```
Start coding or generate with AI.
```

```
!pip install langchain sentence-transformers chromadb llama-cpp-python langchain_community pypdf
```

```
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.12/dist-packages (from scikit-learn->sentence-trans
Requirement already satisfied: cachetools<6.0,>=2.0.0 in /usr/local/lib/python3.12/dist-packages (from google-auth>=1.0.1->kubern
Requirement already satisfied: pyasn1-modules>=0.2.1 in /usr/local/lib/python3.12/dist-packages (from google-auth>=1.0.1->kuberne
Requirement already satisfied: rsa<5,>=3.1.4 in /usr/local/lib/python3.12/dist-packages (from google-auth>=1.0.1->kubernetes>=28.
Requirement already satisfied: zipp>=3.20 in /usr/local/lib/python3.12/dist-packages (from importlib-metadata<8.8.0,>=6.0->opente
Requirement already satisfied: jsonpointer>=1.9 in /usr/local/lib/python3.12/dist-packages (from jsonpatch<2.0,>=1.33->langchain-
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.12/dist-packages (from markdown-it-py>=2.2.0->rich>=10.11.0->
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.12/dist-packages (from sympy->onnxruntime>=1.14.1->ch
Collecting mypy-extensions>=0.3.0 (from typing-inspect<1,>=0.4.0->dataclasses-json<0.7,>=0.6.7->langchain_community)
  Downloading mypy_extensions-1.1.0-py3-none-any.whl.metadata (1.1 kB)
Requirement already satisfied: sniffio>=1.1 in /usr/local/lib/python3.12/dist-packages (from anyio->httpx>=0.27.0->chromadb) (1.3
Collecting humanfriendly>=9.1 (from coloredlogs->onnxruntime>=1.14.1->chromadb)
  Downloading humanfriendly-10.0-py2.py3-none-any.whl.metadata (9.2 kB)
Requirement already satisfied: pyasn1<0.7.0,>=0.6.1 in /usr/local/lib/python3.12/dist-packages (from pyasn1-modules>=0.2.1->googl
Downloading chromadb-1.0.20-cp39-abi3-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (19.8 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 19.8/19.8 MB 118.1 MB/s eta 0:00:00
Downloading langchain_community-0.3.29-py3-none-any.whl (2.5 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.5/2.5 MB 97.8 MB/s eta 0:00:00
Downloading pypdf-6.0.0-py3-none-any.whl (310 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 310.5/310.5 kB 31.0 MB/s eta 0:00:00
Downloading bcrypt-4.3.0-cp39-abi3-manylinux_2_34_x86_64.whl (284 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 284.2/284.2 kB 28.4 MB/s eta 0:00:00
Downloading dataclasses_json-0.6.7-py3-none-any.whl (28 kB)
Downloading diskcache-5.6.3-py3-none-any.whl (45 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 45.5/45.5 kB 4.4 MB/s eta 0:00:00
Downloading kubernetes-33.1.0-py2.py3-none-any.whl (1.9 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.9/1.9 MB 95.4 MB/s eta 0:00:00
Downloading mmh3-5.2.0-cp312-cp312-manylinux1_x86_64.manylinux_2_28_x86_64.manylinux_2_5_x86_64.whl (103 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 103.3/103.3 kB 10.5 MB/s eta 0:00:00
Downloading onnxruntime-1.22.1-cp312-cp312-manylinux_2_27_x86_64.manylinux_2_28_x86_64.whl (16.5 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 16.5/16.5 MB 92.7 MB/s eta 0:00:00
Downloading opentelemetry_exporter_otlp_proto_grpc-1.36.0-py3-none-any.whl (18 kB)
Downloading opentelemetry_exporter_otlp_proto_common-1.36.0-py3-none-any.whl (18 kB)
Downloading opentelemetry_proto-1.36.0-py3-none-any.whl (72 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 72.5/72.5 kB 7.8 MB/s eta 0:00:00
Downloading posthog-5.4.0-py3-none-any.whl (105 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 105.4/105.4 kB 10.8 MB/s eta 0:00:00
Downloading pybase64-1.4.2-cp312-cp312-manylinux1_x86_64.manylinux2014_x86_64.manylinux_2_17_x86_64.manylinux_2_5_x86_64.whl (71
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 71.6/71.6 kB 7.3 MB/s eta 0:00:00
Downloading requests-2.32.5-py3-none-any.whl (64 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 64.7/64.7 kB 5.4 MB/s eta 0:00:00
Downloading backoff-2.2.1-py3-none-any.whl (15 kB)
Downloading durationpy-0.10-py3-none-any.whl (3.9 kB)
Downloading httptools-0.6.4-cp312-cp312-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_64.whl (51
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 510.8/510.8 kB 37.0 MB/s eta 0:00:00
Downloading marshmallow-3.26.1-py3-none-any.whl (50 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 50.9/50.9 kB 4.9 MB/s eta 0:00:00
Downloading typing_inspect-0.9.0-py3-none-any.whl (8.8 kB)
Downloading uvloop-0.21.0-cp312-cp312-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (4.7 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 4.7/4.7 MB 92.3 MB/s eta 0:00:00
Downloading watchfiles-1.1.0-cp312-cp312-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (452 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 452.2/452.2 kB 35.6 MB/s eta 0:00:00
Downloading coloredlogs-15.0.1-py2.py3-none-any.whl (46 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 46.0/46.0 kB 5.2 MB/s eta 0:00:00
Downloading humanfriendly-10.0-py2.py3-none-any.whl (86 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 86.8/86.8 kB 8.1 MB/s eta 0:00:00
Downloading mypy_extensions-1.1.0-py3-none-any.whl (5.0 kB)
Building wheels for collected packages: llama-cpp-python, pypika
```

```
from langchain_community.document_loaders import PyPDFDirectoryLoader
from langchain.text_splitter import CharacterTextSplitter,RecursiveCharacterTextSplitter
from langchain_community.embeddings import HuggingFaceEmbeddings
from langchain.vectorstores import FAISS, Chroma
from langchain_community.llms import LlamaCpp
from langchain.chains import RetrievalQA, LLMChain
```

```
import pathlib
import textwrap
from IPython.display import display
from IPython.display import Markdown



def to_markdown(text):
  text = text.replace('•', '  *')
  return Markdown(textwrap.indent(text, '> ', predicate=lambda _: True))
```

```
# Used to securely store your API key
from google.colab import userdata
```

```
import os
from getpass import getpass

HUGGINGFACEHUB_API_TOKEN = userdata.get("HUGGINGFACEHUB_API_TOKEN")
os.environ["HUGGINGFACEHUB_API_TOKEN"] = "HUGGINGFACEHUB_API_TOKEN"
```

```
loader = PyPDFDirectoryLoader("/content/sample_data/Data")
docs = loader.load()
```

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

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

```
docs
```

```
[Document(metadata={'producer': 'PDFlib+PDI 5.0.0 (SunOS)', 'creator': 'PyPDF', 'creationdate': '2004-12-18T17:00:02-05:00',
'moddate': '2004-12-18T16:15:31-06:00', 'source': '/content/sample_data/Data/Medical_book (1).pdf', 'total_pages': 637, 'page':
0, 'page_label': '1'}, page_content=''),
 Document(metadata={'producer': 'PDFlib+PDI 5.0.0 (SunOS)', 'creator': 'PyPDF', 'creationdate': '2004-12-18T17:00:02-05:00',
'moddate': '2004-12-18T16:15:31-06:00', 'source': '/content/sample_data/Data/Medical_book (1).pdf', 'total_pages': 637, 'page':
1, 'page_label': '2'}, page_content='The GALE\nENCYCLOPEDIA\nof MEDICINE\nSECOND EDITION'),
 Document(metadata={'producer': 'PDFlib+PDI 5.0.0 (SunOS)', 'creator': 'PyPDF', 'creationdate': '2004-12-18T17:00:02-05:00',
'moddate': '2004-12-18T16:15:31-06:00', 'source': '/content/sample_data/Data/Medical_book (1).pdf', 'total_pages': 637, 'page':
2, 'page_label': '3'}, page_content='The GALE\nENCYCLOPEDIA\nof MEDICINE\nSECOND EDITION\nJACQUELINE L. LONGE, EDITOR\nDEIRDRE S.
BLANCHFIELD, ASSOCIATE EDITOR\nVOLUME\nA-B\n1'),
 Document(metadata={'producer': 'PDFlib+PDI 5.0.0 (SunOS)', 'creator': 'PyPDF', 'creationdate': '2004-12-18T17:00:02-05:00',
'moddate': '2004-12-18T16:15:31-06:00', 'source': '/content/sample_data/Data/Medical_book (1).pdf', 'total_pages': 637, 'page':
3, 'page_label': '4'}, page_content='STAFF\nJacqueline L. Longe, Project Editor\nDeirdre S. Blanchfield, Associate
Editor\nChristine B. Jeryan, Managing Editor\nDonna Olendorf, Senior Editor\nStacey Blachford, Associate Editor\nKate
Kretschmann, Melissa C. McDade, Ryan\nThomason, Assistant Editors\nMark Springer, Technical Specialist\nAndrea Lopeman,
Programmer/Analyst\nBarbara J. Yarrow,Manager, Imaging and Multimedia\nContent\nRobyn V . Young,Project Manager, Imaging
and\nMultimedia Content\nDean Dauphinais, Senior Editor, Imaging and\nMultimedia Content\nKelly A. Quin, Editor, Imaging and
Multimedia Content\nLeitha Etheridge-Sims, Mary K. Grimes, Dave Oblender,\nImage Catalogers\nPamela A. Reed, Imaging
Coordinator\nRandy Bassett, Imaging Supervisor\nRobert Duncan, Senior Imaging Specialist\nDan Newell, Imaging
Specialist\nChristine O’Bryan,Graphic Specialist\nMaria Franklin, Permissions Manager\nMargaret A. Chamberlain, Permissions
Specialist\nMichelle DiMercurio, Senior Art Director\nMike Logusz, Graphic Artist\nMary Beth Trimper,Manager, Composition
and\nElectronic Prepress\nEvi Seoud, Assistant Manager, Composition Purchasing\nand Electronic Prepress\nDorothy Maki,
Manufacturing Manager\nWendy Blurton, Senior Manufacturing Specialist\nThe GALE\nENCYCLOPEDIA\nof MEDICINE\nSECOND EDITION\nSince
this page cannot legibly accommodate all copyright notices, the\nacknowledgments constitute an extension of the copyright
notice.\nWhile every effort has been made to ensure the reliability of the infor-\nmation presented in this publication, the Gale
Group neither guarantees\nthe accuracy of the data contained herein nor assumes any responsibili-\nty for errors, omissions or
discrepancies. The Gale Group accepts no\npayment for listing, and inclusion in the publication of any organiza-\ntion, agency,
institution, publication, service, or individual does not\nimply endorsement of the editor or publisher. Errors brought to
the\nattention of the publisher and verified to the satisfaction of the publish-\ner will be corrected in future editions.\nThis
book is printed on recycled paper that meets Environmental Pro-\ntection Agency standards.\nThe paper used in this publication
meets the minimum requirements of\nAmerican National Standard for Information Sciences-Permanence\nPaper for Printed Library
Materials, ANSI Z39.48-1984.\nThis publication is a creative work fully protected by all applicable\ncopyright laws, as well as
by misappropriation, trade secret, unfair com-\npetition, and other applicable laws. The authors and editor of this work\nhave
added value to the underlying factual material herein through one\nor more of the following: unique and original selection,
coordination,\nexpression, arrangement, and classification of the information.\nGale Group and design is a trademark used herein
under license.\nAll rights to this publication will be vigorously defended.\nCopyright © 2002\nGale Group\n27500 Drake
Road\nFarmington Hills, MI 48331-3535\nAll rights reserved including the right of reproduction in whole or in\npart in any
form.\nISBN 0-7876-5489-2 (set)\n0-7876-5490-6 (V ol. 1)\n0-7876-5491-4 (V ol. 2)\n0-7876-5492-2 (V ol. 3)\n0-7876-5493-0 (V ol.
4)\n0-7876-5494-9 (V ol. 5)\nPrinted in the United States of America\n10 9 8 7 6 5 4 3 2 1\nLibrary of Congress Cataloging-in-
Publication Data\nGale encyclopedia of medicine / Jacqueline L. Longe, editor;\nDeirdre S. Blanchfield, associate editor — 2nd
ed.\np. cm.\nIncludes bibliographical references and index.\nContents: V ol. 1. A-B — v. 2. C-F — v. 3.\nG-M — v. 4. N-S — v. 5.
T-Z.\nISBN 0-7876-5489-2 (set: hardcover) — ISBN 0-7876-5490-6\n(vol. 1) — ISBN 0-7876-5491-4 (vol. 2) — ISBN 0-7876-5492-
2\n(vol. 3) — ISBN 0-7876-5493-0 (vol. 4) — ISBN 0-7876-5494-9\n(vol. 5)\n1. Internal medicine—Encyclopedias. I. Longe,
Jacqueline L. \nII. Blanchfield, Deirdre S. III. Gale Research Company.\nRC41.G35 2001\n616’.003—dc21\n2001051245'),
 Document(metadata={'producer': 'PDFlib+PDI 5.0.0 (SunOS)', 'creator': 'PyPDF', 'creationdate': '2004-12-18T17:00:02-05:00',
'moddate': '2004-12-18T16:15:31-06:00', 'source': '/content/sample_data/Data/Medical_book (1).pdf', 'total_pages': 637, 'page':
4, 'page_label': '5'}, page_content='Introduction.................................................... ix\nAdvisory
Board.......................................... xi\nContributors ................................................
xiii\nEntries\nVolume 1: A-B.......................................... 1\nVolume 2: C-
F................................... 625\nVolume 3: G-M.................................... 1375\nVolume 4: N-
S.................................... 2307\nVolume 5: T-Z..................................... 3237\nOrganizations
......................................... 3603\nGeneral Index.......................................... 3625\nGALE
ENCYCLOPEDIA OF MEDICINE 2 V\nCONTENTS'),
 Document(metadata={'producer': 'PDFlib+PDI 5.0.0 (SunOS)', 'creator': 'PyPDF', 'creationdate': '2004-12-18T17:00:02-05:00',
'moddate': '2004-12-18T16:15:31-06:00', 'source': '/content/sample_data/Data/Medical_book (1).pdf', 'total_pages': 637, 'page':
5, 'page_label': '6'}, page_content='The Gale Encyclopedia of Medicine 2is a medical ref-\nerence product designed to inform and
educate readers\nabout a wide variety of disorders, conditions, treatments,\nand diagnostic tests. The Gale Group believes the
```

```
text_splitter = RecursiveCharacterTextSplitter(chunk_size=300, chunk_overlap=50)
chunks = text_splitter.split_documents(docs)
```

```
len(chunks)
```

```
10170
```

```
chunks[0]
```

```
Document(metadata={'producer': 'PDFlib+PDI 5.0.0 (SunOS)', 'creator': 'PyPDF', 'creationdate': '2004-12-18T17:00:02-05:00',
'moddate': '2004-12-18T16:15:31-06:00', 'source': '/content/sample_data/Data/Medical_book (1).pdf', 'total_pages': 637, 'page': 1,
'page_label': '2'}, page_content='The GALE\nENCYCLOPEDIA\nof MEDICINE\nSECOND EDITION')
```

```
chunks[1]
```

```
Document(metadata={'producer': 'PDFlib+PDI 5.0.0 (SunOS)', 'creator': 'PyPDF', 'creationdate': '2004-12-18T17:00:02-05:00',
'moddate': '2004-12-18T16:15:31-06:00', 'source': '/content/sample_data/Data/Medical_book (1).pdf', 'total_pages': 637, 'page': 2,
'page_label': '3'}, page_content='The GALE\nENCYCLOPEDIA\nof MEDICINE\nSECOND EDITION\nJACQUELINE L. LONGE, EDITOR\nDEIRDRE S.
BLANCHFIELD, ASSOCIATE EDITOR\nVOLUME\nA-B\n1')
```

```
chunks[2]
```

```
Document(metadata={'producer': 'PDFlib+PDI 5.0.0 (SunOS)', 'creator': 'PyPDF', 'creationdate': '2004-12-18T17:00:02-05:00',
'moddate': '2004-12-18T16:15:31-06:00', 'source': '/content/sample_data/Data/Medical_book (1).pdf', 'total_pages': 637, 'page': 3,
'page_label': '4'}, page_content='STAFF\nJacqueline L. Longe, Project Editor\nDeirdre S. Blanchfield, Associate Editor\nChristine
B. Jeryan, Managing Editor\nDonna Olendorf, Senior Editor\nStacey Blachford, Associate Editor\nKate Kretschmann, Melissa C.
McDade, Ryan\nThomason, Assistant Editors\nMark Springer, Technical Specialist')
```

```
chunks[3]
```

```
Document(metadata={'producer': 'PDFlib+PDI 5.0.0 (SunOS)', 'creator': 'PyPDF', 'creationdate': '2004-12-18T17:00:02-05:00',
'moddate': '2004-12-18T16:15:31-06:00', 'source': '/content/sample_data/Data/Medical_book (1).pdf', 'total_pages': 637, 'page': 3,
'page_label': '4'}, page_content='Mark Springer, Technical Specialist\nAndrea Lopeman, Programmer/Analyst\nBarbara J.
Yarrow,Manager, Imaging and Multimedia\nContent\nRobyn V . Young,Project Manager, Imaging and\nMultimedia Content\nDean
Dauphinais, Senior Editor, Imaging and\nMultimedia Content')
```

```
chunks[4]
```

```
Document(metadata={'producer': 'PDFlib+PDI 5.0.0 (SunOS)', 'creator': 'PyPDF', 'creationdate': '2004-12-18T17:00:02-05:00',
'moddate': '2004-12-18T16:15:31-06:00', 'source': '/content/sample_data/Data/Medical_book (1).pdf', 'total_pages': 637, 'page': 3,
'page_label': '4'}, page_content='Multimedia Content\nKelly A. Quin, Editor, Imaging and Multimedia Content\nLeitha Etheridge-
Sims, Mary K. Grimes, Dave Oblender,\nImage Catalogers\nPamela A. Reed, Imaging Coordinator\nRandy Bassett, Imaging
Supervisor\nRobert Duncan, Senior Imaging Specialist\nDan Newell, Imaging Specialist')
```

```
chunks[5]
```

```
Document(metadata={'producer': 'PDFlib+PDI 5.0.0 (SunOS)', 'creator': 'PyPDF', 'creationdate': '2004-12-18T17:00:02-05:00',
'moddate': '2004-12-18T16:15:31-06:00', 'source': '/content/sample_data/Data/Medical_book (1).pdf', 'total_pages': 637, 'page': 3,
'page_label': '4'}, page_content='Dan Newell, Imaging Specialist\nChristine O'Bryan,Graphic Specialist\nMaria Franklin,
Permissions Manager\nMargaret A. Chamberlain, Permissions Specialist\nMichelle DiMercurio, Senior Art Director\nMike Logusz,
Graphic Artist\nMary Beth Trimper,Manager, Composition and\nElectronic Prepress')
```

```
embeddings = HuggingFaceEmbeddings(model_name="BAAI/bge-base-en-v1.5")
```

```
/tmp/ipython-input-2576288788.py:1: LangChainDeprecationWarning: The class `HuggingFaceEmbeddings` was deprecated in LangChain 0.2.
  embeddings = HuggingFaceEmbeddings(model_name="BAAI/bge-base-en-v1.5")
/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning: 
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it as
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(
```

```
modules.json: 100%                                   349/349 [00:00<00:00, 35.4kB/s]

config_sentence_transformers.json: 100%              124/124 [00:00<00:00, 8.71kB/s]

README.md:    94.6k/? [00:00<00:00, 6.81MB/s]

sentence_bert_config.json: 100%                      52.0/52.0 [00:00<00:00, 6.44kB/s]

config.json: 100%                                    777/777 [00:00<00:00, 90.8kB/s]

model.safetensors: 100%                              438M/438M [00:01<00:00, 294MB/s]

tokenizer_config.json: 100%                          366/366 [00:00<00:00, 25.6kB/s]

vocab.txt:    232k/? [00:00<00:00, 10.2MB/s]

tokenizer.json:    711k/? [00:00<00:00, 36.1MB/s]

special_tokens_map.json: 100%                        125/125 [00:00<00:00, 12.7kB/s]

config.json: 100%                                    190/190 [00:00<00:00, 16.9kB/s]
```

```
vectorstore = Chroma.from_documents(chunks, embeddings)
```

```
query = "What is fever?"
search = vectorstore.similarity_search(query)
```

```
to_markdown(search[0].page_content)
```

extensive tissue destruction (necrosis).

- Bloodstream. Bloodstream invasion causes high fever (up to 105°F [40.6°C]), chills, a general ill feeling, and is potentially fatal. Diagnosis The diagnosis of anaerobic infection is based pri- marily on symptoms, the patient’s medical history, and

```
retriever = vectorstore.as_retriever(
    search_kwargs={'k': 5}
)
```

```
retriever.get_relevant_documents(query)
```

```
/tmp/ipython-input-3521827203.py:1: LangChainDeprecationWarning: The method `BaseRetriever.get_relevant_documents` was deprecated i
  retriever.get_relevant_documents(query)
[Document(metadata={'page': 181, 'creator': 'PyPDF', 'moddate': '2004-12-18T16:15:31-06:00', 'page_label': '182', 'source': 
'/content/sample_data/Data/Medical_book (1).pdf', 'total_pages': 637, 'producer': 'PDFlib+PDI 5.0.0 (SunOS)', 'creationdate': 
'2004-12-18T17:00:02-05:00'}, page_content='extensive tissue destruction (necrosis).\n• Bloodstream. Bloodstream invasion causes
high fever\n(up to 105°F [40.6°C]), chills, a general ill feeling, and\nis potentially fatal.\nDiagnosis\nThe diagnosis of
anaerobic infection is based pri-\nmarily on symptoms, the patient’s medical history, and'),
 Document(metadata={'source': '/content/sample_data/Data/Medical_book (1).pdf', 'producer': 'PDFlib+PDI 5.0.0 (SunOS)', 
'creationdate': '2004-12-18T17:00:02-05:00', 'creator': 'PyPDF', 'page_label': '343', 'total_pages': 637, 'moddate': '2004-12-
18T16:15:31-06:00', 'page': 342}, page_content='24.\nKEY TERMS\nRheumatic fever —A disease believed to be\ncaused by a bacterium
named group A streptococ-\ncus. This bacterium causes a sore “strep throat”\nand can also result in fever. Infection by this bac-
\nterium can also damage the heart and its valves,'),
 Document(metadata={'creationdate': '2004-12-18T17:00:02-05:00', 'creator': 'PyPDF', 'page_label': '323', 'moddate': '2004-12-
18T16:15:31-06:00', 'total_pages': 637, 'page': 322, 'source': '/content/sample_data/Data/Medical_book (1).pdf', 'producer': 
'PDFlib+PDI 5.0.0 (SunOS)'}, page_content='Heat stroke—A serious condition that results from\nexposure to extreme heat. The body
loses its abili-\nty to cool itself. Severe headache, high fever, and\nhot, dry skin may result. In severe cases, a person\nwith
heat stroke may collapse or go into a coma.'),
 Document(metadata={'producer': 'PDFlib+PDI 5.0.0 (SunOS)', 'creationdate': '2004-12-18T17:00:02-05:00', 'page': 27, 'page_label':
'28', 'moddate': '2004-12-18T16:15:31-06:00', 'creator': 'PyPDF', 'total_pages': 637, 'source': 
'/content/sample_data/Data/Medical_book (1).pdf'}, page_content='Septicemia —The spread of an infectious agent\nthroughout the
body by means of the blood\nstream.\nSinus—A tubular channel connecting one body\npart with another or with the outside.\nlocate
the site of an abscess, but usually something in the'),
 Document(metadata={'total_pages': 637, 'source': '/content/sample_data/Data/Medical_book (1).pdf', 'creator': 'PyPDF', 
'producer': 'PDFlib+PDI 5.0.0 (SunOS)', 'page_label': '61', 'creationdate': '2004-12-18T17:00:02-05:00', 'page': 60, 'moddate': 
'2004-12-18T16:15:31-06:00'}, page_content='are not growing and are in a resting state. Alternatively, a\n“broad spectrum”
antibiotic may be used which would\nkill many different kinds of bacteria.\nAspirin or other medications which reduce the
pain\nand the fever may also be given. Medications which')]
```

```
from google.colab import drive
```

```
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
llm = LlamaCpp(
    model_path= "/content/drive/MyDrive/MediMate/MediMate Model/BioMistral-7B.Q4_K_M.gguf",
    temperature=0.3,
    max_tokens=2048,
    top_p=1)
```

```
llama_model_loader: loaded meta data with 21 key-value pairs and 291 tensors from /content/drive/MyDrive/MediMate/MediMate Model/
llama_model_loader: Dumping metadata keys/values. Note: KV overrides do not apply in this output.
llama_model_loader: - kv   0:                       general.architecture str              = llama
llama_model_loader: - kv   1:                               general.name str              = hub
llama_model_loader: - kv   2:                       llama.context_length u32              = 32768
llama_model_loader: - kv   3:                     llama.embedding_length u32              = 4096
llama_model_loader: - kv   4:                          llama.block_count u32              = 32
llama_model_loader: - kv   5:                  llama.feed_forward_length u32              = 14336
llama_model_loader: - kv   6:                 llama.rope.dimension_count u32              = 128
llama_model_loader: - kv   7:                 llama.attention.head_count u32              = 32
llama_model_loader: - kv   8:              llama.attention.head_count_kv u32              = 8
llama_model_loader: - kv   9:     llama.attention.layer_norm_rms_epsilon f32              = 0.000010
llama_model_loader: - kv  10:                       llama.rope.freq_base f32              = 10000.000000
llama_model_loader: - kv  11:                          general.file_type u32              = 15
llama_model_loader: - kv  12:                       tokenizer.ggml.model str              = llama
llama_model_loader: - kv  13:                      tokenizer.ggml.tokens arr[str,32000]   = ["<unk>", "<s>", "</s>", "<0x00>", "<
llama_model_loader: - kv  14:                      tokenizer.ggml.scores arr[f32,32000]   = [0.000000, 0.000000, 0.000000, 0.0000
llama_model_loader: - kv  15:                  tokenizer.ggml.token_type arr[i32,32000]   = [2, 3, 3, 6, 6, 6, 6, 6, 6, 6, 6, 6,
llama_model_loader: - kv  16:                tokenizer.ggml.bos_token_id u32              = 1
llama_model_loader: - kv  17:                tokenizer.ggml.eos_token_id u32              = 2
llama_model_loader: - kv  18:            tokenizer.ggml.unknown_token_id u32              = 0
llama_model_loader: - kv  19:                    tokenizer.chat_template str              = {{ bos_token }}{% for message in mess
llama_model_loader: - kv  20:               general.quantization_version u32              = 2
llama_model_loader: - type  f32:   65 tensors
llama_model_loader: - type q4_K:  193 tensors
llama_model_loader: - type q6_K:   33 tensors
print_info: file format = GGUF V3 (latest)
print_info: file type   = Q4_K - Medium
print_info: file size   = 4.07 GiB (4.83 BPW)
init_tokenizer: initializing tokenizer for type 1
load: control token:      2 '</s>' is not marked as EOG
load: control token:      1 '<s>' is not marked as EOG
load: special_eos_id is not in special_eog_ids - the tokenizer config may be incorrect
load: printing all EOG tokens:
load:   - 2 ('</s>')
load: special tokens cache size = 3
load: token to piece cache size = 0.1637 MB
print_info: arch             = llama
print_info: vocab_only       = 0
print_info: n_ctx_train      = 32768
print_info: n_embd           = 4096
print_info: n_layer          = 32
print_info: n_head           = 32
print_info: n_head_kv        = 8
print_info: n_rot            = 128
print_info: n_swa            = 0
print_info: is_swa_any       = 0
print_info: n_embd_head_k    = 128
print_info: n_embd_head_v    = 128
print_info: n_gqa            = 4
print_info: n_embd_k_gqa     = 1024
print_info: n_embd_v_gqa     = 1024
print_info: f_norm_eps       = 0.0e+00
print_info: f_norm_rms_eps   = 1.0e-05
print_info: f_clamp_kqv      = 0.0e+00
print_info: f_max_alibi_bias = 0.0e+00
print_info: f_logit_scale    = 0.0e+00
print_info: f_attn_scale     = 0.0e+00
```

```
from langchain.schema.runnable import RunnablePassthrough
from langchain.schema.output_parser import StrOutputParser
from langchain.prompts import ChatPromptTemplate
```

```
template = """
<|context|>
You are personal medical assistant called MediMate developed by Navaz that follows instruction extremely well.
Please be truthful and give direct answers.You can give health advice and insights.
</s>
```

```
<|user|>
{query}
</s>
 <|assistant|>
"""
```

```
prompt = ChatPromptTemplate.from_template(template)
```

```
rag_chain = (
    {"context": retriever,  "query": RunnablePassthrough()}
    | prompt
    | llm
    | StrOutputParser()
)
```

```
response = rag_chain.invoke("What is heart attack??")
```

```
llama_perf_context_print:        load time =   20421.35 ms
llama_perf_context_print: prompt eval time =   20421.08 ms /    75 tokens (  272.28 ms per token,     3.67 tokens per second)
llama_perf_context_print:        eval time =   35866.82 ms /    56 runs   (  640.48 ms per token,     1.56 tokens per second)
llama_perf_context_print:       total time =   56350.33 ms /   131 tokens
llama_perf_context_print:    graphs reused =         59
```

```
to_markdown(response)
```

> A heart attack is a myocardial infarction (MI), which occurs when there is an obstruction of the blood flow to part of the heart. This results in damage or death of the cardiac muscle. It is also known as a coronary heart disease.

```
import sys

while True:
  user_input = input(f"Input Prompt: ")
  if user_input == 'exit':
    print('Exiting')
    sys.exit()
  if user_input == '':
    continue
  result = rag_chain.invoke(user_input)
  print("Answer: ",result)
```

```
Input Prompt: What are symptoms of heart attack?
Llama.generate: 60 prefix-match hit, remaining 17 prompt tokens to eval
```