RAG System (Design & Explanation Document)

1.Chunk Size Selection. Chosen Chunk Size .

Chunk size: 500 words. Overlap: 50 words.

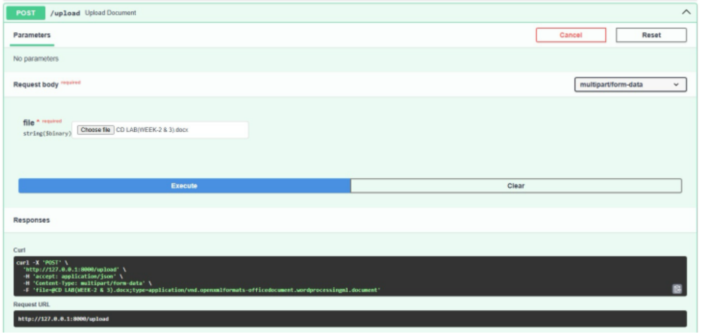
Why This Chunk Size Was Chosen ?

The chunk size was selected to balance semantic completeness and retrieval accuracy . SentenceTransformer embeddings (all-MiniLM-L6-v2) perform best when text chunks contain sufficient semantic context but are not too long. 500 words is obvious sufficient to preserve suggestive concepts and relationships within the text . Overlapping chunks (50 words) help prevent pivotal information from being split across chunks, reducing context loss during retrieval.

2. Retrieval Failure Case Observed

extremely general or wispy queries

Example:“Explain the document”



multipart/form-data

file: CD LAB(WEEK-2 & 3).docx

```
curl -X 'POST' \
  'http://127.0.0.1:8000/upload' \
  -H 'accept: application/json' \
  -H 'Content-Type: multipart/form-data' \
  -F 'file=@CD LAB(WEEK-2 & 3).docx;type=application/vnd.openxmlformats-officedocument.wordprocessingml.document'
```

```
http://127.0.0.1:8000/upload
```

3.Metric Tracked

```
{
  "latency_seconds": 0.21
}
```