Unit 4: Mini Hackathon-2

Retrieval Augmented Generation (RAG)

For this project, we will develop an RAG system that will answer the questions based on the knowledge base. We are going to use two assignment instruction documents [AST-1] and [AST-2] as a knowledge base. You can choose any other dataset/knowledge base as well. You are encouraged to use open source LLMs but you are free to use OpenAI API. Assignments on Langchain and RAG demo sessions will help complete this mini-project.

Steps to build RAG system:

Step 1: Load the document files (1 point)

Step 2: Split the loaded documents into chunks (1 point)

Experiment with different chunking method and parameters and check the final response

Step 3: Create embeddings and store in Vector Database (3 points)

Your are encouraged to use open source embedding models from mteb/leaderboard [Massive Text Embedding Benchmark (MTEB) Leaderboard]

Step 4: Perform the retrieval augmented generation by integrating with LLM (3 points)

Experiment with and without MMR and check the final response. It is encouraged to use open source LLMs.

Step 5: Frame at least five logical questions relevant to the knowledge base and demonstrate relevant answers from the RAG system (1 point)

Experiment with different combinations of different hyperparameters in all above steps to get good results.

Step 6: Create a Gradio App where user can write the query and get the response from the RAG system(1 point)

Step 7: Deploy the application on HF Spaces [Optional]