FastRAG v3.0 - Architecture Overview

# System Components

The FastRAG system consists of several key components working together:

* **UniversalLoader:** Loads documents from multiple formats (TXT, MD, PDF, DOCX)
* **RecursiveCharacterTextSplitter:** Intelligently splits documents into chunks
* **MiniLMEmbedder:** Converts text into semantic vector representations
* **ChromaVectorStore:** Stores and retrieves embeddings efficiently
* **OllamaLLM:** Generates natural language responses (supports multimodal)

# Processing Pipeline

Document Ingestion:

1. Load documents using UniversalLoader
2. Split into chunks with RecursiveCharacterTextSplitter
3. Generate embeddings using MiniLMEmbedder
4. Store in ChromaVectorStore with metadata

Query Processing:

1. Convert user question to embedding
2. Search vector store for similar chunks
3. Construct prompt with context and history
4. Generate response using LLM
5. Return answer with source citations

# v3.0 New Features

**Smart Chunking:** Documents are split intelligently to preserve context while optimizing retrieval

**Multi-Format Support:** Process PDFs and DOCX files in addition to text files

**Multimodal Analysis:** Analyze images alongside text using the llava model

**Conversational Memory:** Maintain context across multiple turns in a conversation

# Configuration

Key configuration parameters can be set in the .env file:

* CHUNK\_SIZE: Size of each text chunk (default: 1000)
* CHUNK\_OVERLAP: Overlap between consecutive chunks (default: 200)
* DEFAULT\_LLM\_MODEL: Text generation model (default: llama3)
* DEFAULT\_MULTIMODAL\_LLM\_MODEL: Image analysis model (default: llava)
* DEFAULT\_TOP\_K: Number of chunks to retrieve (default: 3)