RAG Pipeline Exercise

In this exercise you will build and **compare two simple Retrieval-Augmented Generation (RAG) pipelines**.

You will work with a small collection of PDF documents (e.g. medical guidelines) and:

1. Load and chunk the PDF documents.
2. Create a vector index using **embedding model A** (local `BAAI/bge-m3`).
3. Create a second index using **embedding model B** (e.g. OpenAI or Gemini embeddings).
4. Implement a simple **retriever** and an **answering function** that calls an LLM with retrieved context.
5. Automatically **generate questions** from the documents and use them to **compare two RAG configurations**.

Cells marked with `# TODO` are **for students to implement**. Everything else is provided scaffolding.

0. Setup & Imports

In [1]:
```
# TODO (easy): skim the imports and make sure you understand what each library i

from dotenv import load_dotenv
import os
import glob
from PyPDF2 import PdfReader
from langchain_text_splitters import RecursiveCharacterTextSplitter
import faiss
from sentence_transformers import SentenceTransformer
import pickle
import random
import numpy as np
import pandas as pd

# LLM / API clients (we will mainly use OpenAI here; Gemini can be added as a bo
from openai import OpenAI
```

```
WARNING:tensorflow:From c:\Users\Phil\AppData\Local\Programs\Python\Python313\Lib
\site-packages\tf_keras\src\losses.py:2976: The name tf.losses.sparse_softmax_cro
ss_entropy is deprecated. Please use tf.compat.v1.losses.sparse_softmax_cross_ent
ropy instead.
```

In [2]:
```
# Load API keys from .env (you need to create this file once and add your keys)
load_dotenv()

deepinfra_key = os.getenv("DEEPINFRA_API_KEY")
openai_api_key = os.getenv("OPENAI_API_KEY")
google_api_key = os.getenv("GOOGLE_API_KEY")
anthropic_api_key = os.getenv("ANTHROPIC_API_KEY")
```

```
# For this exercise we mainly use OpenAI for both embeddings (RAG B) and chat co
assert openai_api_key is not None, "Please set OPENAI_API_KEY in your .env file.
openai_client = OpenAI(api_key=openai_api_key)
```

In [3]:
```
# Make pandas show the full table and full cell content
pd.set_option("display.max_rows", None)       # show all rows
pd.set_option("display.max_columns", None)    # show all columns
pd.set_option("display.max_colwidth", None)   # don't truncate cell text
```

1. Load PDF documents

We assume there is a `data/` folder containing one or more PDF files.

Task: implement `load_pdfs(glob_path)` so that it:

- Iterates over all PDF files matching `glob_path`
- Reads them with `PdfReader`
- Concatenates the text of all pages into **one long string**.

In [4]:
```
def load_pdfs(glob_path: str = "data/*.pdf") -> str:
    """Load all PDFs matching the pattern and return their combined text.

    TODO:
    - Use `glob.glob(glob_path)` to iterate over file paths
    - For each file, open it in binary mode and create a `PdfReader`
    - Loop over `reader.pages` and extract text with the extract_text() function
    - Concatenate everything into a single string `text`
    - Be robust: skip pages where `extract_text()` returns None
    """
    # YOUR CODE HERE
    text = ""
    for pdf_path in glob.glob(glob_path):
        with open(pdf_path, "rb") as f:
            reader = PdfReader(f)
            for page in reader.pages:
                page_text = page.extract_text()
                if page_text:
                    text += " " + page_text
    return text
```

In [5]:
```
# Run once and inspect
raw_text = load_pdfs("data/*.pdf")
print("Number of characters:", len(raw_text))
print("Preview:", raw_text[:500])
```

```
Number of characters: 230708
Preview:  Asthma: diagnosis,
moni toring and chr onic
asthma manag emen t (BTS,
NICE, SI GN)
NICE guideline
Published: 27 No vember 202 4
www .nice.or g.uk/guidance/ng2 45
© NICE 202 4. All right s reserved. Subject t o Notice of right s (https://www .n
ice.or g.uk/t erms-and-
conditions#notice-of -right s). Your r esponsi bility
The r ecommendations in t his guideline r epresent t he view of NICE, arriv ed at
aft er car eful
consideration of t he evidence a vailable. When e xercising t heir judg
```

2. Chunk the text

We will split the long text into overlapping chunks.

Later you can **experiment** with different `chunk_size` and `chunk_overlap` to see how it affects retrieval.

Task: start with the given parameters, run once, then try at least one alternative configuration and note the effects.

In [6]:
```python
# Base configuration (RAG A)
chunk_size_a = 2000
chunk_overlap_a = 200

splitter_a = RecursiveCharacterTextSplitter(
    chunk_size=chunk_size_a,
    chunk_overlap=chunk_overlap_a
)

chunks_a = splitter_a.split_text(raw_text)
print(f"RAG A: {len(chunks_a)} chunks produced, first chunk length = {len(chunks

# TODO (mini-experiment): change chunk_size / chunk_overlap for RAG B and compar
chunk_size_b = 1000   # e.g. smaller chunks
chunk_overlap_b = 100

splitter_b = RecursiveCharacterTextSplitter(
    chunk_size=chunk_size_b,
    chunk_overlap=chunk_overlap_b
)

chunks_b = splitter_b.split_text(raw_text)
print(f"RAG B: {len(chunks_b)} chunks produced, first chunk length = {len(chunks
```

```
RAG A: 130 chunks produced, first chunk length = 1995
RAG B: 260 chunks produced, first chunk length = 979
```

3. Create embeddings and a FAISS index

We start with **Embedding model A:** **`BAAI/bge-small-en`** using `sentence-transformers` .

Then, as an optional extension, you can build **Embedding model B** using OpenAI or Gemini and compare.

To keep the exercise manageable, the base version only **requires** BGE.

In [7]:
```
# Embedding model A (local)
model_name_a = "BAAI/bge-small-en"
embedder_a = SentenceTransformer(model_name_a)

# Compute embeddings for all chunks of configuration A
embeddings_a = embedder_a.encode(chunks_a, convert_to_numpy=True)

dimensions_a = embeddings_a.shape[1]
print("Embedding dimensionality (A):", dimensions_a)

index_a = faiss.IndexFlatL2(dimensions_a)
index_a.add(embeddings_a)
print("FAISS index (A) size:", index_a.ntotal)

# Persist index/chunks if you like (optional)
os.makedirs("faiss", exist_ok=True)
faiss.write_index(index_a, "faiss/faiss_index_a.index")
with open("faiss/chunks_a.pkl", "wb") as f:
    pickle.dump(chunks_a, f)
```

```
Embedding dimensionality (A): 384
FAISS index (A) size: 130
```

In [8]:
```
# Embedding model B using OpenAI embeddings.

# TODO:
# - Use `openai_client.embeddings.create(...)` to compute embeddings for `chunks_
# - Create a second FAISS index `index_b`
# - Make sure to check the dimensionality from the first embedding vector

# Example sketch (not complete, adapt & run if you have API access):
# Initialize OpenAI client
openai_client = OpenAI(api_key=openai_api_key)
response = openai_client.embeddings.create(
     model="text-embedding-3-small",
    input=chunks_b
)
embeddings_b = np.array([item.embedding for item in response.data])
dim_b = embeddings_b.shape[1]
index_b = faiss.IndexFlatL2(dim_b)
index_b.add(embeddings_b)
print("FAISS index (B) size:", index_b.ntotal)
```

```
FAISS index (B) size: 260
```

4. Implement a simple retriever

We now implement a generic retrieval function that:

1. Embeds the query.
2. Searches the FAISS index.
3. Returns the corresponding text chunks.

We implement it for configuration A. If you built configuration B, you can reuse the same function.

In [9]:
```python
def retrieve_texts(query: str, k: int, index, chunks, embedder) -> list:
    """Return the top-k most similar chunks for a query.

    TODO (students):
    - Encode the query with `embedder.encode(...)`
    - Call `index.search(query_embedding, k)`
    - Use the returned indices to select the chunks
    - Return a list of strings (chunks)
    """
    # YOUR CODE HERE
    query_emb = embedder.encode([query], convert_to_numpy=True)
    distances, indices = index.search(query_emb, k)
    retrieved = [chunks[i] for i in indices[0]]
    return retrieved

# Quick sanity check
test_query = "What is the most important factor in diagnosing asthma?"
retrieved_text = retrieve_texts(test_query, k=3, index=index_a, chunks=chunks_a,
print("Number of retrieved chunks:", len(retrieved_text))
print("Preview of first chunk:", retrieved_text[0][:400])
```

```
Number of retrieved chunks: 3
Preview of first chunk: and signs of ot her causes of r espirat ory sympt oms but
be awar e that e ven if
examination r esult s are normal, t he person ma y still ha ve ast hma. [NICE 201
7]
Initial tr eatmen t and obje ctive tests f or acu te sym ptoms a t
presen tation
1.1.5 Treat people immediat ely if t hey are acut ely unw ell or highly sympt oma
tic at
presentation, and per form objectiv e tests that ma y help s
```

In [10]:
```python
def openai_embed_query(query: str) -> np.ndarray:
    """
    Compute an OpenAI embedding for a single query string and
    return it as a NumPy array of shape (1, dim).
    """
    resp = openai_client.embeddings.create(
        model="text-embedding-3-small",
        input=[query]  # list with one string
    )
    vec = np.array(resp.data[0].embedding, dtype="float32")
    return vec.reshape(1, -1)
```

In [11]:
```python
def retrieve_texts_b(query: str, k: int, index, chunks) -> list:
    """
    Retrieve top-k chunks using OpenAI embeddings for the query.
    """
    query_emb = openai_embed_query(query)  # shape (1, dim)
    distances, indices = index.search(query_emb, k)
    retrieved = [chunks[i] for i in indices[0]]
    return retrieved
```

5. Implement `answer_query` using an LLM

Now we build the actual RAG call:

1. Use `retrieve_texts` to get top- `k` chunks.
2. Concatenate them into a context string.
3. Build a prompt that:
   - shows the context
   - asks the model to answer the user question based **only** on this context.
4. Call the OpenAI chat completion API.

This is the **core RAG function**.

In [12]:
```
def answer_query_a(query: str, k: int, index, chunks, embedder, client: OpenAI)
    """RAG-style answer: retrieve context and ask an LLM.

    TODO (students):
    - Use `retrieve_texts` to get `k` relevant chunks.
    - Join them into a single context string.
    - Build a chat prompt that instructs the model to answer *only* using the co
    - Call `client.chat.completions.create(...)` with model `"gpt-4o-mini"` (or
    - Return the model's answer text.
    """
    retrieved_chunks = retrieve_texts(query, k, index, chunks, embedder)
    context = "\n\n---\n\n".join(retrieved_chunks)

    system_prompt = (
        "You are a helpful assistant answering questions based ONLY on the provi
        "If the answer is not in the context, say that you do not know."
    )

    messages = [
        {"role": "system", "content": system_prompt},
        {"role": "user", "content": f"Context:\n{context}\n\nQuestion: {query}"}
    ]

    completion = client.chat.completions.create(
        model="gpt-4o-mini",
        messages=messages
    )

    return completion.choices[0].message.content.strip()

# Quick manual test
answer = answer_query_a(test_query, k=3, index=index_a, chunks=chunks_a, embedde
print("RAG answer:", answer)
```

```
RAG answer: The most important factor in diagnosing asthma is the presence of a h
istory suggestive of asthma, supported by the results of objective tests such as
eosinophil count, fractional exhaled nitric oxide (FeNO), spirometry, or peak exp
iratory flow (PEF). Even if examination results are normal, a person may still ha
ve asthma.
```

Answer Function with the other embedding type

In [13]:
```
def answer_query_b(query: str, k: int, index, chunks, client: OpenAI) -> str:
    """
    RAG-style answer for configuration B (OpenAI embeddings + chunking B).
    """
```

```
    retrieved_chunks = retrieve_texts_b(query, k, index, chunks)
    context = "\n\n---\n\n".join(retrieved_chunks)

    system_prompt = (
        "You are a helpful assistant answering questions based ONLY on the provi
        "If the answer is not in the context, say that you do not know."
    )

    messages = [
        {"role": "system", "content": system_prompt},
        {"role": "user", "content": f"Context:\n{context}\n\nQuestion: {query}"}
    ]

    completion = client.chat.completions.create(
        model="gpt-4o-mini",
        messages=messages,
    )

    return completion.choices[0].message.content.strip()
```

6. Generate questions from random chunks (automatic evaluation set)

To compare two RAG configurations, we need **questions**.

We will:

- randomly sample a few chunks from the corpus,
- ask an LLM to generate a **good question** whose answer is contained in the chunk.

Then we can use these question–chunk pairs as a small evaluation set.

We provide most of the implementation. Your job is mainly to:

- inspect the code,
- understand the prompt,
- maybe tweak the number of chunks or retries.

In [14]:
```
def generate_questions_for_random_chunks(chunks, num_chunks: int = 5, max_retrie
    selected_chunks = random.sample(chunks, num_chunks)
    qa_pairs = []

    for chunk in selected_chunks:
        prompt = prompt = (
            "Based on the following text, generate an insightful question that c
            "Text:\n" + chunk + "\n\n"
            "Question:"
        )

        question = None
        for attempt in range(max_retries):
            try:
                completion = openai_client.chat.completions.create(
                    model="gpt-4o-mini",
                    messages=[{"role": "user", "content": prompt}]
                )
```

```
                question = completion.choices[0].message.content.strip()
                if question:
                    break
            except Exception as e:
                print("Error while generating question, retrying...", e)

        if question is None:
            question = "Error: could not generate question."

        qa_pairs.append((chunk, question))

    return qa_pairs

questions = generate_questions_for_random_chunks(chunks_a, num_chunks=5, max_ret
for i, (chunk, q) in enumerate(questions, 1):
    print(f"Q{i}: {q}\n  From chunk preview: {chunk[:120]}...\n")
```

```
Q1: What are the key research recommendations related to the management of hypert
ension in adults, particularly regarding the effectiveness of relaxation therapie
s, same-day specialist assessments, and optimal blood pressure targets for those
with aortic aneurysms and prior strokes?
  From chunk preview: step 1 treatment .
5 Relaxa tion ther apies
What is t he clinical and cost eff ectiv eness of r elaxation t herapies f...

Q2: What are the roles of long-acting medications and maintenance therapy in the
management of asthma, and how do they differ in their mechanisms of action and us
age compared to short-acting therapies?
  From chunk preview: A long-acting medicine t hat act s on beta-r ecept ors in t
he air way to relax air way smoot h
muscle and r elieve sym...

Q3: What are the key criteria and recommendations for the immediate management an
d referral of patients with severely elevated blood pressure and signs of target
organ damage, according to the 2019 NICE guidelines?
  From chunk preview: higher), but no sympt oms or signs indicating same-da y ref
erral (see
recommendation 1 .5.2), carr y out in vestigation...

Q4: What are the key recommendations from the BTS, NICE, and SIGN 2024 guidelines
regarding the pharmacological management of asthma in children aged 5 to 11, part
icularly concerning the initial treatment and the use of maintenance and reliever
therapy?
  From chunk preview: dose of ICS t o a specialist in ast hma car e. [BTS/NICE/SI
GN 202 4]
For a shor t explanation of wh y the committ ee ma...

Q5: What are the recommended steps for managing adult hypertension that is not co
ntrolled with initial treatments, and how should adherence to medication be evalu
ated?
  From chunk preview: • review t he person's medications t o ensur e they are bei
ng tak en at t he
optimal t olerat ed doses and
• discuss a...
```

7. Compare two RAG configurations

Now we can:

- Use the generated questions,
- Answer them with RAG configuration A (BGE + chunking A),
- (Optional) Answer them with RAG configuration B (e.g. different chunking and/or different embeddings),
- Compare the answers qualitatively.

To keep the exercise manageable, we start with config A only. If you implemented config B, reuse `answer_query` with `index_b`, `chunks_b`, and your second embedder.

In [15]:
```
def answer_generated_questions_a(question_tuples, k, index, chunks, embedder, cl
    results = []
    for chunk, question in question_tuples:
        answer = answer_query_a(question, k, index, chunks, embedder, client)
        results.append({
            "chunk": chunk,
            "question": question,
            "answer": answer
        })
    return results

results_a = answer_generated_questions_a(
    questions,
    k=5,
    index=index_a,
    chunks=chunks_a,
    embedder=embedder_a,
    client=openai_client,
)

for item in results_a:
    print("Question:", item["question"])
    print("Answer A:", item["answer"])
    print("Source chunk preview:", item["chunk"][:150], "...")
    print("-" * 60)
```

Question: What are the key research recommendations related to the management of hypertension in adults, particularly regarding the effectiveness of relaxation therapies, same-day specialist assessments, and optimal blood pressure targets for those with aortic aneurysms and prior strokes?
Answer A: The key research recommendations related to the management of hypertension in adults include:

1. **Relaxation Therapies**: There is a recommendation for research on the clinical and cost-effectiveness of relaxation therapies for managing primary hypertension in adults, particularly in terms of reducing cardiovascular events and improving quality of life.

2. **Same-Day Specialist Assessment**: Research is recommended to identify which individuals with extreme hypertension (220/120 mmHg or higher) or emergency symptoms should be referred for same-day hospital specialist assessment.

3. **Blood Pressure Targets for Aortic Aneurysm**: Research is needed to determine the optimal blood pressure targets in adults with hypertension and aortic aneurysm, including whether these targets vary by age.

4. **Blood Pressure Targets for Prior Stroke**: Research is called for to establish the optimal blood pressure targets in adults with a history of ischaemic or haemorrhagic stroke.

These recommendations highlight the need for further investigation to inform future guidelines and improve the management of hypertension in specific populations.
Source chunk preview: step 1 treatment .
5 Relaxa tion ther apies
What is t he clinical and cost eff ectiv eness of r elaxation t herapies f or man aging primar y
hyper t ...

Question: What are the roles of long-acting medications and maintenance therapy in the management of asthma, and how do they differ in their mechanisms of action and usage compared to short-acting therapies?
Answer A: Long-acting medications in the management of asthma, such as long-acting beta-2 agonists (LABAs) and long-acting muscarinic receptor antagonists (LAMAs), act on different receptors to provide relief from asthma symptoms. LABAs relax airway smooth muscle by acting on beta receptors, while LAMAs act on muscarinic receptors to achieve a similar effect. Both types of long-acting medications are used for maintaining control of asthma over an extended period, helping to reduce symptoms and prevent exacerbations.

Maintenance therapy, such as the Maintenance and Reliever Therapy (MART) approach, involves using a combination inhaler that contains inhaled corticosteroids (ICS) along with a fast-acting LABA (like formoterol) for daily use and symptomatic relief as needed. This form of therapy is designed to provide consistent management of asthma while allowing for quick relief from symptoms.

In contrast, short-acting therapies, primarily short-acting beta-2 agonists (SABAs), are used on an as-needed basis for quick relief of acute symptoms. They act rapidly to relieve bronchospasm but do not provide long-term control of asthma symptoms, necessitating regular use of maintenance therapy to manage the condition effectively.

Thus, the main differences lie in their duration of action, mechanism of action, and role in overall asthma management, with long-acting medications and maintenance therapies being crucial for ongoing symptom control and prevention of exacerbations, while short-acting therapies are reserved for immediate relief.
Source chunk preview: A long-acting medicine t hat act s on beta-r ecept ors in t

he air way to relax air way smoot h
muscle and r elieve sympt oms of ast hma.
Long-ac ti ...

Question: What are the key criteria and recommendations for the immediate managem ent and referral of patients with severely elevated blood pressure and signs of t arget organ damage, according to the 2019 NICE guidelines?
Answer A: According to the 2019 NICE guidelines, the key criteria and recommendat ions for the immediate management and referral of patients with severely elevated blood pressure (clinic blood pressure of 180/120 mmHg and higher) and signs of ta rget organ damage are as follows:

1. If target organ damage is identified, consider starting antihypertensive drug treatment immediately, without waiting for the results of Ambulatory Blood Pressu re Monitoring (ABPM) or Home Blood Pressure Monitoring (HBPM).
2. If no target organ damage is identified, confirm diagnosis by:
   - Repeating clinic blood pressure measurement within 7 days, or
   - Considering monitoring using ABPM (or HBPM if ABPM is not suitable or not to lerated), following recommendations and ensuring a clinical review within 7 days.
3. Refer people for specialist assessment, carried out on the same day, if they h ave a clinic blood pressure of 180/120 mmHg and higher with:
   - Signs of retinal hemorrhage or papilloedema (accelerated hypertension) or
   - Life-threatening symptoms such as new onset confusion, chest pain, signs of heart failure, or acute kidney injury.
4. Refer people for specialist assessment, carried out on the same day, if they h ave suspected pheochromocytoma (for example, labile or postural hypotension, head ache, palpitations, pallor, abdominal pain, or diaphoresis).
Source chunk preview: higher), but no sympt oms or signs indicating same-da y ref erral (see
recommendation 1 .5.2), carr y out in vestigations f or tar get or gan damage (. ..

Question: What are the key recommendations from the BTS, NICE, and SIGN 2024 guid elines regarding the pharmacological management of asthma in children aged 5 to 1 1, particularly concerning the initial treatment and the use of maintenance and r eliever therapy?
Answer A: The key recommendations from the BTS, NICE, and SIGN 2024 guidelines re garding the pharmacological management of asthma in children aged 5 to 11 are as follows:

1. **Initial Treatment**: Offer a twice-daily pediatric low-dose inhaled corticos teroid (ICS), with a short-acting beta 2 agonist (SABA) as needed, as the initial treatment for children aged 5 to 11 years with newly diagnosed asthma.

2. **Maintenance and Reliever Therapy (MART)**: Consider pediatric low-dose MART (maintenance and reliever therapy) for children with asthma that is not controlle d on pediatric low-dose ICS plus SABA.

These recommendations aim to effectively manage asthma in children and guide thei r treatment plans based on controlling symptoms and improving overall health outc omes.
Source chunk preview: dose of ICS t o a specialist in ast hma car e. [BTS/NICE/SI GN 202 4]
For a shor t explanation of wh y the committ ee made t hese r ecommendations and ...

Question: What are the recommended steps for managing adult hypertension that is not controlled with initial treatments, and how should adherence to medication be evaluated?
Answer A: For managing adult hypertension that is not controlled with initial tre

```
atments, the following steps are recommended:

1. **Check Adherence**: Before considering the next step in treatment, discuss wi
th the person if they are taking their medicine as prescribed and support adheren
ce. This evaluation aligns with recommendation 1.4.40.

2. **Step 2 Treatment**: If hypertension is not controlled in adults taking step
1 treatment of an ACE inhibitor or ARB, offer one of the following drugs in addit
ion to step 1 treatment:
   - A CCB (Calcium Channel Blocker) or
   - A thiazide-like diuretic.

3. **Alternative for Step 1 CCB**: If hypertension is not controlled in adults ta
king step 1 treatment of a CCB, offer one of the following drugs in addition:
   - An ACE inhibitor or
   - An ARB or
   - A thiazide-like diuretic.

4. **Special Consideration for Certain Groups**: For adults of Black African or A
frican-Caribbean family origin who do not have type 2 diabetes and whose hyperten
sion is not controlled on step 1 treatment, consider an ARB in preference to an A
CE inhibitor in addition to step 1 treatment.

5. **Step 3 Treatment**: If hypertension remains uncontrolled in adults taking st
ep 2 treatment, offer a combination of:
   - An ACE inhibitor or ARB,
   - A CCB, and
   - A thiazide-like diuretic.

6. **Step 4 Treatment**: If hypertension is still not controlled with the optimal
tolerated doses of an ACE inhibitor or ARB plus a CCB and a thiazide-like diureti
c, regard them as having resistant hypertension. Confirm elevated blood pressure
measurements using ambulatory or home blood pressure recordings, assess for postu
ral hypotension, and discuss adherence again.

7. **Consider Adding a Fourth Drug**: For those with confirmed resistant hyperten
sion, consider adding a fourth antihypertensive drug as step 4 treatment or seeki
ng specialist advice.

Throughout this process, adherence to medication should be evaluated regularly as
per recommendation 1.4.40, ensuring that individuals are taking their medications
as prescribed.
Source chunk preview: • review t he person's medications t o ensur e they are bei
ng tak en at t he
optimal t olerat ed doses and
• discuss adher ence (see recommendation  ...
------------------------------------------------------------
```

Extension: add RAG B and create a comparison table

If you implemented a second configuration (e.g. different chunking + OpenAI embeddings):

1. Build `index_b` using OpenAI embeddings and `chunks_b`.
2. Implement `openai_embed_query`, `retrieve_texts_b`, and `answer_query_b`.
3. Run `results_b = answer_generated_questions_b(questions, k=5, index=index_b, chunks=chunks_b, client=openai_client)`.

4. For each question, compare:
    - Which answer is more complete / specific?
    - Which one is better grounded in the source chunk?
5. Summarise your findings in a short **markdown cell** or a small table.

This concludes the core RAG exercise.

In [16]:
```
def answer_generated_questions_b(question_tuples, k, index, chunks, client):
    """
    Use RAG B to answer a list of (chunk, question) pairs.
    """
    results = []
    for chunk, question in question_tuples:
        answer = answer_query_b(question, k, index, chunks, client)
        results.append({
            "chunk": chunk,
            "question": question,
            "answer": answer,
        })
    return results
```

In [17]:
```
results_b = answer_generated_questions_b(
    questions,
    k=5,
    index=index_b,
    chunks=chunks_b,
    client=openai_client,
)
```

In [18]:
```
rows = []

for qa_a, qa_b in zip(results_a, results_b):
    rows.append({
        "Question": qa_a["question"],
        "Answer A (BGE + config A)": qa_a["answer"],
        "Answer B (OpenAI + config B)": qa_b["answer"],
        "Source chunk (A) preview": qa_a["chunk"][:200] + "..."
    })

df_comparison = pd.DataFrame(rows)
display(df_comparison)
```

	Question	Answer A (BGE + config A)	Answer B (OpenAI + config B)	Source chunk (A) preview
0	What are the key research recommendations related to the management of hypertension in adults, particularly regarding the effectiveness of relaxation therapies, same-day specialist assessments, and optimal blood pressure targets for those with aortic aneurysms and prior strokes?	The key research recommendations related to the management of hypertension in adults include:\n\n1. **Relaxation Therapies**: There is a recommendation for research on the clinical and cost-effectiveness of relaxation therapies for managing primary hypertension in adults, particularly in terms of reducing cardiovascular events and improving quality of life.\n\n2. **Same-Day Specialist Assessment**: Research is recommended to identify which individuals with extreme hypertension (220/120 mmHg or higher) or emergency symptoms should be referred for same-day hospital specialist assessment.\n\n3. **Blood Pressure Targets for Aortic Aneurysm**: Research is needed to determine the optimal blood pressure targets in adults with hypertension and aortic aneurysm, including whether these targets vary by age.\n\n4. **Blood Pressure Targets for Prior Stroke**: Research is called for to establish the optimal blood pressure targets in adults with a history of ischaemic or haemorrhagic stroke.\n\nThese recommendations highlight the need for further investigation to inform future guidelines and improve the management of hypertension in specific populations.	The key research recommendations related to the management of hypertension in adults include:\n\n1. **Relaxation Therapies**: Further research is needed to determine the clinical effectiveness of relaxation therapies for managing primary hypertension, particularly in terms of reducing cardiovascular events and improving the quality of life. The committee noted that a larger study would be needed to obtain meaningful results.\n\n2. **Same-Day Specialist Assessment**: The committee recommends investigating which individuals with extreme hypertension (220/120 mmHg or higher) or emergency symptoms should be referred for same-day hospital specialist assessment to inform future recommendations.\n\n3. **Blood Pressure Targets**: There is no specific mention of research recommendations regarding optimal blood pressure targets for individuals with aortic aneurysms and prior strokes in the provided context.\n\nIf you need more detailed information on any specific aspect, please let me know!	step 1 treatment .\n5 Relaxa tion ther apies \nWhat is t he clinical and cost eff ectiv eness of r elaxation t herapies f or managing primar y \nhyper tension in adult s in t erms of r educing car diova...

	Question	Answer A (BGE + config A)	Answer B (OpenAI + config B)	Source chunk (A) preview
1	What are the roles of long-acting medications and maintenance therapy in the management of asthma, and how do they differ in their mechanisms of action and usage compared to short-acting therapies?	Long-acting medications in the management of asthma, such as long-acting beta-2 agonists (LABAs) and long-acting muscarinic receptor antagonists (LAMAs), act on different receptors to provide relief from asthma symptoms. LABAs relax airway smooth muscle by acting on beta receptors, while LAMAs act on muscarinic receptors to achieve a similar effect. Both types of long-acting medications are used for maintaining control of asthma over an extended period, helping to reduce symptoms and prevent exacerbations. \n\nMaintenance therapy, such as the Maintenance and Reliever Therapy (MART) approach, involves using a combination inhaler that contains inhaled corticosteroids (ICS) along with a fast-acting LABA (like formoterol) for daily use and symptomatic relief as needed. This form of therapy is designed to provide consistent management of asthma while allowing for quick relief from symptoms. \n\nIn contrast, short-acting therapies, primarily short-acting beta-2 agonists (SABAs), are used on an as-needed basis for quick relief of acute symptoms. They act rapidly to relieve bronchospasm but do not provide long-term control of asthma symptoms, necessitating regular use of maintenance therapy to manage the condition	I do not know.	A long-acting medicine t hat act s on beta-r ecept ors in t he air way to relax air way smoot h \nmuscle and r elieve sympt oms of ast hma. \nLong-ac ting m uscarinic r eceptor an tagonist \nA long-actin...

	Question	Answer A (BGE + config A)	Answer B (OpenAI + config B)	Source chunk (A) preview
		effectively.\n\nThus, the main differences lie in their duration of action, mechanism of action, and role in overall asthma management, with long-acting medications and maintenance therapies being crucial for ongoing symptom control and prevention of exacerbations, while short-acting therapies are reserved for immediate relief.		
2	What are the key criteria and recommendations for the immediate management and referral of patients with severely elevated blood pressure and signs of target organ damage, according to the 2019 NICE guidelines?	According to the 2019 NICE guidelines, the key criteria and recommendations for the immediate management and referral of patients with severely elevated blood pressure (clinic blood pressure of 180/120 mmHg and higher) and signs of target organ damage are as follows:\n\n1. If target organ damage is identified, consider starting antihypertensive drug treatment immediately, without waiting for the results of Ambulatory Blood Pressure Monitoring (ABPM) or Home Blood Pressure Monitoring (HBPM).\n2. If no target organ damage is identified, confirm diagnosis by:\n - Repeating clinic blood pressure measurement within 7 days, or\n - Considering monitoring using ABPM (or HBPM if ABPM is not suitable or not tolerated), following recommendations and ensuring a clinical review within 7 days.\n3. Refer people for specialist assessment, carried out on the same day, if they	The key criteria and recommendations for the immediate management and referral of patients with severely elevated blood pressure (clinic blood pressure of 180/120 mmHg or higher) according to the 2019 NICE guidelines include:\n\n1. **Referral for Same-Day Specialist Review**: If a person has severe hypertension but shows no symptoms or signs indicating the need for same-day referral, investigations for target organ damage should be carried out as soon as possible. \n\n2. **Investigations**: Prompt investigations for target organ damage are recommended for those with severe hypertension and no concerning symptoms.\n\n3. **Follow-Up**: Checking blood pressure again within 7 days for individuals	higher), but no sympt oms or signs indicating same-da y referral (see \nrecommendation 1 .5.2), carr y out in vestigations f or tar get or gan damage (see \nrecommendation 1.3.3) as soon as possible: \n...

	Question	Answer A (BGE + config A)	Answer B (OpenAI + config B)	Source chunk (A) preview
		have a clinic blood pressure of 180/120 mmHg and higher with:\n - Signs of retinal hemorrhage or papilloedema (accelerated hypertension) or\n - Life-threatening symptoms such as new onset confusion, chest pain, signs of heart failure, or acute kidney injury.\n4. Refer people for specialist assessment, carried out on the same day, if they have suspected pheochromocytoma (for example, labile or postural hypotension, headache, palpitations, pallor, abdominal pain, or diaphoresis).	with severe hypertension and no target organ damage is advised to ensure proper monitoring and treatment.\n\n4. **Research Consideration**: There is an acknowledgment that further research is needed, particularly for individuals with extreme hypertension (220/120 mmHg or higher) or those showing emergency symptoms.\n\nThese recommendations are intended to ensure timely assessment and management to prevent complications associated with severely elevated blood pressure.	
3	What are the key recommendations from the BTS, NICE, and SIGN 2024 guidelines regarding the pharmacological management of asthma in children aged 5 to 11, particularly concerning the initial treatment and the use of maintenance and reliever therapy?	The key recommendations from the BTS, NICE, and SIGN 2024 guidelines regarding the pharmacological management of asthma in children aged 5 to 11 are as follows:\n\n1. **Initial Treatment**: Offer a twice-daily pediatric low-dose inhaled corticosteroid (ICS), with a short-acting beta 2 agonist (SABA) as needed, as the initial treatment for children aged 5 to 11 years with newly diagnosed asthma. \n\n2. **Maintenance and Reliever Therapy (MART)**: Consider pediatric low-dose MART (maintenance and reliever therapy) for children with asthma that is not controlled on pediatric low-dose ICS plus SABA.	The key recommendations from the BTS, NICE, and SIGN 2024 guidelines regarding the pharmacological management of asthma in children aged 5 to 11 include: \n\n1. **Initial Treatment**: Offer a twice-daily paediatric low-dose inhaled corticosteroid (ICS) with a short-acting beta 2 agonist (SABA) as needed, as the initial treatment for children aged 5 to 11 years with newly diagnosed asthma. \n\n2. **Maintenance and Reliever Therapy**: Refer children to a specialist in asthma care if their asthma is not controlled on	dose of ICS t o a specialist in ast hma car e. [BTS/ NICE/SIGN 202 4] \nFor a shor t explanation of wh y the committ ee made t hese r ecommendations and ho w \nthey might aff ect practice, see t he ratio...

	Question	Answer A (BGE + config A)	Answer B (OpenAI + config B)	Source chunk (A) preview
		\n\nThese recommendations aim to effectively manage asthma in children and guide their treatment plans based on controlling symptoms and improving overall health outcomes.	paediatric moderate-dose Maintenance And Reliever Therapy (MART) or paediatric moderate-dose ICS/ LABA maintenance treatment (with or without an LTRA, depending on previous response). \n\nThese recommendations aim to improve asthma management through careful use of medication in this age group.	
4	What are the recommended steps for managing adult hypertension that is not controlled with initial treatments, and how should adherence to medication be evaluated?	For managing adult hypertension that is not controlled with initial treatments, the following steps are recommended: \n\n1. **Check Adherence**: Before considering the next step in treatment, discuss with the person if they are taking their medicine as prescribed and support adherence. This evaluation aligns with recommendation 1.4.40. \n\n2. **Step 2 Treatment**: If hypertension is not controlled in adults taking step 1 treatment of an ACE inhibitor or ARB, offer one of the following drugs in addition to step 1 treatment:\n - A CCB (Calcium Channel Blocker) or\n - A thiazide-like diuretic.\n\n3. **Alternative for Step 1 CCB**: If hypertension is not controlled in adults taking step 1 treatment of a CCB, offer one of the following drugs in addition:\n - An ACE inhibitor or\n - An ARB or\n - A thiazide-like diuretic.\n\n4. **Special	The provided context does not include information on the recommended steps for managing adult hypertension that is not controlled with initial treatments or how to evaluate adherence to medication. Therefore, I do not know.	• review t he person's medications t o ensur e they are being tak en at t he \noptimal t olerat ed doses and \n• discuss adher ence (see recommendation 1.4.40). [2019] \n1.4.45 If hyper tension is not c...

Question	Answer A (BGE + config A)	Answer B (OpenAI + config B)	Source chunk (A) preview
	Consideration for Certain Groups**: For adults of Black African or African-Caribbean family origin who do not have type 2 diabetes and whose hypertension is not controlled on step 1 treatment, consider an ARB in preference to an ACE inhibitor in addition to step 1 treatment.\n\n5. **Step 3 Treatment**: If hypertension remains uncontrolled in adults taking step 2 treatment, offer a combination of:\n - An ACE inhibitor or ARB, \n - A CCB, and\n - A thiazide-like diuretic. \n\n6. **Step 4 Treatment**: If hypertension is still not controlled with the optimal tolerated doses of an ACE inhibitor or ARB plus a CCB and a thiazide-like diuretic, regard them as having resistant hypertension. Confirm elevated blood pressure measurements using ambulatory or home blood pressure recordings, assess for postural hypotension, and discuss adherence again.\n\n7. **Consider Adding a Fourth Drug**: For those with confirmed resistant hypertension, consider adding a fourth antihypertensive drug as step 4 treatment or seeking specialist advice. \n\nThroughout this process, adherence to medication should be evaluated regularly as per recommendation 1.4.40, ensuring that individuals are taking their medications as prescribed.		

In [ ]: