

AgentRAG Sample PDF Fixture

This document is used as a test fixture for the AgentRAG ingestion pipeline. It verifies that the reader module can extract text from PDF files using PyMuPDF.

The ingestion pipeline reads this file, extracts its text content, then chunks the text into overlapping token windows before generating dense vector embeddings with a local sentence-transformers model.

Semantic retrieval over PDF content is one of the primary use cases for AgentRAG. Users can ingest research papers, contracts, manuals, and other PDF documents, then query them using natural language through the MCP tool interface.

This fixture is intentionally kept simple: a single page of plain text. Complex PDF layouts, embedded images, and multi-column text are handled by PyMuPDF but are not required for the basic ingestion test to pass.