1 - Data

Collection of structured and unstructured documents from multiple sources + loading Data
techniques used: langchain, pypdf2

2- Document Preprocessing

cleaning and standardization of document formats for optimal processing efficiency **techniques used:** regex and text processing libraries

3 - Chunking & Embedding

Intelligent document segmentation and high-dimensional vector encoding for semantic understanding.
techniques used: langchain recursivecharactertextsplitter (for chunking) + openaiembedding (embedding) + faiss (to create vectordbs)

4 - Vector Database Storage

High-performance vector storage with optimized indexing for rapid semantic similarity queries
techniques used: FAISS, Qdrant

5 - Semantic Retrieval Module

Advanced semantic search identifying contextually relevant private clauses for each public document section.
techniques used: Cosine SimSimilarity

6 - LLM Analysis Engine

Large Language Model performs sophisticated comparative analysis using retrieved context and structured prompts
techniques used: LLM

7 - Gap Analysis & Reporting

Comprehensive analysis interpretation with structured findings.
techniques used: Web interface