SK네트웍스 Family AI과정 14기 1Team

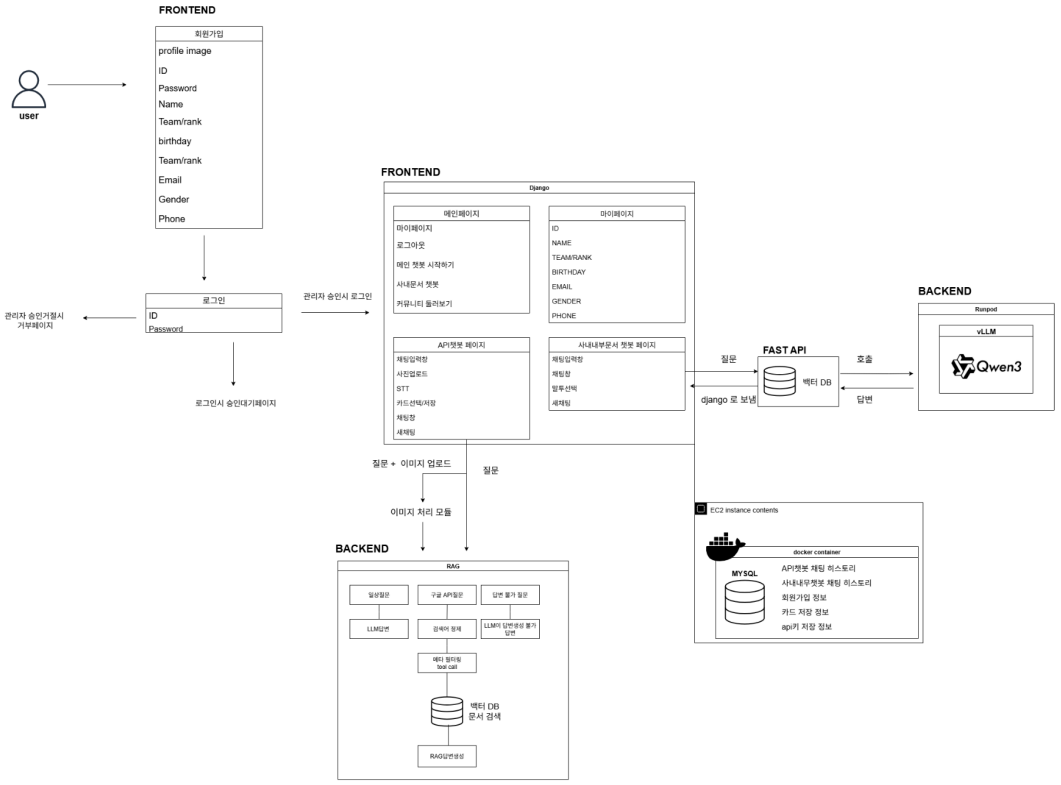
모델링 및 평가 시스템 아키텍처

☐ 개요

- 산출물 단계 : 모델링 및 평가
- 평가 산출물 : 시스템 아키텍처
- 제출 일자 : 2025. 09. 12.
- 깃허브 경로 : SKN14-Final-1Team
- 작성 팀원 : 이원지희

시스템 아키텍쳐 다이어그램	• 구성 요소 • 설명
프로세스 다이어그램	• 사용자흐름

1. 시스템 아키텍쳐 다이어그램



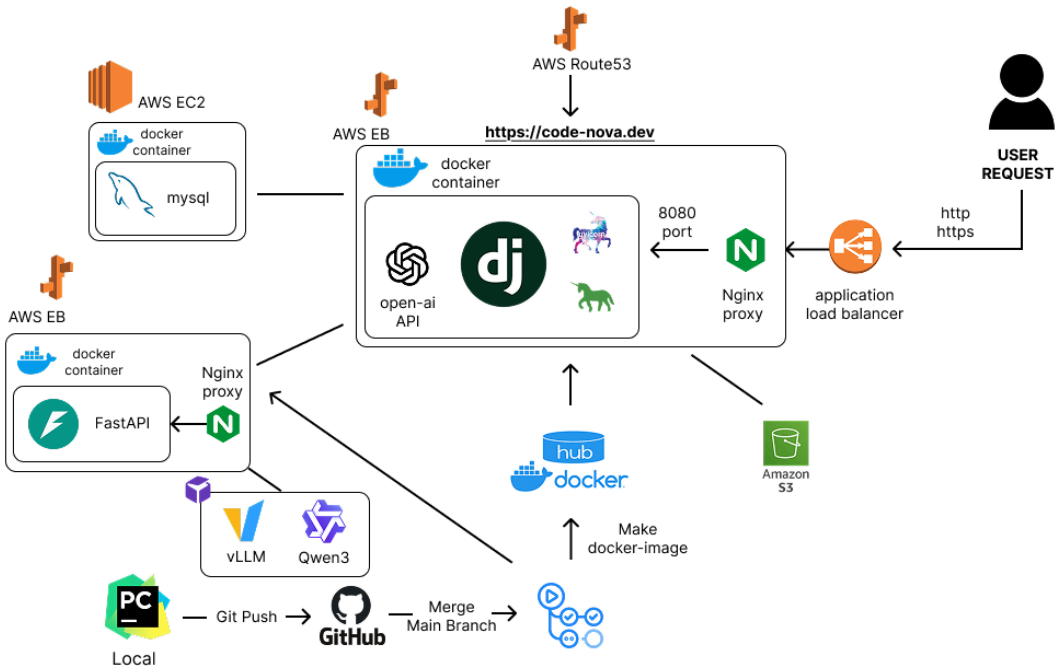


Figure 1. System architecture

구성 요소

1. 클라이언트(Client):

   - 컴포넌트:

     - 웹브라우저

2. LangGraph 모델:

   - Django 기반 LangGraph 애플리케이션
   - OpenAI 연동
   - Hugging Face 임베딩모델 bge-m3 모델 사용
   - Chroma Vector DB
   - Google APIs → 문서 수집/벡터화
   - Prompt Engineering (AI 모델 최적화)

3. 데이터베이스:

   Amazon EC2 (MySQL)

4. AWS Cloud:

   - AWS Elastic Beanstalk:
     - Nginx (리버스 프록시 및 정적 파일 제공)
     - Gunicorn (WSGI 서버)

        - Uvicorn (ASGI 엔드포인트)
        - Docker 기반 컨테이너 배포
        - Django (웹 애플리케이션)
  - EC2 instance contents:
      - FastAPI
      - LangChain
      - Chroma Vector DB (사내문서)
- S3
  - 프로필이미지
  - 채팅 이미지 첨부

5.외부서비스:
  - RunPod (sLLM 실행 환경)
      - Qwen3
      - vLLM
  - OpenAI API

설명

이 시스템은 AWS Elastic Beanstalk 위에 구동되는 Django 웹 애플리케이션을 중심으로 구성되어 있습니다. Django는 사용자의 질문을 받아 주제에 따라 두 가지 경로로 라우팅합니다.

첫 번째 경로는 구글 API 관련 질문입니다. Django 내부의 LangGraph가 질문을 임베딩(BGE-M3)하여 ChromaDB에서 관련 문서를 검색합니다. 검색된 문서와 질문을 조합해 프롬프트를 만들고, 이를 OpenAI LLM에 전달하여 답변을 생성합니다. 생성된 답변은 Django를 통해 사용자에게 전달됩니다. 이 경로는 공개 문서를 대상으로 하는 RAG 기반 처리 방식으로, 검색 품질은 임베딩과 문서 청킹에 크게 좌우됩니다.

두 번째 경로는 사내문서 관련 질문입니다. 이 경우 Django가 별도의 EB에서 운영되는 FastAPI 서버로 요청을 전달합니다. FastAPI는 LangChain 파이프라인을 통해 프롬프트를 구성한 뒤, RunPod의 vLLM 파드에서 서빙되는 Qwen3 파인튜닝 모델을 호출합니다. 이 모델은 사내 데이터에 최적화되어 있으며, FastAPI가 응답을 정리해 Django로 다시 전달하고 최종적으로 사용자에게 보여줍니다.

정리하면, Django는 게이트웨이 역할을 하며, 구글 API 질문은 RAG + OpenAI 모델, 사내문서 질문은 Qwen3 파인튜닝 모델로 각각 처리됩니다. 이 구조는 공개 지식과 내부 지식을 분리해 정확도와 보안을 확보하면서도, vLLM 기반 서빙으로 성능을 최적화할 수 있다는 장점이 있습니다.

2. 프로세스 다이어그램

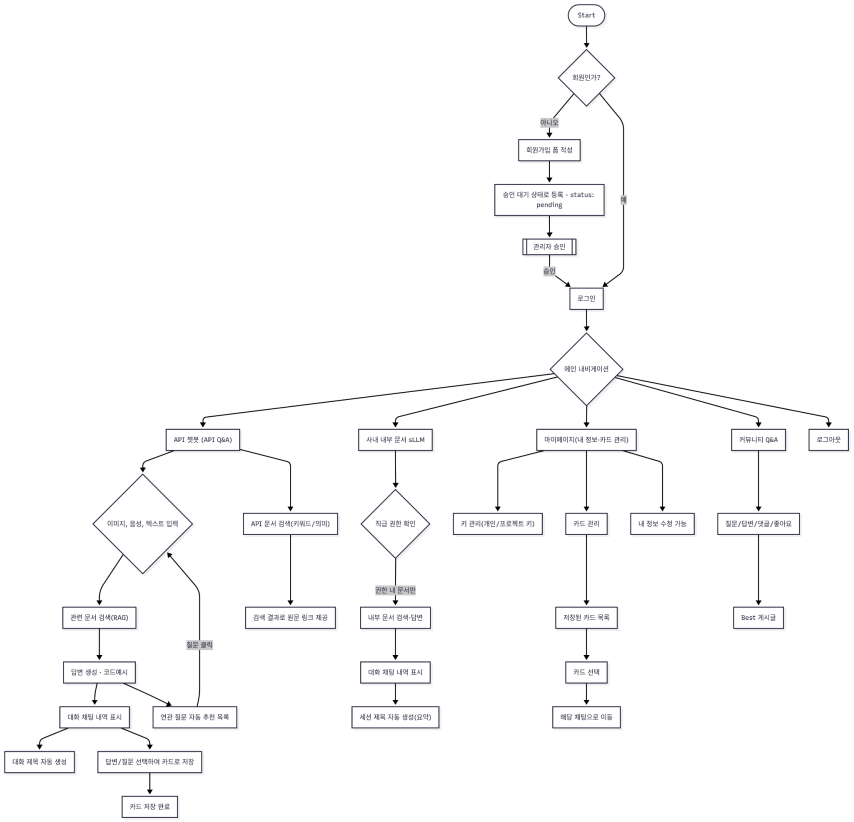


Figure 2. 프로세스 다이어그램 | 별첨1 사진 참조

참여자 (Actors)

1.회원가입자 / 사용자

시스템을 처음 사용하는 유저 (회원가입, 로그인, 내부 기능 이용).

로그인 후 내부 애플리케이션 기능(API Q&A, 문서 검색, 커뮤니티 등)을 사용하는 사람

2.관리자

회원가입 승인 권한을 가진 사람.

주요 흐름

1 회원가입 및 승인

- 사용자가 최초 접근 시 회원가입 여부를 확인한다.
- 신규 사용자는 회원가입 폼 작성 → 승인 대기 상태(pending)로 사용자 등록이 된다.
- 승인대기 상태로 로그인을 시도하면면 승인대기 페이지가 보인다
- 관리자가 거부시 거부페이지가 보인다.
- 관리자의 승인 후 정상적으로 로그인이 가능하다.

2. 마이페이지(회원 정보/카드 관리)

- 회원 정보 수정 가능
- 카드 관리(사용자가 저장한 채팅 메세지)
- Api 키 관리

3. 메인 챗봇 (API Q&A)

- 사용자가 질의 입력 시 관련 문서 검색(RAG) 수행
- 검색 결과 기반 응답 (이전 대화 맥락 반영)
- 연관 질문 자동 추천 기능
- 채팅 세션 제목 자동 요약 기능
- 채팅 메세지 선택 후 카드로 저장 가능(마이페이지에서 확인 가능)

4. API 문서 검색

- 검색 질의 입력 → 결과 원문 링크 제공

5. 커뮤니티 (Q&A 게시판)

- 질문/답변/댓글/좋아요 기능 제공

- Best 게시글 큐레이션을 통해 품질 높은 정보 공유

6. 사내 내부 문서 검색 (sLLM 기반)

- 직급/권한에 따라 검색할 수 있는 문서 차별화
- 채팅 세션 제목 자동 요약 기능
- 채팅 메세지 선택 후 카드로 저장 가능(마이페이지에서 확인 가능)
- 연관 질문 자동 추천 기능

7. 로그아웃

- 모든 활동이 종료되면 로그아웃한다.