Memic RAG Pipeline Test Document for DOCX Conversion

====================================================

Introduction

------------

This document tests the DOCX file processing capabilities of the Memic RAG pipeline.

It will be converted to DOCX format and then processed through the entire pipeline.

Key Features

------------

- File Conversion using LibreOffice

- Document Parsing and Text Extraction

- Semantic Chunking with Overlap

- Vector Embedding Generation

- Full-text Semantic Search

Technical Architecture

---------------------

The pipeline consists of four main stages:

1. Conversion: LibreOffice converts documents to standardized PDF format

2. Parsing: Extracts text, metadata, and structure from documents

3. Chunking: Splits documents into semantic chunks with configurable overlap

4. Embedding: Generates vector embeddings for semantic search capabilities

Supported File Formats

----------------------

PDF - No conversion required, goes directly to parser

DOCX - Modern Word format, can be parsed directly

DOC - Legacy Word format, requires conversion to PDF

XLSX - Modern Excel format, can be parsed directly

XLS - Legacy Excel format, requires conversion

PPTX - Modern PowerPoint format, can be parsed directly

PPT - Legacy PowerPoint format, requires conversion

JPG/PNG - Images are converted to PDF with OCR

TXT - Text files are converted to PDF

MP3/WAV - Audio files are transcribed

Testing Results

---------------

This document serves as a test case for DOCX file processing through the RAG pipeline.

The system should successfully handle this file through all processing stages.

Expected Processing Flow:

1. Upload to Azure Blob Storage

2. Conversion check (DOCX may skip if parser supports it directly)

3. Text extraction and parsing

4. Semantic chunking into manageable segments

5. Vector embedding generation

6. Storage in vector database for search

Document Metadata

-----------------

Document Type: Technical Test Document

Format: DOCX (converted from TXT)

Version: 1.0

Created: October 2024

Purpose: RAG Pipeline Testing