DEVISIC AIDEVISERS

Knowledge Extraction & Retrieval

Step	Task	Tools / Models	
1. Data Ingestion	Text Documents	• PyMuPDF • PDFPlumber • Apache Tika • Azure Form Recognizer • Google Document AI • GPT-4.1 / GPT-4o] • Google Gemini 2.0 Flash / Pro	
	Image-Based Documents	• Tesseract • LayoutLM • TrOCR • EasyOCR • PaddleOCR • Donut • Google Vision OCR • Microsoft Azure OCR • Claude 3.5 Sonnet	
	Internal System Text	Working on it	
2. Preprocessing	• Cleaning & Normalization	• spaCy • NLTK	
3. Domain Understanding	• Named Entity Recognition (NER) • Key Phrase & Topic Extraction • Relation Extraction • Knowledge Graph Construction	• spaCy custom NER • HuggingFace transformers • Prodigy (annotation tool)	• RoBERTa-NER • BioBERT (medical NER) • FinBERT (finance NER) • med7 (medical entities) • BERT-base multilingual NER

4. Embedding Generation	• Generate Text Embeddings		• OpenAI text-embedding-3-large / 3-small • Llama 3 Instruct Models • e5-large-v2 • all-mpnet-base-v2
5. Vector Database	• Store Text Embeddings	• Pinecone • Qdrant • Milvus • Redis Vector Store • Weaviate • ChromaDB	
6. Retrieval System	• Semantic Search • Hybrid Search (Keyword + Semantic) • Contextual Retrieval (RAG)	• LangChain retrievers • LlamaIndex • Qdrant/Pinecone query APIs • Elasticsearch Hybrid Search	ReRankers: • Cohere ReRank • bge-reranker-large • cross-encoder/ms-marco-MiniLM
7. Content Generation (LLM Layer)	• Document Summarization • Q&A From Knowledge Base • Drafting New Content • Multi-step Reasoning		• GPT-4/5 • GPT-4o • Llama 3 70B • Claude 3.5 Sonnet
8. Continuous Learning & Feedback Loop	• Reinforcement learning from human feedback • Data augmentations • Active learning • Flow improvements		• Periodic fine-tuning of BERT/LLM models • Lightweight adapters (LoRA/QLoRA)