**Testing Different RAG Models: Strengths and Weaknesses**

In this project, the objective is to evaluate the performance of various Retrieval-Augmented Generation (RAG) models in document-based question-answering systems. RAG models integrate document retrieval with text generation, enhancing the quality and relevance of generated responses by grounding them in factual information. The focus will be on testing different RAG architectures and analysing their strengths and weaknesses. The performance of each model will be measured in terms of accuracy, answer relevancy, and context awareness when responding to user queries. Additionally, speed and computational efficiency will also be key considerations. Some models may excel at generating more precise and accurate answers, while others might retrieve information more efficiently, depending on the complexity of the questions and documents. This comparative analysis will highlight the trade-offs between different approaches, such as whether a model is better at handling longer documents or excels at delivering quicker responses. By identifying these strengths and weaknesses, this project aims to recommend the most effective RAG model for specific use cases, depending on the needs of the application.

**Project Plan Overview**

* **Data Collection & Cleaning** (Oct 10 - 24):  
  Gather and review the documents and question files. Clean the data to ensure consistency and prepare it for the next phase.
* **Data Pre-processing** (Oct 25 - Nov 7):  
  Perform text tokenization, generate embeddings, and normalize the dataset. Pre-process the data for efficient retrieval and model input.
* **Database & Architecture Setup** (Nov 8 - 21):  
  Set up a retrieval database (e.g., Elasticsearch or Pinecone) and design the architecture that integrates the RAG model for document retrieval and response generation.
* **LLM Selection** (Nov 22 - Dec 5):  
  Test various large language models to determine the most suitable one based on performance, relevancy, and speed for the RAG system.
* **RAG Testing & Model Evaluation** (Dec 6 - 19):  
  Evaluate different RAG models based on answer relevancy, context awareness, accuracy, and latency. Identify strengths and weaknesses.
* **Final Testing & Refinement** (Dec 20 - Jan 7):  
  Refine the selected models based on the evaluation results, improving performance and stability for final deployment. Prepare the research report.

**Data Sources:**

<https://www.kaggle.com/datasets/thedevastator/rag-financial-legal-evaluation-dataset>

<https://www.kaggle.com/datasets/samuelmatsuoharris/single-topic-rag-evaluation-dataset>