Task4

Question Generation using LLM (hugging face library)

Type of question generated

1.Analytical

2.Factual

3.Inferential

From documents provided by URL

https://www.3gpp.org/ftp/Specs/2023-06/Rel-18/23_series/23501-i20.zip

Approach:

I used RAG (Retrieval Augmented Generation) is a technique that combines a retriever and a generative language model to deliver accurate response. It involves retrieving relevant information from a large corpus and then generating contextually appropriate responses to queries. Here we use the quantized version of the meta Llama2 13B LLM(`model_name = "heBloke/Llama-2-13b-Chat-GPTQ"`)

with Lang Chain to perform generative Question Generation with RAG. The notebook file has been tested in Google Colab with T4 GPU.

Note:- Please change the runtime type to T4 GPU before running the notebook

Here's a high-level overview of how LLMs can work with document chunks and vector search(FAISS in my case)

1. **Chunking Documents**: Large documents are split into smaller, manageable chunks or segments. Each chunk contains a portion of the original document's content. The size of each chunk can vary depending on factors like the model's memory limitations and the desired granularity of analysis in my case it was 512(chunk size) and in the documents there are 6224 chunks has created.

2. **Vector Representation**: Each chunk of the document is converted into a vector representation using hugging face embedding technique to convert each word or token in the chunk into a high-dimensional vector representation.

3. **Vector Search**: Once the document chunks are converted into vector representations, I have used the vector database (FAISS) to store the embedding to find relevant information or chunks based on a query, we'll

first convert the query into a vector representation. Then, we will I use FAISS to search for the most similar vectors to the query vector in the index.

4. **Model Interaction**: In the context of LLMs like, we can interact with the model using the vector representations of the document chunks. For example, you can input the vector representations of the document chunks to the model and generate question based on the context provided by the chunks.

Prompt Structure:

For generating each type of Question its very important to give correct prompt to model

For that I first give some example to model that these are some example of each type of Question along with Documents you can see below how does prompt looks like

```
prompt_template = """
<s>[INST] <<SYS>>
Use the following context to Answer the question at the end. Do not use
any other information. If you can't find the relevant information in
the context, just say you don't have enough information to answer the
question. Don't try to make up an answer.
{context}

Analytical Question Example:
Question: How does the deployment of a Service Capability Exposure
Function (SCEF) contribute to the overall efficiency of NF service
discovery in the 5G Core network architecture? [/INST]

Factual Question Example:
Question: What is the purpose of a Notification Correlation ID in the
context of "Subscribe-Notify" NF Service interactions, as described in
the document? [/INST]

Inferential Question Example:
Question: Considering the described mechanisms for NF service
authorization, what potential challenges might arise in ensuring
seamless NF service access across different operator networks in a
roaming scenario? [/INST]


{context}
you have to generate question other than provided example
{context}
Question: {question} [/INST]
```

Also consider all examples are extracted from the documents itself not anywhere else

Below are some Question generated by the LLM model(llama 2 13b)

1. Analytical Question
    1. How does the deployment of a Service Capability Exposure Function (SCEF) contribute to the overall efficiency of NF service discovery in the 5G Core network architecture? [/INST]
    2. What is the purpose of a Notification Correlation ID in the context of "Subscribe-Notify" NF Service interactions, as described in the document? [/INST]
    3. Considering the described mechanisms for NF service authorization, what potential challenges might arise in ensuring seamless NF service access across different operator networks in a roaming scenario? [/INST]
    4 How does the usage data reporting mechanism for secondary RATs in the 5G core?
    5 network architecture support the optimization of QoS models and UDR management, considering the
    6 identified identifiers and their respective roles in the system?
    7 What potential challenges might arise in ensuring seamless NF service acess across different operator networks in a roaming scenario?
2. Factual Question
    1) What is the purpose of (Application Function) in influencing SMF routing decisions?
    2) How does the AF influence UPF (re)selection and (I-)SMF (re)selection?
    3) Can the AF request route user traffic to a local access to a Data Network (identified by a DNAI)?
    4) Is the AF allowed to access the network directly, or must it use the NEF to interact with the 5GC?
    5) What is the purpose of the local part of the DN (as defined in TS 23.548 [130])?
    6) What kind of events related to PDU Sessions can the AF request to be notified about?
    7) In the case of AF instance change, what information may the AF send in a request for AF relocation?
3. Inferential Question
    i. What is the purpose of the "PCF Selection" procedure described in point e) of the document ?4
    ii. Is it related to the selection of a specific PCF instance for a particular UE or PDU session?
    iii. How does the AMF determine which PCF instance to select for a given UE or PDU session, as mentioned in point f) of the document? Is it based on specific criteria or requirements?
    iv. What is the significance of the "PCF Group ID" provided by the AMF to the SMF ,as mentioned in point g) of the document? Does it play a role in the PCF selection process? Can the same PCF instance be selected for both

the UE and the PDU session, as suggested in point i) of the document? If so, what are the implications of this selection?

v. In the case of delegated discovery and selection in SCP, how does the AMF determine which NRF to use, as mentioned in point of the document? Is it based on specific criteria or requirements?

vi. What potential benefits might the SMF discovery in the 5G core network architecture ,considering the principles outlined in clause 6.3.1 and the specific factors considered in the discovery procedure with NRF?

Instructions for running the code on Colab:

I have done this on colab please change your runtime to T4GPU ,then run cell by cell, all the package necessary to installed are already there you just need one click.Also as My GPU has running out of time I was only able to generate few question again and again by running query we can generate more Question of different Type .For more accurate then the question I generated I need more context and example to feed into model to understand .