Educational Chatbot

Timeline for this project is between 8 to 12 weeks.

Phase 1: Data Preparation

1. Collect curriculum materials

1. Download official textbooks (PDF) from the Ministry of Education’s website for all primary school subjects.

2. Optionally add external books (e.g., "Al-Emtihan", "Al-Adwaa") later.

2. Extract text from PDFs

1. Use Python libraries such as pypdf or pdfplumber to extract clean text.

3. Dividing Data into Chunks

1. Using Langchain text splitter

2. Ensures Efficient retrieval and embedding.

Phase 2: Converting Data

1. Converting Data from text to Vectors

1. Embeddings : Generate embeddings using OllamaEmbeddings (as example).

2. Store The chunks with Metadata(Point 3).

2. Vector Database

1. We can start with FAISS which used in facebook.

2. We use this sothat we can compare betweeen texts.

3. Metadata Tagging

1. Adding bonus information like {Grade, Subject, Chapter}.

2. This helps us in filtering results by subject/grade.

Goal of this phase

When student asks "هو 6*5 بكام ؟".Now chatbot searches in Vector (FAISS) DB using filtering by "Maths" and "Grade".Then chatbot will find the result and send it to LLM Which sends it to User.

Phase 3: RAG Pipeline

1. LLM

1. Start with a pre-trained model available in Falcon, Ollama, OpenAI, Groq or Gemma.

2. Retriever

1. Query goes → Vector DB retrieves top-k relevant chunks → pass to the LLM

3. Prompt Engineering

1. Adding Explanation prompt to LLM to know what should be done.

2. Example for a Prompt :

   You are a smart educational assistant for Egyptian students. Answer in simple Arabic (Egyptian dialect).
   If the question is about the school curriculum, answer accurately from the textbooks.
   If the question is outside the curriculum, respond: ‘This question is not part of the curriculum.’

Phase 4: Frontend

1. Basic web app

1. Simple UI/UX using chanlit (as example).

2. UI Features

1. Dropdown menus: (Subject - Grade).

2. "Listen to Answer" button using Text-to-Speech (TTS).

3. Suggestions Section to Provide User's experience.

Phase 5: Testing & Quality Assuranceon

1. Internal Testing

1. Create a set of test questions from each subject and grade.

2. Compare chatbot answers against official textbooks.

3. Document mismatches between expected and actual answers.

2. Error Logging

1. Implement a logging system to capture:
   1. User question.
   2. Retrieved content/chunks.
   3. Final model response.
   4. Whether the response was correct or incorrect.

2. Use Logs to identify weak areas.

3. Feedback Loop & Improvements

1. Analyze error logs + user feedback.

2. Improve :
   1. Data preprocessing (cleaner chunks)
   2. Embeddings (better chunk size)
   3. Prompt engineering (clearer instructions to the LLM)

3. Re-test until accuracy is consistently high.

Phase 6: Expansion

(Once MVP is stable)

1. Expand to preparatory and secondary levels.

2. Integrate external educational resources.

3. Make Collaborations with MOE and other books producers like "Al-Adwaa" or "Al-Emtihan".