## **AI Engineer (Level-1) — Technical Assessment**

### Develop a Simple Multilingual Retrieval-Augmented Generation (RAG) System

**Objective**

Design and implement a basic RAG pipeline capable of understanding and responding to both English and Bengali queries. The system should fetch relevant information from a **pdf document** corpus and **generate** a meaningful answer grounded in retrieved content.

### **Core Task**

* Build a basic RAG application that:
  + Accepts user queries in English and Bangla
  + Retrieves relevant document chunks from a small knowledge base
  + Generates answers based on the retrieved information
* Build a knowledge base
  + Use the following Bangla Book - [HSC26 Bangla 1st paper](https://drive.google.com/file/d/19h7t3xVNBS5KR9o2sLMh3i2uaje7ww7n/view?usp=drive_link)
  + Proper Pre-Processing & data cleaning for better chunk accuracy
  + Document Chunking & Vectorize
* Maintain Long-Short term memory
  + "Short-Term" : Recent inputs in the chat sequence
  + "Long-Term" : Pdf document corpus in vector database

### Sample Test Case:

User Question: অনুপমের ভাষায় সুপুরুষ কাকে বলা হয়েছে?

Expected Answer: শুম্ভুনাথ

User Question: কাকে অনুপমের ভাগ্য দেবতা বলে উল্লেখ করা হয়েছে?

Expected Answer: মামাকে

User Question: বিয়ের সময় কল্যাণীর প্রকৃত বয়স কত ছিল?

Expected Answer: ১৫ বছর

### 

### 

### 

### 

### **Bonus Tasks**

#### Simple Conversation API

* + Build a lightweight REST API to enable interaction with the RAG system.
  + Endpoint should accept user input and return model-generated responses.

#### RAG Evaluation

* + Implement a basic evaluation of your RAG system using any of the following: You may use simple metrics (e.g., cosine similarity scores, human-labeled examples, etc.)
    - Groundedness (Is the answer supported by retrieved context?)
    - Relevance (Does the system fetch the most appropriate documents?)

Use Industry-Standard Tools and Practices .Tools & libraries are not limited to Langchain/Langflow/N8n/diffy/other. You can use any vector database (postgres/mongodb/pinecone/other) and LLM model (openai/gemini/mistrail/ollama/other).

**Submission Requirements**

* Source code (on GitHub Public Repo) & README file with:
  + Setup guide
  + Used tools,library,package
  + Sample queries and outputs (Bangla & English)
  + API Documentation (if implement)
  + Evaluation Matrix (if implement)
  + Must Answer following Questions
  + What method or library did you use to extract the text, and why? Did you face any formatting challenges with the PDF content?
  + What chunking strategy did you choose (e.g. paragraph-based, sentence-based, character limit)? Why do you think it works well for semantic retrieval?
  + What embedding model did you use? Why did you choose it? How does it capture the meaning of the text?
  + How are you comparing the query with your stored chunks? Why did you choose this similarity method and storage setup?
  + How do you ensure that the question and the document chunks are compared meaningfully? What would happen if the query is vague or missing context?
  + Do the results seem relevant? If not, what might improve them (e.g. better chunking, better embedding model, larger document)?