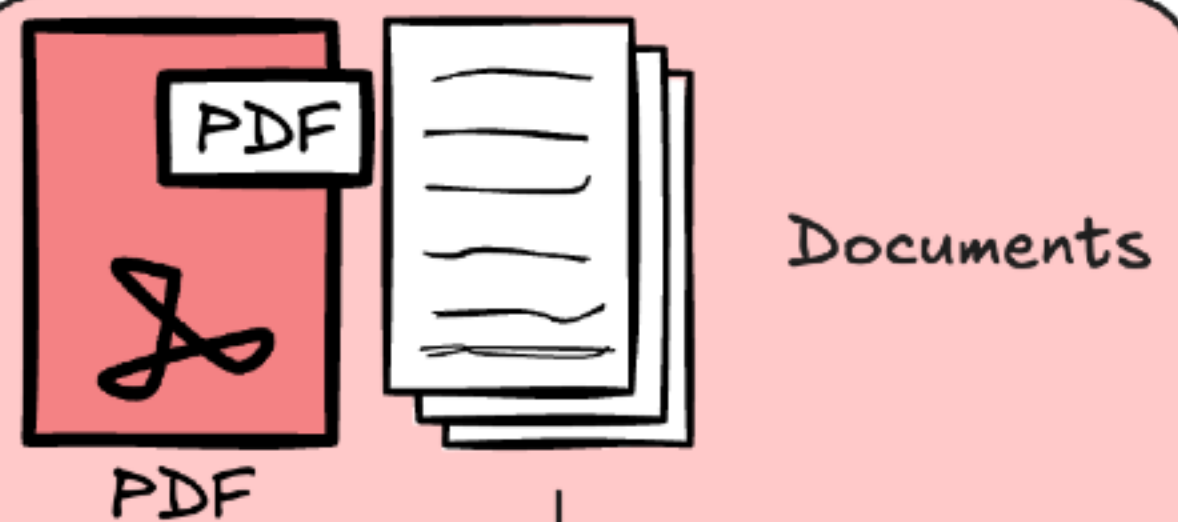
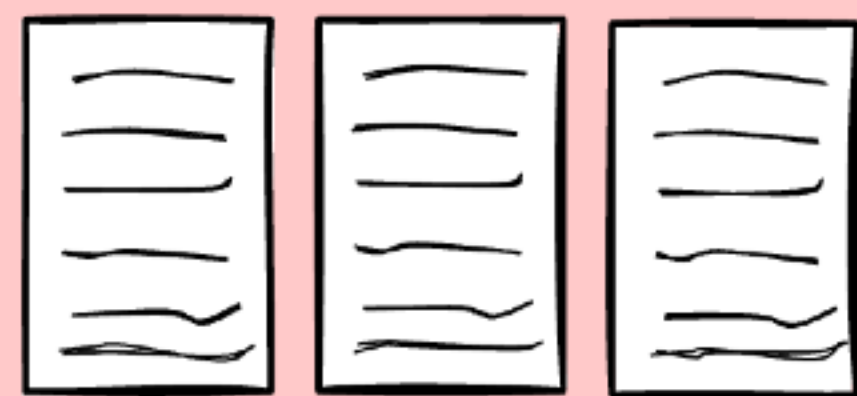


Batch



Parsing + Preprocessing



Chunking



Embedding Model

Vectorize

Indexing

Retrieve

Vector Store

Search

Embedding Model

Vectorize

Query

Client



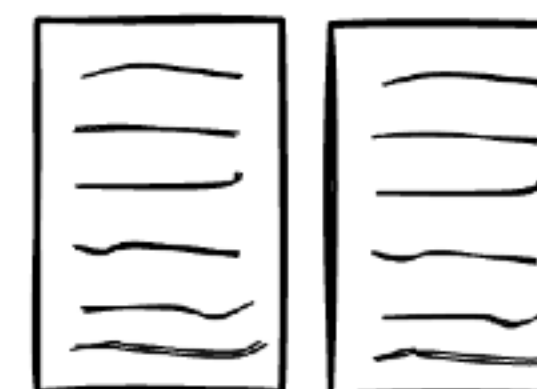
코드로 제어 가능한 부분
'좋은 문서'가 중요

파싱 전략은 이미
langchain이라는
LLM을 다루는 프레임워크의
API에 구현되어있기 때문에
문서에 알맞은 전략을 선택하여
파싱하는게 필요

Embedding Model과 Vector Store, LLM은
항상 변경가능한 구조



User Query
Relevant Docs Context



Prompt Engineering

RAG DEMO with OpenWebUI

- <https://github.com/open-webui/open-webui>

- Open WebUI is an **extensible**, feature-rich, and user-friendly self-hosted AI platform designed to operate entirely offline. It supports various LLM runners like Ollama and OpenAI-compatible APIs, with **built-in inference engine for RAG**, making it a powerful AI deployment solution.

<http://localhost:3000>

- 우리팀의 업무 프로세스는?
- 우리팀의 목표는?