PART 3- REPORT

About Data set:

This dataset was constructed on 1 Aug 2023, which contains news in English published from 10 May 2023 to 1 Aug 2023 from various sources.

Rationale for its selection:

The dataset was chosen because it contains recent information from May-August 2023
And from the assumption that the cohere model, which was trained in 2023, will not contain up-to-date information as we have chosen, and if so, not necessarily in the domain of news articles in the dataset.
After testing some questions and seeing that the standard model answers them incorrectly, we finally decided to use this dataset.

Anecdote examples of cases where the standard QA model failed, and the RAG pipeline succeeded:

1. Query:

   ```
   England beat Australia by how many runs in the fifth and
   final Ashes Test at The Oval in 2023?
   ```

   Answers:

   Standard model:

   ```
   196 runs.
   ```

   RAG model:

   ```
   England beat Australia by 49 runs in the fifth and final
   Ashes Test at The Oval in 2023.
   ```

   When the source of knowledge from which the Rag model drew the answer is:

   ```
   " Home SportsAshes 2023: Stuart Broad Gets Winning Send-Off
   As England Beat Australia By 49 Runs; Draw Series 2-2
   Veteran fast-bowler Stuart Broad got his perfect fairytale
   ending to a glorious cricketing career by picking the last
   two wickets as England beat Australia by 49 runs in the
   fifth and final Ashes Test at The Oval…"
   ```

2. Query:

   ```
   England beat Australia by how many runs in the fifth and
   final Ashes Test at The Oval in 2023?
   ```

Answers:

Standard model:

```
The specific issue that puts Wairarapa Hospital at risk
during a major earthquake is its proximity to the Wellington
Fault Line, which is one of the most active and dangerous
fault lines in New Zealand. The hospital is located in the
town of Masterton, which is in the Wairarapa region, just
north of Wellington. This region is known for its high
seismic activity and has a high risk of experiencing strong
earthquakes.\n\nThe Wellington Fault Line poses a
significant threat to the hospital's infrastructure and the
safety of its patients and staff. During a major earthquake,
the fault line could generate strong ground shaking, soil
liquefaction, and surface rupturing, which could severely
damage the hospital's buildings and disrupt its critical
services.\n\nThe age and structural design of the hospital
buildings are also factors that contribute to the risk. Some
of the hospital's buildings are older and may not meet the
latest seismic standards, making them more vulnerable to
damage or col.
```

RAG model:

```
Wairarapa Hospital is at risk during a major earthquake due
to the almost 3000 structural defects in the building, which
means it does not comply with the building code.
```

When the source of knowledge from which the Rag model drew the answer is:

```
Wairarapa Hospital could be at risk of failing in a major
earthquake, with almost 3000 structural defects found in the
building, which does not comply with the building code.
```

Describe the retrieval system and the prompts used in the RAG pipeline.

Given dataset and user prompt, firstly they were converted into vector using all-MiniLM-L6-v2 embedding model.

The retrieval system used in the RAG pipeline is a vector database, which stores and organizes vectors in a way that allows for efficient retrieval of similar vectors.
In our case, the vector database used is Pinecone and the

The formatted query used to retrieve documents from the vector database by measuring the similarity between the documents and the queries using cosine similarity metric, we choose to retrieve 3 documents, and then the user query and the selected documents are combined to form a prompt for the Cohere generation model to generate a coherent and contextually relevant response.

insights or observations on the effectiveness of the RAG pipeline in reducing hallucinations and improving answer accuracy:

Reducing hallucinations:

by using relevant documents as a basis for response the RAG pipeline grounds the generated content in factual and contextual information, and this helps the generation model to stay within the boundaries of the provided information, thereby reducing the likelihood of hallucinations.

Improving Answer Accuracy:

fetches the most relevant documents to ensure accurate information.
updates to the document store provide the latest data.
adds context to queries to better understand intent.

Git link :
https://github.com/AmraniBar/Lab-2-part-3.git