자체 sLLM 개발 통한 기업 업무 활용 생성형 AI 플랫폼

프로젝트 기획서

Version : V.0.1
Last Update : 2024.12.20
팀원: 이호재, 안준용,
변가원
멘토: 김유진

목차

프로젝트 소개

1. 수행 목표

- 프로젝트 명 : 자체sLLM을 활용한 Q&A시스템 구축

- 목표 :
- 파인튜닝 적용한 sLLM을 활용해서 Agent형 RAG Application을 구축
- 직원들의 정보검색시간을 단축하여 업무 효율성 향상을 기대

1. 일정

- 24.12.06~25.02.04(8주)

1. 기대효과

- 향후 AI 기술을 확대 적용을 위한 토대 마련
- 회사 내 비정형 데이터들을 활용하여 신속한 업무정보 제공
- 회사의 업무 효율성 향상 및 시간 절약

- 반복적인 검색이나 사내 문서를 찾는 비효율적인 작업을 줄임으로써, 현업 관리자들이 자신의 핵심 업무에 더 집중할 수 있는 환경을 조성 - 자원 관리를 최적화하고, 회사의 생산성을 높이는 데 중요한 역할

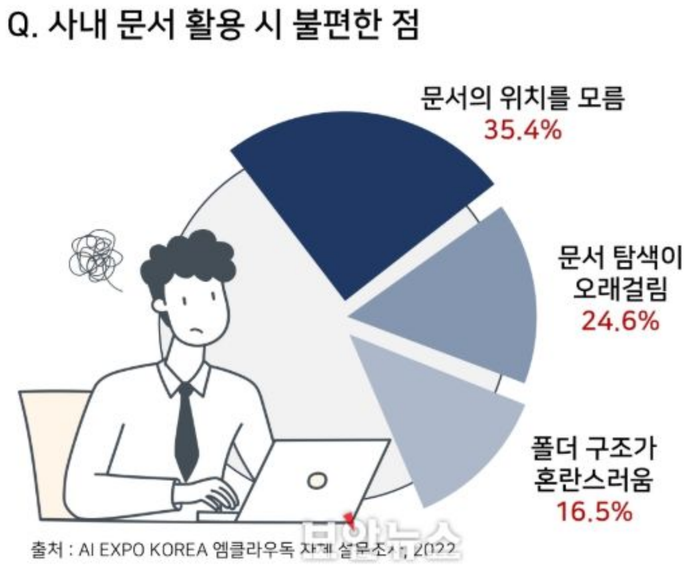
- 기업 경쟁력 향상

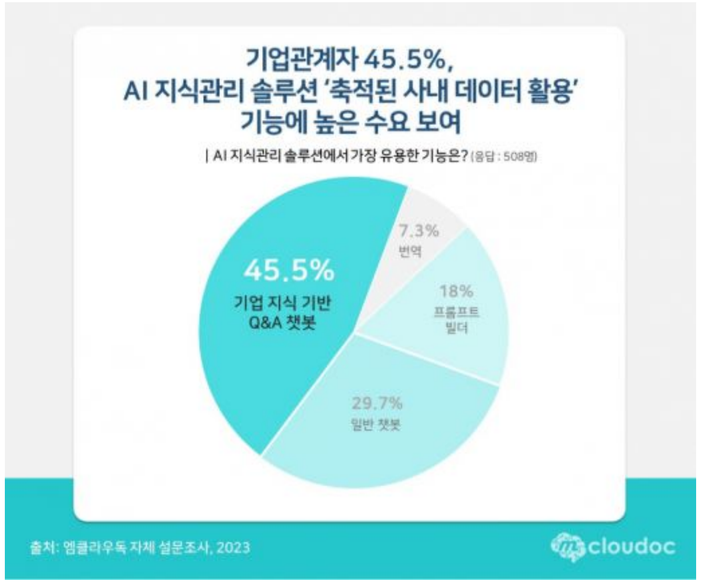
- 기업의 혁신적인 이미지를 강화하고, 고객 및 투자자의 신뢰를 통한 기업 경쟁력 확보

- 의사결정 도출 가능 - 데이터의 양이 많아질 수록 기업 데이터를 활용한 qa 시스템이 필수적, 더 좋은 의사결정 도출 가능
- 인건비 감소 - 기존 업무를 자동화하여 인건비를 줄일 수 있음

프로젝트 소개

1. 추진 배경 및 필요성





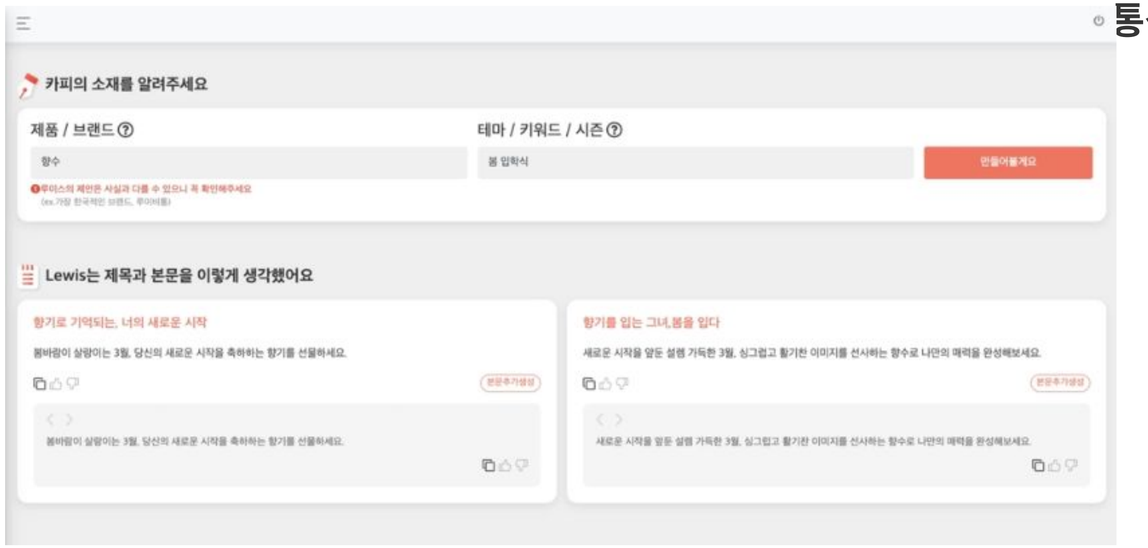
전자문서의 활용에 대한 과제는 현대 기업들의 고질적인 문제 중 하나이며 기존에 축적된 방대한 기업 문서를 인공지능을 통해 효과적으로 활용할 수 있는 기능에 높은 수요

- https://www.boannews.com/media/view.asp?idx=106452

프로젝트 소개

1. 추진 배경 및 필요성

1.현대백화점 루이스(Louis)



통상 2주가량 소요되던 카피라이팅 업무시간 -> 루이스 도입 후 평균 3~4시간 내로 줄음

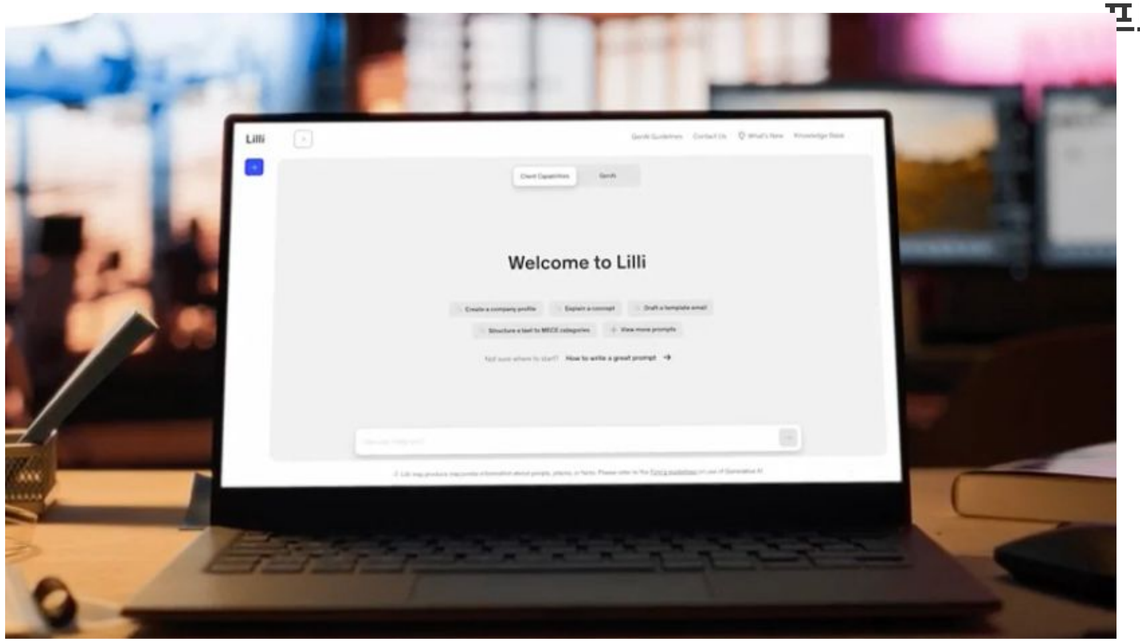
- 최근 3년간 사용한 광고 카피, 판촉 행사에서 쓴 문구 등에서 고객 호응을 얻었던 데이터 1만여 건을 집중적으로 학습 -> 현대백화점만의 색깔을 담아 작성 가능
- OpenAI GPT-3 대비 한국어 데이터를 6,500배 이상 학습한 초거대언어모델(LLM) -> 핵심 키워드를 조합하고 타깃 연령대를 고려해 문구의 톤과 어투까지 조절

- https://www.yna.co.kr/view/AKR20230224138700003?utm_source=chatgpt.com
- https://aiheroes.ai/community/172

프로젝트 소개

1. 추진 배경 및 필요성

1. McKinsey 릴리(Lilli)



프로젝트 초기 단계를 2주에서 몇분으로 단축 지식 검색 및 종합에 30%의 시간 절감

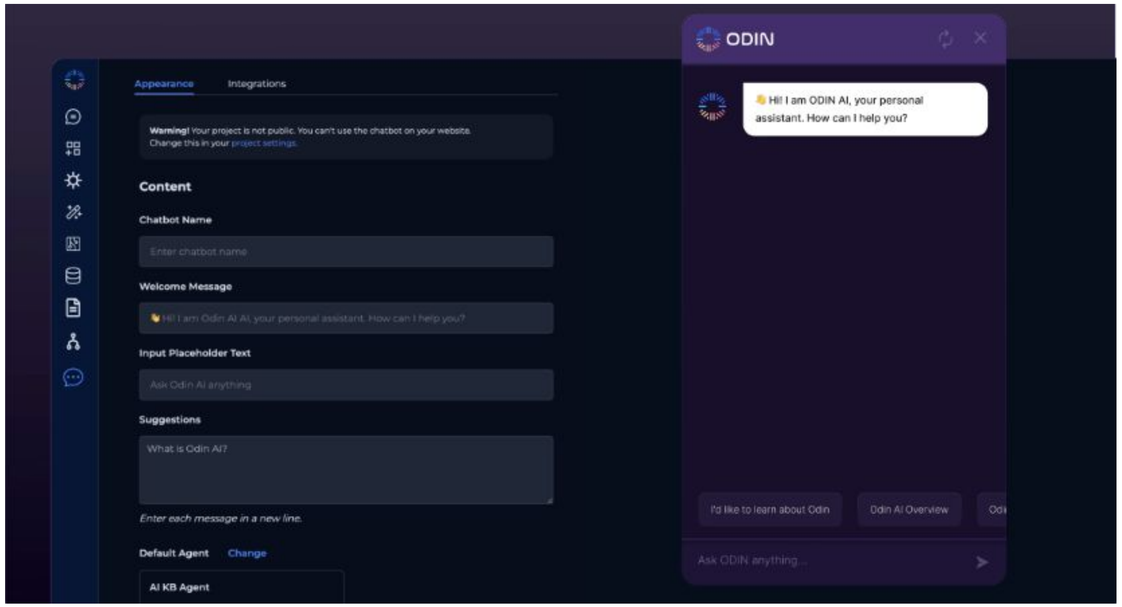
- 40개 이상의 세심하게 큐레이션 된 지식 자산과 70개국에 걸친 전문가 네트워크를 효과적으로 연결하여, 글로벌 인사이트를 실시간으로 활용 가능 -> 컨설턴트의 시간 절약 및 프로젝트 초기 시간 단축
- 회사의 72%가 이 플랫폼에서 활동 및 매달 500,000 이상의 프롬프트 -> 정보 수집 및 종합에 소요되는 시간 절약 및 전세계 여러 동료에게 도움

- https://brunch.co.kr/%40brunchk1wj/219?utm_source=chatgpt.com

프로젝트 소개

1. 추진 배경 및 필요성

1. Odin AI



평균 15분의 검색 시간 -> 검색 시간이 문서당 1-2분으로 단축되어 주당 10시간 절약

- 포괄적인 데이터 세트를 확보하기 위해 6,000개 이상의 기술 문서와 URL을 수집 -> 35%의 응답 일관성이 90%로 올라 균일하고 신뢰할 수 있음
- 40분 평균 쿼리 해결 -> 쿼리 해결 시간이 7분으로 단축
- 오래된 문서로 인한 오류율 20% -> 문서화 오류 15% 감소
- 평균 해결 시간 1.5일 -> 해결 시간이 몇 시간으로 단축

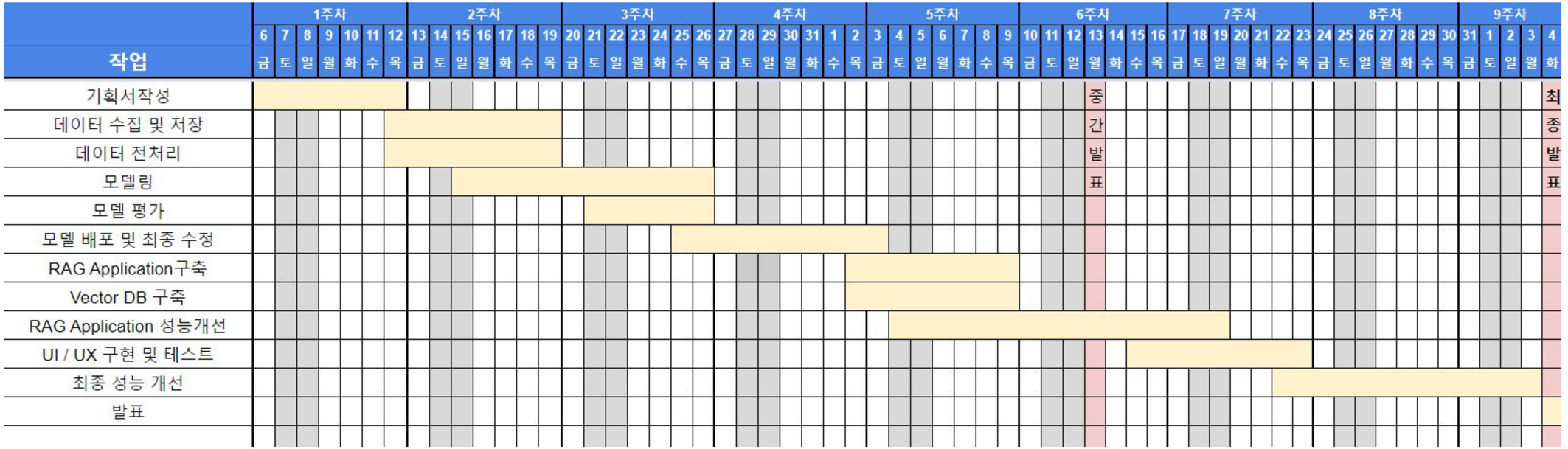
- https://blog.getodin.ai/using-ai-agents-for-technical-document-search-a-detailed-case-study/

참여 인원 및 일정

1. 참여 인원

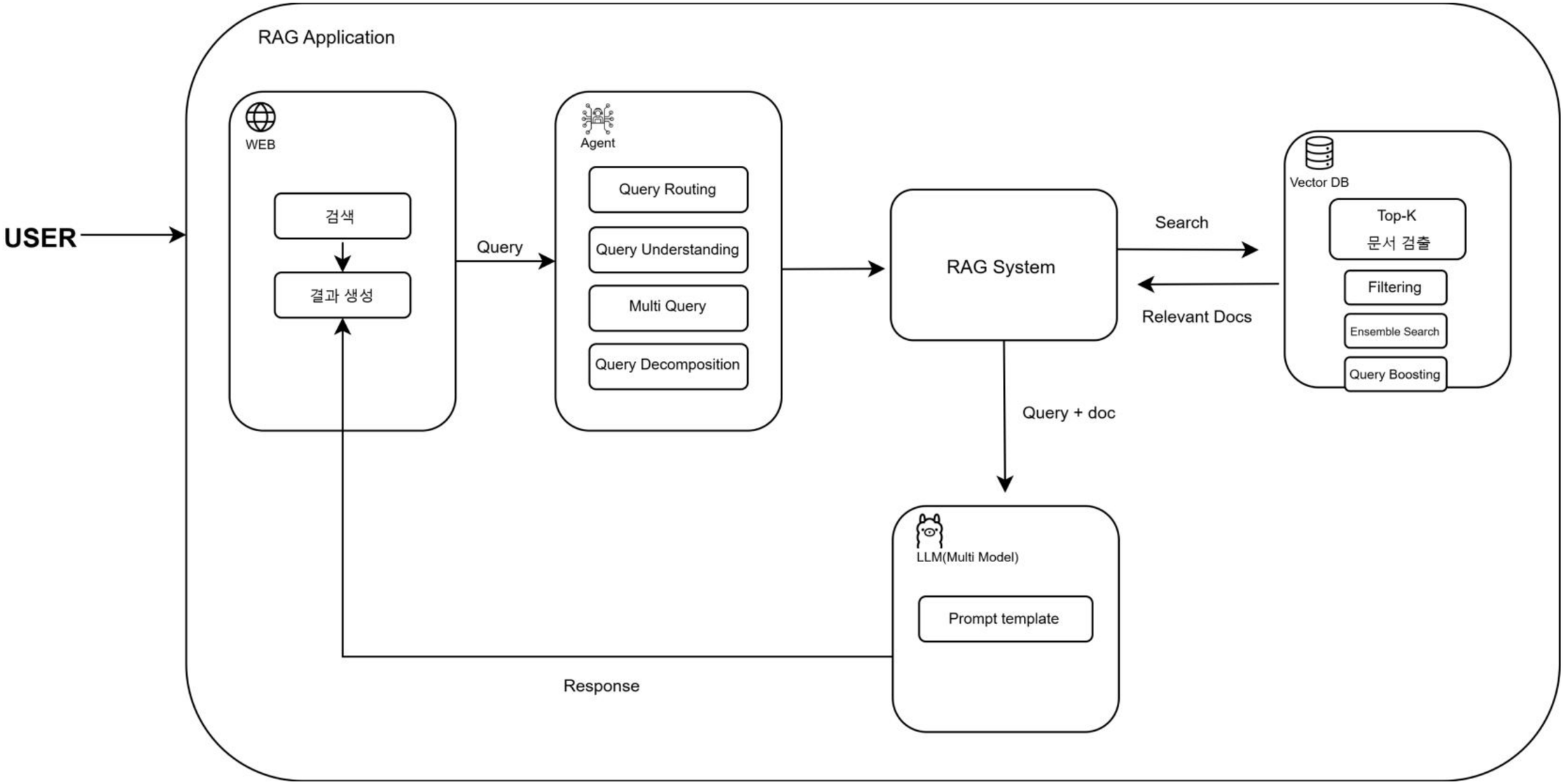
- 안준용 : 프로젝트 리더 | AI Engineer
- 이호재 : AI Engineer | Backend
- 변가원 : Frontend

1. 개발 일정



시스템 구조도

***시스템 아키텍쳐**



상세 수행 내용

1. 데이터

- 파인튜닝 데이터
- Context QA에 적합한 데이터를 Open Source에서 수집.
- 데이터 수집에 제한이 발생할 시, GPT와 같은 대형 언어 모델을 활용하여 Data Augmentation을 진행.

- RAG 데이터
- 타겟 기업 또는 공공기관의 데이터를 수집.
- 수집한 데이터에 대해 Metadata를 생성.
- Data Augmentation을 통해 특정 도메인에 적합한 양질의 Q/A를 생성.

1. **모델링**

- 모델 셀렉션을 위해 LLAMA, Mistral, Phi 등의 모델을 비교 테스트 예정.
- 선택된 베이스라인과 튜닝된 모델을 비교하여 성능을 지속적 개선 예정.

상세 수행 내용

1. **RAG**

- 사용자의 쿼리 의도를 파악하기 위해 Query Translation 기술을 적용.
- Ensemble Search, Filtering, Query Boosting 등의 기술을 활용하여 단독 Semantic Search의 성능 개선예정.
- Search Performance Evaluation: 검색 테스트 데이터셋을 활용해 검색 성능을 평가.
- Generated Dataset Evaluation: 생성 데이터셋을 기반으로 테스트셋 생성 성능을 평가.
- Prompt Engineering: Multi-Model과 Multi-Prompt Template을 사용하여 최적의 성능을 도출.

1. **UI/UX**

- 유저가 쿼리를 입력할수있는 검색창.
- 히스토리, 문서 리스트를 볼수있는 사이드바 메뉴.
- 이전 대화 히스토리를 유지하여 문맥에 반영하는 멀티턴 기능.
- 사용자가 입력한 내용과 답변의 내용을 대화형식으로 표시하는 대화창.
- 검색시에 지연 시간동안 사용자가 시각적으로 지루하지 않게 로딩바로 상태 표시.
- 자세한검색 (예: 카테고리별 검색)을 위한 고급검색 기능 추가.
- 답변 출력중에는 검색을 블락 할 수 있는 기능 추가.

상세 수행 내용

1. **회원관리**

    - 구글 연동을 통한 회원가입 기능
    - 구글계정으로 쉽게 로그인 할 수 있는 기능
    - 로그아웃 버튼을 눌러 로그아웃 하는 기능

평가 방법

1. 튜닝된 모델 평가

- 테스트 데이터로 테스트하여 반환된 문서와 정답의 Context Recall, Context Precision을 구하여 비교.
- 루지 스코어(Rouge Score)와 같은 정량적 평가 지표를 참고하며, 결과를 직접 확인하여 정성적으로 평가.

2. Retriever 평가

- 테스트 데이터로 테스트하여 반환된 문서와 정답의 Context Recall, Context Precision을 구하여 비교.

3. 생성 평가

- RAGAS를 이용하여 Context Recall, Context Precision, Answer Relevancy, Faithfulness와 같은 지표를 참고하여 평가 예정.
- 평가 데이터는 튜닝데이터에서 1000천개 가량 빼서 실행.
- 사람이 직접평가하는 정성적 평가도 100건 가량 실시.