**Breaking\_down\_NCERT\_concepts**

**Context:** So far, the agent seems to working very well. Flow, RAG, content seem to be in place.

Refer:

* <https://docs.google.com/document/d/1Kl_Fi7d7asrxjyJarllO8cSs8NfZwCSzrKKDBvyTx_s/edit?usp=sharing>
* <https://educative-rag-agent.streamlit.app/>

What next: Inorder to enable this agent for all NCERT Science concepts for 7th std, we need to break the concepts and store in a DB such that retrieval should be easier. For retrieval, we might or might not use vector search. Maybe dictionaries are sufficient. Either way we need to break the concept and store in a db. Along with images and youtube links.

Steps involved:

1. Extract the concept, **image** and it’s description as given in the NCERT pdf. (Pdf is already extracted. Refer: <https://docs.google.com/document/d/1zKIhgYu2duoPSEbLKkKc-A7-PGF3aIV8dpOw-0dc1is/edit?usp=sharing>) (Explore libraries to extract text and image)
2. Remove any activities, Fascinating facts, Experiments, Dive Deeper, Exploratory projects, Let us enhance our learning, In a nutshell.
3. Send only the concept content into LLM Gemini 2.5 using a prompt. (prompt given in the end.). Note: Image should be tagged along with the concept.
4. Get the response in json format
5. Store in Mongodb (Explore and come up with DB Schema)
6. Repeat for all Chapters in Class 7th Science Textbook

Reference:

* <https://chatgpt.com/c/68171d0b-dec8-8007-9660-0952af2709ad>
* <https://chatgpt.com/c/68ab0a18-00cc-8330-9c77-a0e6b60edf44>

### 

### 

### Tool Options:

### PyMuPDF (fitz) or pdfplumber for text extraction.

### OCR (Tesseract) for pages with embedded images or scanned text.

### 

**Prompt that can be used:**

Break the description into Detail Working Intuition / Logical Flow Critical Thinking Open-Ended MCQs Real-Life Applications Relation Between Sub-Concepts Cross-Concept Critical Thinking Exam-Oriented Questions Also, metrics estimation Estimated Avg time taken for the 1st time learners: 10 min Estimated Avg time taken for the 2nd time learners: 3 min # of revisions needed to master: 3 (interval of revisit in days: 5) Importance for the rest of the chapter: 10% Importance for the Olympiad: Rare (1 out of 500 question) Importance for the IIT/NEET: Very very rare (1 out of 50,000 question) Definition getting asked in the unit test: low Which aspect is important? Def/concept/word problem…? Should you remember this?

**Retrieval Design:**

Option A: Dictionary-Based

* Store {chapter -> concept\_name -> details} in MongoDB.

Option B: Vector Search (Optional)

* Use embeddings (e.g., OpenAI text-embedding-3-small) for key\_idea + explanation fields.
* Store vectors in Pinecone/Weaviate/FAISS for semantic retrieval.