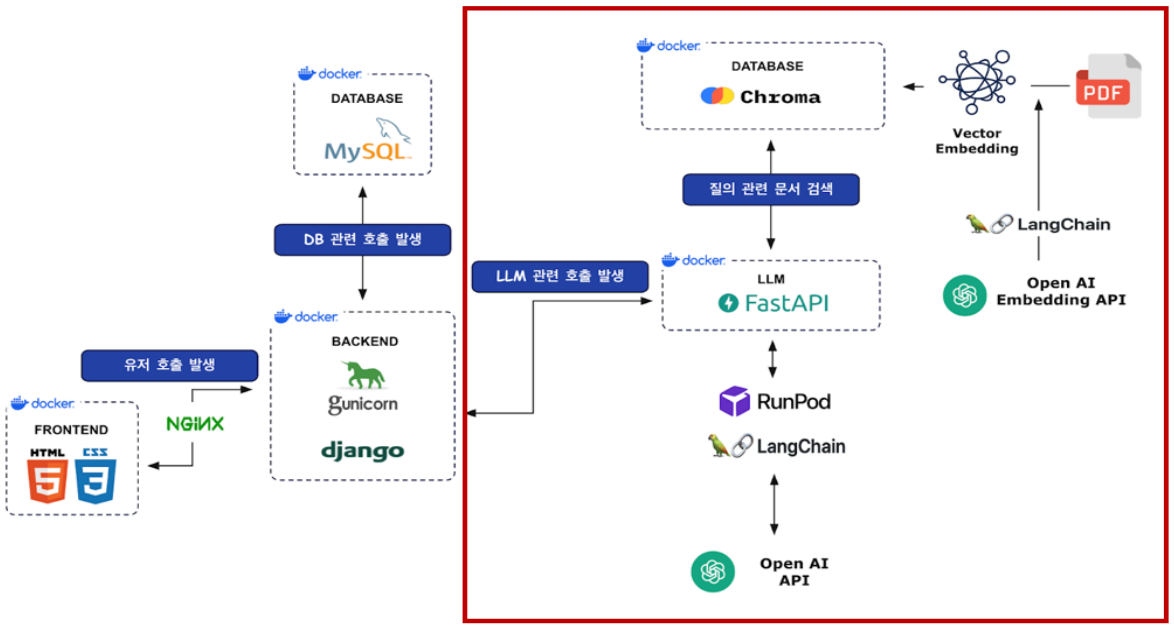
AI 푸드닥터(푸닥) 시스템 아키텍처

1. 개요

본 시스템은 문서 검색 및 LLM 기반 응답 생성을 담당하는 모듈로, ChromaDB, FastAPI, LangChain, RunPod, OpenAI API 등을 활용하여 AI 기반 검색 및 질의응답 기능을 수행한다.

2. 시스템 구성 및 동작 방식



(1) 문서 임베딩 및 검색 (ChromaDB + LangChain)

- **PDF 문서 처리 및 벡터 임베딩**
  - 문서를 LangChain을 사용하여 Chunk(청크) 단위로 분할
  - OpenAI Embedding API를 활용하여 문서를 벡터화
  - 벡터 임베딩된 데이터를 ChromaDB에 저장

- **질의 관련 문서 검색**
  - 사용자의 질의(Query)를 벡터화하여 ChromaDB에서 유사 문서 검색
  - LangChain의 Retriever를 활용하여 사용자의 질의와 ChromaDB 상에 저장된 벡터 임베딩 간의 유사도를 계산
  - 유사도가 높은 상위 N개의 chunk를 LLM의 입력 context로 활용하여 최적의 응답을 생성 (RAG)

(2) LLM 기반 응답 생성 (FastAPI + RunPod + OpenAI API)

- **FastAPI를 통한 AI 서비스 제공**
  - FastAPI 기반 LLM 서비스가 사용자 요청을 처리
  - 검색된 문서를 바탕으로 RunPod에서 실행되는 LangChain LLM 호출
- **RunPod을 활용한 고성능 AI 연산**
  - RunPod의 GPU 최적화 환경에서 sLLM 실행
  - OpenAI API 또는 자체 sLLM을 활용하여 자연어 응답 생성
- **최종 응답 반환**
  - AI가 생성한 답변을 FastAPI를 통해 반환
  - 백엔드로 최적화된 응답을 제공

3. 주요 기술 스택

기술 스택	역할
ChromaDB	문서 벡터 임베딩 및 검색
LangChain	문서 임베딩, 검색 최적화 및 LLM 연동
OpenAI Embedding API	문서 데이터 벡터화

FastAPI	AI API 서비스 제공
RunPod	GPU 기반 sLLM 배포
OpenAI API	GPT 기반 자연어 응답 생성