ISSUES FACED

1-Data processing:

Issue	Fix
Extracted text was just a header	Changed extraction to line-wise processing instead of slicing a flattened string; picked last occurrence of the matching header to get main content
Content started after newlines or pages	Extracted all lines between header lines instead of slicing by character position
Misleading indexes due to text normalization	Used line indices to slice content, avoiding normalized text index issues

2-Noise reduction:

Problem	Fix
Nearly all chunks filtered out due to presence of noisy text.	Pre-clean extracted text by removing known noisy patterns.
Chunk size too small leading to fragmented chunks	Increase chunk size (~800 chars) for more meaningful chunks.
Filtering too aggressive on chunks containing some noise text	Relax the noise filter; exclude only obviously junk chunks.
Extraction slicing on normalized text causing index mismatch	Use line-wise extraction and pick last topic header occurrence.

3-Output after batch processing:

(.venv) PS F:\rdsharma_extractor> python .\src\batch_llm_extract.py
Loading checkpoint shards:
100%|[redacted]| 2/2
[00:00<00:00, 15.53it/s]
Device set to use cpu
Processing chunk 1/13...
Processing chunk 2/13...
Processing chunk 3/13...
Processing chunk 4/13...
Processing chunk 5/13...

Processing chunk 6/13...
Processing chunk 7/13...
Processing chunk 8/13...
Processing chunk 9/13...
Processing chunk 10/13...
Processing chunk 11/13...
Processing chunk 12/13...
Processing chunk 13/13...
Extraction complete. Saved all questions to F:\rdsharma_extractor\extracted_questions.tex

4-problems faced during cleaning the latex document:

Problem	Solution
Filter removed almost all chunks	Raw text pre-cleaning + relaxed noise filter
No output saved for LLM batch	Explicitly save filtered chunks as a text file
LLM batch slow	Wait for CPU, or switch to GPU/parallel/smaller chunks
Prompt echoes or malformed LaTeX output	Stripping prompt echoes during extraction/validation
LaTeX syntax warnings in validation	Checked are ignorable for macros like `\frac`; compiles fine