PREREQUISITES NEEDED TO RUN THE CHATBOT

* USE GOOGLE COLAB TO RUN THE NOTEBOOK
* ACTIVATE THE GPU MODE ( T4 GPU ) FOR FASTER INFERRENCE
* **HUGGING FACE TOKEN** : GET THE HUGGING FACE TOKEN FROM YOUR HUGGING FACE ACCOUNT AND GET THE ACCESS FOR THE GATED REPOSITORY OF THE **META-LLAMA/LLAMA-3.2-3B-INSTRUCT** MODEL.
* **NGROK ACCOUNT :**OBTAIN THE NGROK AUTH TOKEN TO EXPOSE THE APP PUBLICALLY.

INSTALL THE DEPENDENCIES

SETTING UP THE AUTHENTICATION TOKEN

* IN THE CODE YOU NEED TO REPLACE "INCLUDE YOUR HF TOKEN HERE" IN THE load\_models FUNCTIONS.
* REPLACE THE "INCLUDE YOUR NGROK AUTH TOKEN HERE" WITH YOUR NGROK AUTH TOKEN

NOW RUN THE CELL AFTER SETTING UP ALL THE ABOVE DEPENDENCIES

* A PUBLIC URL WILL BE GENERATED USING THE 8501 PORT AND THE LINK WILL APPEAR AS OUTPUT , CLICK ON IT AND JUST FOLLOW TO THE STREAMLIT WEBPAGE
* UPLOAD ANY PDF CONTENT AND START ASKING QUESTIONS TO IT

TECHNICAL WRITEUP

* TO DEVELOP THE FRONTEND THE LOW CODE STREAMLIT HAS BEEN USED. FOR THE LLM RESPONSE GENERATION I HAVE USED LLAMA 3.2 3B INSTRUCT MODEL.
* THE SENTENCE TRANSFORMER MODEL IS BAAI/BGE-M3 MODEL WHICH HAS THE CAPACITY FO GENERATING DENSE EMBEDDINGS (1024 DIMENSIONS) , DOCMENT CHUNKS.
* USED QDRANT VECTOR DATABASE TO PERFORM THE SIMILARITY SEARCH
* PYPDF2 HAS BEEN USED TO EXTRACT THE TEXT FROM THE PDF
* RANK -MB25 HAS BEEN USED TO PERFROM THE SPARSE RETRIVAL ALONG WITH THE DENSE VECTOR SEARCH . TO IMPROVE THE RETRIVAL PROCESS
* NGROK HAS BEE USED SO THAT THE CODE CAN BE RUN EASILY IN COLAB USING THE GPU AND ALSO THE WEBPAGE CAN BE RUN IN A PUBLIC URL

RESPONSE STRUCTING APPROACH

* RAG PIPELINE HAS BEEN IMPLEMENTED
* DOCUMENT PROCESSING IS DONE WHERE IN THE TEXT IS EXTRACTED FROM PDFS, AND IS SPLIT IN CHUNKS ( 700 CHARS WITH 200 OVERLAP).
* EMBEDDING IS DONE USING THE BGE-M3 MODEL AND IS STORED IN QDRANT VECTOR DB
* RETRIVAL APPROACH IS A HYBRID METHOD WHICH USES THE DENSE VECTOR METHOD ( QDRANT ) AND THE SPARSE BM25 RANKING APPROACH TO FETCH THE TOP 5 RELEVENT CHUNKS WITH A COSINE SIMILARITY OF 0.85
* FOR RESPONSE GENERATION THE LLAMA 3.2 3B MODEL HAS BEEN PROMPT TUNED IN A WELL STRCUTURED MANNER SO AS TO GET THE BEST RESPONSES

PROMPT DESIGN

* PROPMT TUNED TO EXPLAIN EACH CONCEPT PROPERLY , ALSO TUNED TO MAKE A CONCISE SUMMARY
* RULES HAVE BEEN IMPLENETED TO NOT TO USE ANY EXTERNAL KNOWLEDGE , AND NO REPETITION
* TEMPERATURE OF THE MODEL HAS BEEN SET TO 0.3 SO THAT MODEL DOESN’T DEVIATE MUCH ALONG WITH REPETITION\_PENALTY=1.2

CHALLENGES FACED :

* TO GENERATE THE APPROPRIATE RESPONSE FINE TUNING THE PARAMETERS OF THE RETRIVAL AND THE PARAMETERS OF THE MODEL WAS TOUGH
* GETTING THE ACCESS OF THE GATED REPO MODELS
* PLAYING WITH DIFFERENT RETRIVAL TECHNIQUES TO GET THE PERFECT RESPONSE