Multilingual Knowledge Extraction & Exploration Assistant

Problem Statement:

You are given a **digitized (scanned or OCR-ed) copy** of a multilingual, handwritten textbook. Your goal is to **design and prototype an AI assistant** that can:

- Understand and structure the knowledge inside the book.
- Allow for flexible exploration or questioning,
- Handle noisy or multilingual content.
- Use GenAI tools appropriately to augment understanding.

Key Modules to Implement (Modular Grading):

1. Document Understanding + Preprocessing (Core)

- Input: Sample textbook file.
- Task: Use OCR / vision-language / layout-aware models to extract and clean the content.
- Bonus: Support Gujarati/Sanskrit and messy handwriting.
- Output: Clean text/markdown-like structure preserving hierarchy.

2. Chunking + Embedding Strategy

- Task: Break the content into meaningful chunks for future indexing or analysis.
- Suggest a chunking method (e.g., structure-based, semantic, etc.)
- Choose a multilingual embedding strategy

3. Question Answering or Exploration Tool (Stretch Goal)

- Task: Enable basic QA or exploration over the parsed book.
- This could be as simple as: ○ Retrieve passages by keyword.

○ Or run small LLM-based answers over structured content.

Final Deliverables:

Please submit the following:

1. A Python notebook or GitHub repo with: ○ Your code,

○ Sample outputs,

○ Any custom scripts or helper functions.

2. A **README or report** that clearly outlines: ○ The models, tools, and techniques you tried,

○ What worked well and what didn't?

○ How you handled multilingual or handwritten content,

○ Trade-offs or design decisions you made,

○ What you would improve or extend with more time.