RAG System Evaluation on RAGBench Dataset

- Design the retriever using all the documents present in the dataset
  - Suggestion: use test dataset of each domain to reduce the computational overload
- For a given question in the dataset, get top k relevant documents
- Generate Response using LLM for a given Question and Relevant documents
- Use the (Relevant documents, Question and Response) to get the following attributes by prompting the same or a different LLM (refer to the prompt template used in RAGBench paper)
  - "relevance_explanation"
  - "all_relevant_sentence_keys"
  - "overall_supported_explanation"
  - "overall_supported"
  - "sentence_support_information"
  - "all_utilized_sentence_keys"
- Use these attributes to compute the metrics: Context Relevance, Context Utilization, Completeness and Adherence. (refer RAGBench paper)
- Compute RMSE and AUCROC by comparing with the original Context Relevance, Context Utilization, Adherence and Completeness scores in the dataset. (refer RAGBench paper)

Suggestions

1. Read and understand the RAGBench research paper thoroughly
2. Explore and understand the Dataset
3. Recommended to experiment with open-source LLMs
4. Use HuggingFace or Groq API’s to use LLMs
5. Refer to Best Practices in RAG paper to incorporate different techniques for each of the RAG components