

RAG Test Document (PDF)

Generated: 2025-12-21 01:43:54

Purpose: Test /ingest_pdf and /ask (RAG) with a known source document.

1) Key Facts

- Ollama default API base URL: http://localhost:11434
- Default Ollama API port is 11434.
- Example embedding model name: nomic-embed-text
- Example small generation model: llama3.2:3b
- ChromaDB stores embeddings for similarity search (vector database).
- RAG pipeline: Embed question -> Retrieve top_k chunks -> Generate answer using retrieved context.

2) Mini Glossary

- Embedding: A numeric vector representation of text for similarity search.
- top_k: Number of relevant chunks retrieved from the vector DB.
- Chunking: Splitting long text into smaller pieces for better retrieval.
- Overlap: Repeating a tail of a chunk at the start of the next chunk.

Sample Questions to Ask After Ingestion

- Q1) What is the default Ollama API port?
- Q2) What is the default Ollama base URL?
- Q3) Name one embedding model mentioned in this document.
- Q4) What does top_k mean in RAG retrieval?
- Q5) Summarize the RAG pipeline in one line.

Tip: Image-based PDFs need OCR; text-based PDFs work with PyMuPDF `get_text()`.