SK네트웍스 Family AI과정 14기

모델링 및 평가 시스템 아키텍처

□ 개요

- 산출물 단계 : 모델링 및 평가
- 평가 산출물 : 시스템 아키텍처
- 제출 일자 : 2025. 09.12 .
- 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN14-FINAL-6Team

컴포넌트 다이어그램	• 구성 요소 • 설명
시스템 워크플로우	• 참여자 (Actors) • 주요 흐름

1. 컴포넌트 다이어그램

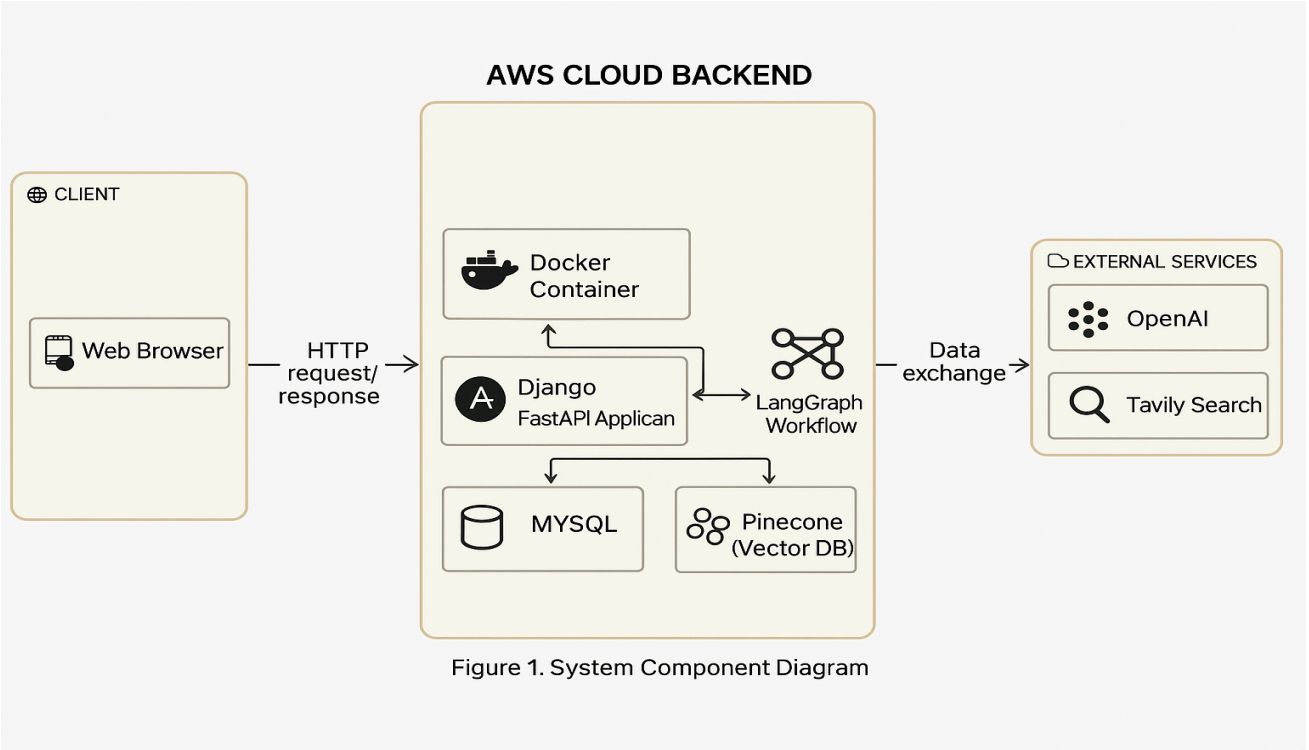


Figure 1. System Component Diagram

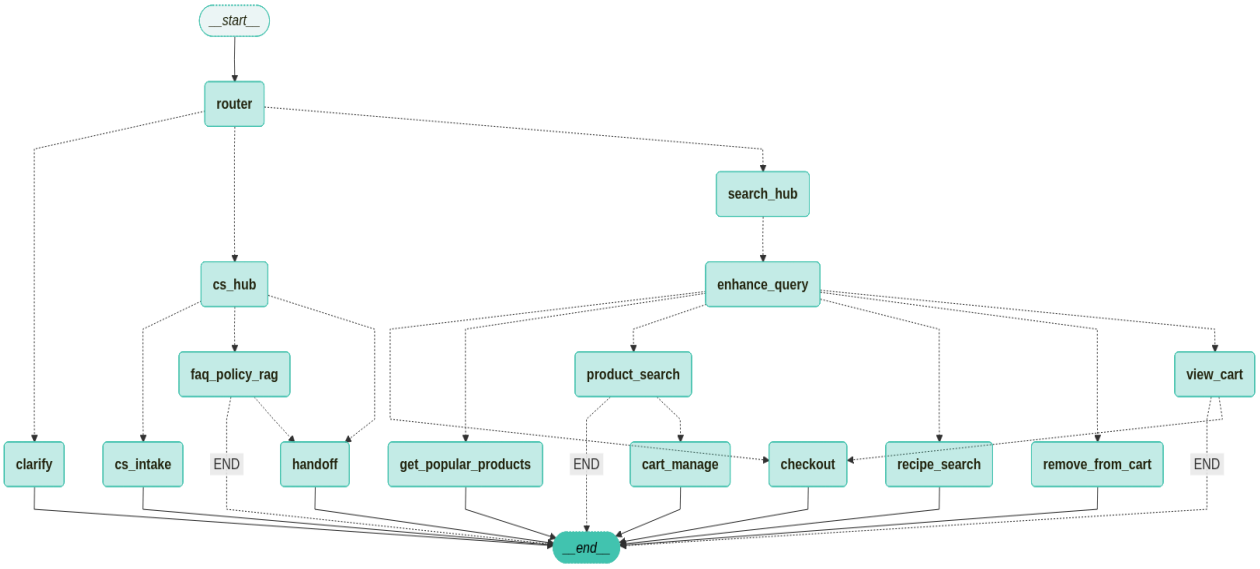
구성 요소

1. **클라이언트 (CLIENT)**
    - 인터페이스: HTTP request/response
2. **AWS CLOUD BACKEND**
    - Docker Container
    - FastAPI Application
    - LangGraph Workflow
3. **데이터베이스 (Database)**
    - MySQL (RDB)
    - Pinecone (Vector DB)
4. **외부 서비스 (EXTERNAL SERVICES)**
    - OpenAI
    - Tavily Search

설명

1. 클라이언트(Web Browser)는 HTTP 요청을 AWS CLOUD BACKEND로 전송합니다.
2. 백엔드의 Docker 컨테이너에서 실행되는 FastAPI 애플리케이션이 이 요청을 수신하여 처리합니다.
3. FastAPI는 LangGraph Workflow를 통해 비즈니스 로직을 실행하며, 이 과정에서 필요에 따라 MySQL(RDB) 및 Pinecone(Vector DB)에서 데이터를 조회하거나 저장합니다.
4. LangGraph 워크플로우는 OpenAI의 언어 모델 기능이나 Tavily의 검색 기능이 필요할 경우, 외부 서비스와 데이터를 교환합니다.
5. 모든 처리가 완료되면, 최종 결과를 HTTP 응답 형태로 클라이언트에게 반환합니다.

2. 시스템 워크플로우



구성

1. 시작 지점: 사용자가 쿼리를 입력하며 워크플로우 시작
2. 의도 분석 및 라우팅
    - 라우터가 사용자 쿼리 의도와 신뢰도를 분석합니다.
    - 신뢰도가 높으면 의도에 따라 Search Hub 또는 CS Hub로 작업을 전달합니다.
    - 신뢰도가 낮으면 Clarify 노드로 전달하여 사용자에게 질문을 명확히 하도록 요청합니다.
3. Search Hub 내부 활동
    - Enhance Query를 통해 사용자 쿼리를 검색에 용이하도록 구체화합니다
    - Product Search, Cart Manage, Recipe Search 등 쇼핑 및 정보 검색 관련 작업을 수행합니다.
4. CS Hub 내부 활동
    - CS Intake를 통해 CS 문의 유형을 먼저 분류합니다.
    - FAQ/Policy RAG 노드가 Pinecone 벡터 DB에서 관련 문서를 찾고, OpenAI LLM을 통해 자연스러운 답변을 생성하거나 이미지를 읽고 답변을 생성합니다.

5. 데이터베이스

- Search Hub와 CS hub는 작업 중 필요한 상품 정보, 사용자 데이터, 장바구니 내역 등을 조회하거나 저장하기 위해 MySQL(RDB)에 접근합니다.

6. 외부 서비스 호출
   - 필요에 따라 외부 검색 서비스인 Tavily를 호출하여 추가 정보를 얻을 수 있습니다
7. 결과 반환
   - 각 노드에서 처리된 결과는 사용자에게 텍스트 형태로 변환됩니다.
8. 종료 지점
   - 사용자에게 최종 응답이 전달되거나, 시스템이 해결할 수 없어 상담원 연결로 안내된 후 작업이 종료됩니다.

주요 액션 노드

1. 의도 분석 : 사용자 쿼리를 받아 의도를 파악하고 신뢰도에 따라 처리 경로를 동적으로 결정하는 핵심 분기점입니다.
2. 기능 수행 허브 : 명확한 의도를 가진 쿼리를 받아, 검색/쇼핑 기능 또는 고객 서비스 기능을 각각 전담하여 처리합니다.
3. RAG 기반 답변 생성 : CS Hub의 핵심 기능으로, Pinecone 벡터 DB와 OpenAI LLM을 연동하여 정책/FAQ 기반의 정확하고 자연스러운 답변을 생성합니다.
4. 데이터 상호작용 : 정형 데이터(MySQL)와 비정형 벡터 데이터(Pinecone)를 활용하여 워크플로우 전반에 필요한 데이터를 공급하고 결과를 저장합니다.