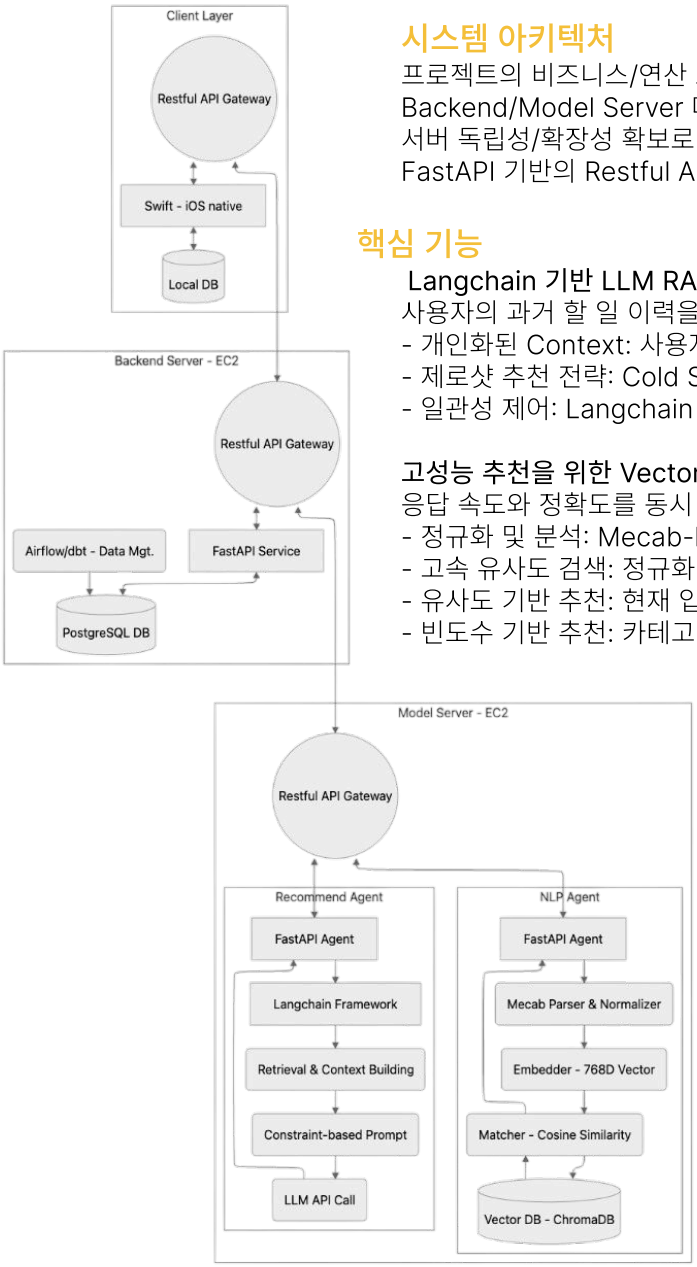
홍승완

www.hosugator.com

Dotodo

음성 입력 기반 인공지능 개인화 할 일 추천 앱 서비스



시스템 아키텍처

프로젝트의 비즈니스/연산 로직 분리를 위해 AWS EC2 Multi-Instance MSA 설계
Backend/Model Server 다중화로 트래픽 부하 분산 및 장애 격리
서버 독립성/확장성 확보로 LLM 모델/NLP 로직 변경에도 메인 백엔드 서비스 안정
FastAPI 기반의 Restful API JSON/HTTP 프로토콜 표준화로 통신 일관성 극대화

핵심 기능

Langchain 기반 LLM RAG 시스템 구현 및 최적화

사용자의 과거 할 일 이력을 활용하여 RAG 구축하고, 개인화된 할 일 추천 기능을 제공
- 개인화된 Context: 사용자의 최근 3개월간의 데이터를 프롬프트에 동적 삽입
- 제로샷 추천 전략: Cold Start Problem 방지로 일반적인 패턴을 명시적 주입
- 일관성 제어: Langchain 튜닝을 통해 창의성을 제어해 문맥상 일관성 있는 결과 보장

고성능 추천을 위한 VectorDB 및 NLP 파이프라인

응답 속도와 정확도를 동시 확보를 위해 코사인 유사도 기반 추천 시스템 병행 구축
- 정규화 및 분석: Mecab-ko 사용하여 입력값을 정제하고 일관된 형태로 정규화
- 고속 유사도 검색: 정규화된 텍스트를 고차원 벡터로 저장하여 실시간 코사인 유사도 계산
- 유사도 기반 추천: 현재 입력된 할 일과 유사도가 높은 과거 항목을 찾아 추천
- 빈도수 기반 추천: 카테고리별/시간대별 공통 빈도수 기반 추출로 추천 다양성 확보

문제 해결 및 성능 분석

Langchain 기반 RAG 시스템 응답 속도 최적화

Langchain RAG 시스템의 특성상 LLM 추론 시간으로 인해 응답 지연 발생

- 비동기 처리(Async/Await):

FastAPI의 비동기 기능을 활용하여, LLM API 호출 및 VectorDB I/O 작업을 병렬 처리하여 전체 응답 시간을 단축

- Context 크기 조정:

프롬프트에 삽입되는 사용자 이력(Context)의 크기를 최소화하여 LLM 추론 비용과 응답 시간의 균형화

추천 결과의 품질 및 유용성 개선

초기 추천 로직이 날짜 필터링 미작동 문제로 인해, 입력값과 동일 혹은 의미 없는 항목을 반환하여 유용성 저하

- 동일 항목 제외 규칙

로직을 수정하여, 사용자 입력값과 일치하는 과거 항목 제외하여 추천의 다양성 확보

- Top-K 제한

사용자에게 불필요한 정보를 제공하는 것을 방지하고 실질적인 선택을 유도하기 위해, 최종 추천 항목을 상위 3개로 제한

회고

AWS, Docker, MSA 분리, NLP Agent 등 현업에서 요구되는 핵심 기술 스택을 성공적으로 통합하고 구현하여, 복잡한 시스템 구축 역량을 효과적으로 입증할 수 있었습니다. 다만, 시간 및 협업의 한계로 인해 당장의 결과물에는 반영하지 못했지만, Agent의 고도화된 기능 개발에 대한 필요성을 확인했습니다. 향후 개선점으로 아래의 목표점을 잡고 추진 중에 있습니다.

- LLM 평가 시스템: `LLM as a judge` 개념을 도입하여, 추천 결과를 생성하는 LLM과 그 결과를 평가하는 별도의 LLM을 두어 추천 품질을 자율적으로 검증하고 개선하는 메커니즘을 구축할 수 있습니다.
- 다중 추천 로직 고도화: 현재의 코사인 유사도/빈도수 추천 외에, "패턴/문맥/연관성" 기반의 다중 추천 모듈을 세분화하여 (예: 운동 시 패턴 추천, 요리 시 연관성 추천), 상황별 추천 정확도를 비약적으로 향상시킬 수 있습니다.