PDFQandA Sample Document

This document demonstrates various content types for testing the ingestion pipeline.

It includes narrative paragraphs, bullet lists, tables described in text, figure captions, and footnotes.

The introduction references a concept called ’hybrid retrieval’ which combines vector search and keyw

Key Concepts

- Hybrid retrieval blends semantic and lexical signals.
- Citations must always be present when claims are made.
- The embedding dimension in this project is 3072, matching text-embedding-3-large.
- Chunking aims for roughly one thousand tokens with overlap to maintain context.

Table Overview

Table 1: Evaluation Metrics

Metric | Description | Baseline

Latency | Time to first token | under 3s

Accuracy | Relevance of answers | >= 0.85

Coverage | Percentage of sections covered | 92%

Observations: Latency remained stable while accuracy improved after adding HNSW indexing.

Graphics and Notes

Figure 2 illustrates the retrieval workflow: ingesting documents, embedding markdown, and storing me

Although the actual figure is not embedded, this caption ensures the graphics pipeline has text to anch

Nearby paragraphs mention the fallback to GIN-based filtering when vectors are inconclusive.

Footnotes and References

The system records footnotes detected near the bottom of pages.

It also captures references such as [1] Hybrid Retrieval Research Notes.

1) Hybrid retrieval references internal design discussions.

2) Citations are enforced for every answer.

Summary

Ingesting a PDF populates kb.documents, kb.sections, and kb.markdowns.

Asking a question triggers vector search for top twelve candidates, optional FTS refine, and passes si

Answers must include citations similar to doc:section L1-L3 to satisfy the hard-fail guard.

This final page confirms the integration of ingestion, retrieval, and citation enforcement.