: Evaluate if the answer follows the metric completely.

The answer provides a comprehensive protocol for

  Sure, I can help you rate the context using the evaluation criteria.
Based on the provided context, here are my ratings:

1. Relevance: The context is highly relevant to the question as it
provides specific guidelines for managing sepsis in a critical care
unit. All the important aspects of sepsis management are covered,
including early recognition, prompt empiric therapy, and surgical
intervention when necessary. Rating: 5 (completely followed).
2. Completeness: The context provides a comprehensive protocol for
managing sepsis in a critical care unit, including
```

- For query 1, the model has done well with both set of parameters: one with temperature of 0.1 and the other with temperature of 0.3.

```
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?"
ground, rel = generate_ground_relevance_response(query_2, 3, 1024,
0.3, 0.95, 50)

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

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

  Sure, I can help you rate the answer based on the evaluation
criteria provided. Here are the steps to evaluate the answer:

Step 1: Evaluate if the answer is derived only from the information
presented in the context.

The answer mentions that the common symptoms of appendicitis include
abdominal pain, anorexia, and abdominal tenderness, which are all
present in the context. However, the answer also mentions that there
is no curative medication for appendicitis, which is not explicitly
stated in the context. Therefore, the answer partially adheres to the
metric.
```

Step 2: Evaluate the extent to which the metric is followed.

The answer partially follows the metric because it derives some information from the context but not all of it. The answer mentions that there is no curative medication for appendicitis, which is not explicitly stated in the context. Therefore, the answer should be rated as follows:

Rating: 2 - The metric is followed only to a limited extent.

Step 3: Use the previous information to rate the answer using the evaluation criteria.

Based on the rating of 2, I would assign a score of 2 to the answer. This score indicates that the answer partially follows the metric and could be improved by providing more information derived only from the context.

Here's the final rating and score for the answer:

Rating: 2 - The metric is followed only to a limited extent
Score: 2 - Partially follows the metric, could be improved by providing more information derived only from the context.

Sure, I can help you rate the context and answer provided.

First, let me break down the steps needed to evaluate the context as per the metric:

1. Identify the main aspects of the question: The main aspects of the question are the symptoms of appendicitis, the possibility of curative medicine, and the surgical procedure to treat it.
2. Check if all important aspects are contained in the answer: The answer addresses all the main aspects of the question, including the common symptoms of appendicitis, the lack of curative medication, and the recommended surgical procedures (appendexomy and antibiotics).
3. Evaluate relevance: The context is highly relevant to the question since it provides detailed information about the treatment of acute appendicitis, including the use of antibiotics and surgical procedures.

Now, let me rate the extent to which the metric is followed based on the provided context:

The context follows the metric completely. It provides a comprehensive overview of the treatment of acute appendicitis, covering all the important aspects mentioned in the question. The answer is well-structured and addresses all the main points of the question.

Based on the above information, I would rate the context using the

```
evaluation criteria as follows:

5 - The metric is followed completely.

The context provides a detailed and comprehensive overview of the
treatment of acute appendicitis, covering all the important aspects
mentioned in the question. It is well-structured and addresses all the
main points of the question, making it an excellent example of
following the metric completely.
```

```
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?"
ground, rel = generate_ground_relevance_response(query_2, 3, 1024,
0.3, 0.8, 40)

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

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

  Sure, I'd be happy to help you evaluate the answer!

To rate the answer, we need to consider how well it adheres to the
metric, which is that the answer should be derived only from the
information presented in the context.

Step-by-step explanation:

1. The answer mentions the common symptoms of appendicitis, which are
present in the context.
2. The answer correctly states that there is no curative medication
for appendicitis and that surgical removal of the inflamed appendix is
necessary.
3. The answer refers to the use of antibiotics if the appendix is
perforated, which is also mentioned in the context.

Based on the above steps, it appears that the answer adheres well to
the metric and derives all information from the provided context.

Rating:

Based on the evaluation criteria, I would rate the answer as 5 - The
metric is followed completely. The answer provides a clear and
accurate summary of the treatment of appendicitis, including the use
of antibiotics, and only refers to information presented in the
context.

Score: 5/5
```

```
  Sure, I'd be happy to help you evaluate the context and answer!

Step 1: Evaluate the context as per the metric.

The context provides relevant information about the treatment of acute
appendicitis, including the surgical removal of the inflamed appendix
and the use of antibiotics. The text also mentions the classification
of abdominal hernias and their symptoms. Therefore, the context
adheres well to the metric, as it addresses all the important aspects
of the question.

Step 2: Evaluate the extent to which the metric is followed.

The answer provides a detailed description of the common symptoms of
appendicitis, including abdominal pain, anorexia, and abdominal
tenderness. The answer also mentions the surgical procedure for
treating appendicitis, which is appendectomy, and the use of
antibiotics in case of a perforated appendix. Therefore, the answer
follows the metric to a good extent.

Step 3: Rate the context using the evaluation criteria.

Based on the information provided in the context and the answer, I
would rate the context as 4 (mostly followed) and the answer as 4
(mostly followed) on the evaluation criteria. The context provides
relevant information about the treatment of acute appendicitis and
classifies abdominal hernias, while the answer describes the common
symptoms of appendicitis and the surgical procedure for treating it.
However, the answer could have provided more information about the use
of antibiotics in case of a perforated appendix.

Overall, the context and answer provide a good overview of the
treatment of acute appendicitis and the classification of abdominal
hernias, but could benefit from additional information about the use
of antibiotics in case of a perforated appendix.
```

- For the second query, the model has done well with the set of parameters which uses lower values of top_p(0.8) and top_k(40).

```
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?"
ground, rel = generate_ground_relevance_response(query_3, 4, 1024,
0.01, 0.95, 60)

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

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

  Sure, I'd be happy to help you evaluate the answer! Here are the
steps I will follow:

Step 1: Identify the metric that needs to be followed.
The metric is "the answer should be derived only from the information
presented in the context."

Step 2: Evaluate the answer to see if it adheres to the metric.
The answer provides a list of effective treatments for sudden patchy
hair loss, but it does not derive any information from the context.
The answer is based on general knowledge and does not take into
account the specific information provided in the context. Therefore,
the answer does not adhere to the metric.

Step 3: Evaluate the extent to which the metric is followed.
The answer does not follow the metric at all. The answer provides a
list of treatments that are not specifically related to the
information presented in the context.

Based on the above evaluation, I would rate the answer as follows:

Rating: 1 (The metric is not followed at all)

Please let me know if you have any further questions or if there's
anything else I can help with!

  Sure, I can help you evaluate the context and answer based on the
given criteria. Here are the steps to follow:

Step 1: Evaluate the relevance of the answer to the question.

The question asks about effective treatments or solutions for sudden
patchy hair loss, and the answer provides a list of medications and
therapies that can be used to treat the condition. The answer is
relevant to the question as it addresses the main aspects of the
question, which is the treatment of sudden patchy hair loss.

Step 2: Evaluate the extent to which the context adheres to the
metric.

The context provides information about the possible causes of sudden
patchy hair loss and the treatments that can be used to address the
condition. The information is presented in a clear and concise manner,
and it covers all the important aspects of the question. Therefore,
the context adheres well to the metric.

Step 3: Evaluate the answer based on the evaluation criteria.
```

```
Based on the given criteria, I would rate the answer as follows:

* Relevance: 5 (The answer is completely relevant to the question and
addresses all the main

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?"
ground, rel = generate_ground_relevance_response(query_3, 8, 1024,
0.01, 0.95, 60)

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

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

  Sure, I'd be happy to help you evaluate the answer! Here are the
steps I will follow:

Step 1: Identify the metric that needs to be followed.
The metric is "the answer should be derived only from the information
presented in the context."

Step 2: Evaluate the answer to see if it adheres to the metric.
The answer provides a list of effective treatments for sudden patchy
hair loss, but it does not derive any information from the context.
The answer is based on general knowledge and does not take into
account the specific information provided in the context. Therefore,
the answer does not adhere to the metric.

Step 3: Evaluate the extent to which the metric is followed.
The answer does not follow the metric at all. The answer provides a
list of treatments that are not specifically related to the
information presented in the context.

Based on the above evaluation, I would rate the answer as follows:

Rating: 1 (The metric is not followed at all)

Please let me know if you have any further questions or if there's
anything else I can help with!

  Sure, I can help you evaluate the context and answer based on the
given criteria. Here are the steps to follow:

Step 1: Evaluate the relevance of the answer to the question.

The question asks about effective treatments or solutions for sudden
```

```
patchy hair loss, and the answer provides a list of medications and
therapies that can be used to treat the condition. The answer is
relevant to the question as it addresses the main aspects of the
question, which is the treatment of sudden patchy hair loss.

Step 2: Evaluate the extent to which the context adheres to the
metric.

The context provides information about the possible causes of sudden
patchy hair loss and the treatments that can be used to address the
condition. The information is presented in a clear and concise manner,
and it covers all the important aspects of the question. Therefore,
the context adheres well to the metric.

Step 3: Evaluate the answer based on the evaluation criteria.

Based on the given criteria, I would rate the answer as follows:

* Relevance: 5 (The answer is completely relevant to the question and
addresses all the main
```

- For the above query, the model's answer is rated high in Relavence but low in groundedness.

```
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?"
ground, rel = generate_ground_relevance_response(query_4, 2, 1024, 0,
0.8, 40)

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

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

  Sure, I'd be happy to help you evaluate the answer!

Step 1: Evaluate the answer against the metric.

The metric is that the answer should be derived only from the
information presented in the context.

Based on the answer provided, it appears that the answerer has derived
the recommended treatments for a person who has sustained a physical
injury to brain tissue from the information provided in Chapter 324 of
The Merck Manual of Diagnosis & Therapy, 19th Edition. The answer
mentions supportive care, good nutrition, and prevention of pressure
ulcers as recommended treatments, which are all consistent with the
information presented in the context.
```

Step 2: Evaluate the extent to which the metric is followed.

The answerer has followed the metric completely, as all the recommended treatments mentioned in the answer are based on the information provided in the context.

Step 3: Rate the answer using the evaluation criteria.

Based on the evaluation criteria provided, I would rate the answer as a 5, as the answerer has followed the metric completely and derived the recommended treatments only from the information presented in the context.

Here's the final rating:

Rating: 5 (Completely Followed)

The answerer has provided a complete and accurate response to the question, deriving all the recommended treatments for a person who has sustained a physical injury to brain tissue from the information presented in the context. The answer is well-supported by the information provided in Chapter 324 of The Merck Manual of Diagnosis & Therapy, 19th Edition, and demonstrates a thorough understanding of the topic.

Sure, I'd be happy to help you evaluate the context and answer!

Step 1: Evaluate the context as per the metric.

The context provided is a passage from The Merck Manual of Diagnosis & Therapy, 19th Edition, specifically from Chapter 324, which discusses traumatic brain injury (TBI). The passage provides a comprehensive overview of TBI, including its definition, causes, pathology, clinical manifestations, and treatment options.

Step 2: Consider whether all and only the important aspects are contained in the answer.

The answer provided is a list of recommended treatments for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function. The treatments listed include supportive care, good nutrition, and prevention of pressure ulcers. These treatments are all relevant to the context of TBI and are important aspects of managing the condition.

Step 3: Evaluate the extent to which the metric is followed.

The answer provided follows the metric well, as it lists specific treatments that are relevant to the context of TBI. The answer also provides a brief explanation for each treatment, which helps to

```
reinforce the relevance of the treatments to the context.

Step 4: Use the previous information to rate the context using the
evaluation criteria.

Based on the information provided, I would rate the context as
follows:

* Relevance: 5 (completely followed) - The answer provides specific
treatments that are relevant to the context of TBI, and the
explanations provided reinforce the relevance of the treatments.
* Completeness: 4 (mostly followed) - The answer provides a
comprehensive list of treatments for TBI, but does not cover all
aspects of the condition, such as rehabilitation and long-term
outcomes.
* Accuracy: 5 (completely accurate) - The information provided in the
answer is accurate and based on a reliable source, The Merck Manual of
Diagnosis & Therapy, 19th Edition.

Overall, I would rate the context as 4.5 out of 5, as it provides a
comprehensive list of treatments for TBI that are relevant to the
context, and the information is accurate and complete.

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?"
ground, rel = generate_ground_relevance_response(query_4, 6, 1024, 0,
0.8, 40)

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

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

  Sure, I'd be happy to help you evaluate the answer!

Step 1: Evaluate the answer against the metric.

The metric is that the answer should be derived only from the
information presented in the context.

Based on the answer provided, it appears that the answerer has derived
the recommended treatments for a person who has sustained a physical
injury to brain tissue from the information provided in Chapter 324 of
The Merck Manual of Diagnosis & Therapy, 19th Edition. The answer
mentions supportive care, good nutrition, and prevention of pressure
ulcers as recommended treatments, which are all consistent with the
information presented in the context.
```

Step 2: Evaluate the extent to which the metric is followed.

The answerer has followed the metric completely, as all the recommended treatments mentioned in the answer are based on the information provided in the context.

Step 3: Rate the answer using the evaluation criteria.

Based on the evaluation criteria provided, I would rate the answer as a 5, as the answerer has followed the metric completely and derived the recommended treatments only from the information presented in the context.

Here's the final rating:

Rating: 5 (Completely Followed)

The answerer has provided a complete and accurate response to the question, deriving all the recommended treatments for a person who has sustained a physical injury to brain tissue from the information presented in the context. The answer is well-supported by the information provided in Chapter 324 of The Merck Manual of Diagnosis & Therapy, 19th Edition, and demonstrates a thorough understanding of the topic.

  Sure, I'd be happy to help you evaluate the context and answer!

Step 1: Evaluate the context as per the metric.

The context provided is a passage from The Merck Manual of Diagnosis & Therapy, 19th Edition, specifically from Chapter 324, which discusses traumatic brain injury (TBI). The passage provides a comprehensive overview of TBI, including its definition, causes, pathology, clinical manifestations, and treatment options.

Step 2: Consider whether all and only the important aspects are contained in the answer.

The answer provided is a list of recommended treatments for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function. The treatments listed include supportive care, good nutrition, and prevention of pressure ulcers. These treatments are all relevant to the context of TBI and are important aspects of managing the condition.

Step 3: Evaluate the extent to which the metric is followed.

The answer provided follows the metric well, as it lists specific treatments that are relevant to the context of TBI. The answer also provides a brief explanation for each treatment, which helps to

```
reinforce the relevance of the treatments to the context.

Step 4: Use the previous information to rate the context using the
evaluation criteria.

Based on the information provided, I would rate the context as
follows:

* Relevance: 5 (completely followed) - The answer provides specific
treatments that are relevant to the context of TBI, and the
explanations provided reinforce the relevance of the treatments.
* Completeness: 4 (mostly followed) - The answer provides a
comprehensive list of treatments for TBI, but does not cover all
aspects of the condition, such as rehabilitation and long-term
outcomes.
* Accuracy: 5 (completely accurate) - The information provided in the
answer is accurate and based on a reliable source, The Merck Manual of
Diagnosis & Therapy, 19th Edition.

Overall, I would rate the context as 4.5 out of 5, as it provides a
comprehensive list of treatments for TBI that are relevant to the
context, and the information is accurate and complete.
```

- The model has done well in answering query 4 with both set of parameters.

```
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?"
ground, rel = generate_ground_relevance_response(query_5, 3, 1024, 0,
0.7, 35)

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

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

  Sure, I'd be happy to help you evaluate the answer!

First, let me summarize the question and context:

Question: 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?

Context: The patient has fractured their leg during a hiking trip.

Now, let me evaluate the answer based on the metric you provided:

1. The metric is followed to a good extent: The answer provides a list
```

```
of necessary precautions and treatment steps for a person who

  Sure, I'd be happy to help you rate the context!

To evaluate the extent to which the metric is followed, we need to
consider the following steps:

1. Relevance: Does the answer address the main aspects of the
question?
2. Completeness: Is all relevant information included in the answer?
3. Accuracy: Are the facts and information provided accurate?
4. Clarity: Is the answer clear and easy to understand?

Based on the provided
```

```
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?"
ground, rel = generate_ground_relevance_response(query_5, 5, 1024,
0.09, 0.95, 50)

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

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

  Sure, I'd be happy to help you rate the answer!

To evaluate the extent to which the metric is followed, we need to
consider the following steps:

1. The answer should be derived only from the information presented in
the context.
2. The answer should address all aspects of the question.
3. The answer should be concise and to the point.

Based on the provided answer, here's my evaluation:

Step 1: The answer is derived only from the information presented in
the context. Yes, the answer provides a detailed list of necessary
precautions and treatment steps for a person who has

  Sure, I'd be happy to help you rate the context!

To evaluate the extent to which the metric is followed, we need to
consider the following steps:

1. Relevance: Does the answer address the main aspects of the
question?
```

```
2. Completeness: Is all relevant information included in the answer?
3. Accuracy: Are the facts and information provided accurate?
4. Clarity: Is the answer clear and easy to understand?

Based on the provided context
```

- The model has also answered the query 5 with good score but the explanation is not clearly provided.

Conclusions and Recommendations

- The LLM model provides information using the general data from the internet over which the Llama-2 model was trained.
- Using prompt Engineering, we can shape the model's output answer's as per our convenience, especially in formatting and getting clear responses.
- The RAG model provides information based on the text document which was fed to the LLM model.
- The same method can be used on other models like Mistral, etc. and check if those models do better.
- With higher computational power, better models can be built in lesser amount of time.
- Restricting the top_p and top_k value has made the model's answers more concise.
- The LLM model can not be fully trusted since it is prone to hallucination.
- RAG model performed better than the LLM since the answer is derived from the context.
- Different chunking, embedding and retrieval strategies can be used to check if further improvement can be made.
- Further research is needed for improving the groundedness and relavance of the RAG model.

```

```