<u>Exercise 5</u>

<u>Lab 2:</u>

<u>Indexing time:</u> 3.77 sec
<u>AVG response time:</u> 6.76 seconds
<u>Accuracy of questions:</u> 100%
The llm indexed the file properly and gave sufficient answers

CLI Results:

```
Response:
According to the context, the two parts for a warm-up are:

1. Part One: General Cardiovascular (CV) - 5 minutes on a treadmill, cross trainer, rower, or bike at a steady pace.
2. Part Two: Dynamic movements - three patterns:
   - Quadruped T-spine rotations (12 per side)
   - Mountain climbers (20 reps)
   - Cat camels (10 reps)
Elapsed time: 7.8005 seconds
INFO:httpx:HTTP Request: POST http://127.0.0.1:11434/api/chat "HTTP/1.1 200 OK"
INFO:langchain_classic.retrievers.multi_query:Generated queries: ['Here are five different versions of the original question:', '1. What be
nefits does performing a warmup bring to the vector database retrieval process?', "2. Can you explain why it's necessary to run a warmup be
fore searching for documents in a vector database?", '3. How does the warmup process improve the accuracy and relevance of search results i
n a vector database?', '4. What are the consequences of not performing a warmup on the performance of a vector database-based search system
?', '5. Under what circumstances is it recommended to perform a warmup before searching for documents in a vector database, and what benefi
ts can be expected from doing so?']
INFO:httpx:HTTP Request: POST http://127.0.0.1:11434/api/embed "HTTP/1.1 200 OK"
INFO:httpx:HTTP Request: POST http://127.0.0.1:11434/api/embed "HTTP/1.1 200 OK"
INFO:httpx:HTTP Request: POST http://127.0.0.1:11434/api/embed "HTTP/1.1 200 OK"
INFO:httpx:HTTP Request: POST http://127.0.0.1:11434/api/embed "HTTP/1.1 200 OK"
INFO:httpx:HTTP Request: POST http://127.0.0.1:11434/api/embed "HTTP/1.1 200 OK"
INFO:httpx:HTTP Request: POST http://127.0.0.1:11434/api/embed "HTTP/1.1 200 OK"
INFO:httpx:HTTP Request: POST http://127.0.0.1:11434/api/chat "HTTP/1.1 200 OK"
Response:
According to the text, you need to do a warm-up because your body needs to be warm for it to perform optimally. Just like a car's engine ne
eds to be warmed up before driving, your body needs to be warm before physical activity.
Elapsed time: 5.8554 seconds
INFO:httpx:HTTP Request: POST http://127.0.0.1:11434/api/chat "HTTP/1.1 200 OK"
INFO:langchain_classic.retrievers.multi_query:Generated queries: ['Here are five different versions of the original question:', '1. How can
 I optimize my workout routine to achieve greater muscle growth and development?', '2. What nutritional supplements or dietary adjustments
can help support muscle hypertrophy and strength gains?', '3. Can you recommend a combination of exercises, weights, and training protocols
 that effectively stimulate muscle protein synthesis and increase muscle mass?', '4. What is the role of adequate rest and recovery in buil
ding and maintaining muscle size, and how can I prioritize this aspect of my fitness routine?', '5. Are there any specific lifestyle or env
ironmental factors (e.g. temperature, humidity, sleep quality) that can impact muscle growth and development, and how can I optimize these
variables to support my goals?']
INFO:httpx:HTTP Request: POST http://127.0.0.1:11434/api/embed "HTTP/1.1 200 OK"
INFO:httpx:HTTP Request: POST http://127.0.0.1:11434/api/embed "HTTP/1.1 200 OK"
INFO:httpx:HTTP Request: POST http://127.0.0.1:11434/api/embed "HTTP/1.1 200 OK"
INFO:httpx:HTTP Request: POST http://127.0.0.1:11434/api/embed "HTTP/1.1 200 OK"
INFO:httpx:HTTP Request: POST http://127.0.0.1:11434/api/embed "HTTP/1.1 200 OK"
INFO:httpx:HTTP Request: POST http://127.0.0.1:11434/api/embed "HTTP/1.1 200 OK"
INFO:httpx:HTTP Request: POST http://127.0.0.1:11434/api/chat "HTTP/1.1 200 OK"
Response:
According to the provided context, to increase muscle size, you need to go into the gym with a specific purpose and engage in heavy resista
nce training. This is because heavy weight stimulates type II muscle fibers, which have potential for growth. Additionally, research confir
ms that mechanical tension, muscle damage, and metabolic stress are the factors that induce hypertrophy (muscle growth).
Elapsed time: 6.7232 seconds
```

Part B

A. Switching to a Smaller Model

The LLM was changed from Llama 3.2 : 3B to the smaller Llama 3.2 : 1B model.
Contrary to expectations, the model loading time remained effectively unchanged (≈ 3.4 seconds), indicating that model initialization overhead dominated over model size differences in this setup. In terms of output quality, the accuracy of answers was preserved. The primary observed difference was stylistic: responses from the smaller model were less detailed and less verbose, but still factually correct.

B. Reducing Chunk Size

The chunk size was reduced from 1200 to 600 tokens, while keeping other parameters unchanged. This change had a noticeable negative impact on answer accuracy. For example, when asked *"What are the parts for a warmup?"*, both the baseline configuration and Experiment A successfully retrieved and answered the question. In contrast, with the smaller chunk size, the model failed to retrieve sufficient context and responded that the information was not explicitly stated, despite it being present across multiple chunks.

This indicates that overly small chunks fragmented semantically related information, reducing retrieval effectiveness and causing relevant context to be missed.

Results:

B	A	Baseline	**מדד**
6.8	4.83	6.76	זמן תשובה ממוצע
600	1200	1200	Chunk size
No	Yes	Yes	האם תשובה מדוייקת
			בעיות טכניות

Part C

In this experiment, the local Ollama-based LLM was replaced with a cloud-based model using Hugging Face Inference. This change eliminated the need to download, manage, and run large models locally, significantly reducing local compute and setup complexity.

However, this architectural shift resulted in a substantial increase in inference latency, with the average response time rising to 32.92 seconds. This increase is primarily attributed to network overhead, remote request queuing, and the shared compute resources allocated to the hosted inference endpoint. Additionally, this approach introduces a dependency on network connectivity and therefore cannot operate offline.

In terms of output quality, the answers produced by the cloud-based model were comparable to those of the baseline local model, with no significant qualitative improvement observed.

Conclusion:
Cloud-based inference simplifies deployment and reduces local resource usage but introduces higher latency and loss of offline capability, while providing similar answer quality for this use case.

Part D

Context

On the first step I switched to using an embedding for each chunk as his. File: pdf-rag-clean-basic.py
On the second one I use contextual retrieval. File: pdf-rag-clean-context.py

Contextual RAG performed better on ambiguous or abstract queries because the added context helps retrieval understand *what the chunk is about*, not just *what words it contains*.
It typically retrieves the correct chunk in fewer attempts and improves answer accuracy.
Tradeoff: slightly higher indexing cost (LLM call before embedding), but no extra cost at query time. (avg response time for questions were similar)

Ranking:

First I switch to using similarity search (k=5) this is in file: pdf-clean-rank.py
On the second option I switched to similarity search (k=10) and used only the top 3 for the llm

On the second option I could clearly see that the answers were more in context and more aligned with the pdf content, the answers were more specific and provided example and quotes from the pdf file

Reranking is more accurate, especially when:

- Relevant chunks are not top-ranked by embeddings
- Chunks are semantically similar but contextually wrong

Tradeoff: higher latency and cost (extra LLM/reranker step).
Helps avoid common failure modes like: *Missed Top Rank, Relevant chunk not included in context*

Bottom line:
Basic retrieval is faster and simpler, but reranking significantly improves answer quality and reliability when retrieval alone is insufficient

Screenshot of executed Ollama commands: