[LG U+ Why not SW CAMP 2기]

상담 지원 chatbot

모델 정의서

팀명 : 용&우 컴퍼니

팀장 : 이용우

팀원 : 김찬유 남궁진 성채린 박지민 박소엽

목차

I. 모델 개요 및 선정 이유

1.1 KMS Embedding 모델

모델 개요

사용 모델	**Upstage embedding 모델 (solar-embedding-1-large)**
사용목적	**상담 지식정보(KMS) 데이터를 단어 의미적 유사성을 기반으로 벡터화하여 적재하고, 이후 질의가 들어올 때 유사성이 높은 문서들을 반환하기 위해 사용**
활용 방식	**VectorStore 기반으로 mmr 알고리즘을 사용해 Retriever로 문서 검색**
성능과 효율성	**높은 정확도로 텍스트 유사도 계산 및 검색 속도를 개선**
경량화	**대규모 데이터셋을 처리하면서도 메모리 효율성을 유지**
유연성	**다양한 데이터 유형(KMS 텍스트, 문서)과 다국어 지원**

모델 선정 이유

비교 항목	**solar-embedding-1-large**	**kobert**	**snowflake-arctic-embed2**
임베딩 차원수	**4096**	**768**	**1024**
검색 정확도	**Top3 Accuracy: 91.42%**	**85%**	
적합성	**RAG, 다국어 검색, 임베딩에 적합**	**한국어 데이터 최적화**	**RAG, 다국어 검색, 임베딩에 적합**

결론

지식정보(KMS) embedding 모델을 위한 최종 모델로 **solar-embedding-1-large** (upstage) 모델 선정

- 테스트 환경 상 비용 측면의 이점.

- Retriever를 통한 유사도 문서 검색 속도 및 정확성 우수.

1.2 LLM

모델 개요

사용 모델	**Gemini-1.5-flash**
사용목적	**Retriever를 통해 검색된 관련 문서들을 구어체로 변환해 상담 시 활용하기 편하게 하기 위해 사용**
활용 방식	**Retriever방식으로 가져온 문서를 기반으로, 상담사의 요구사항에 맞춘 프롬프트를 통해 변환 챗봇 UI에 전달**
성능과 효율성	**약 6~7초 가량 소요됨.**

모델 선정 이유

비교 항목	**gemini**	**OpenAI**
언어 범위	**다국어 지원**	**다국어 지원**
비용	**무료**	**5$/ 100만 토큰**
적합성	**대규모 구어체 변환 및 다국어 작업에 적합**	**대규모 구어체 변환 및 다국어 작업에 적합**

결론

처리 속도는 OpenAI API를 사용하는 것이 좀 더 빠르지만, 가격 상의 문제로 인해 gemini LLM 모델을 사용함.

1.3 Vector DB

DB 개요

사용 DB	**pinecone**
사용 목적	**지식정보(KMS) 문서를 구어체로 변환**
활용 방식	**Retriever 방식으로 가져온 문서를 구어체로 변환하여 챗봇 UI에 전달**
성능과 효율성	**SaaS 기반이라 환경에 구애 받지 않고 구축할 수 있고, 배포에 용이함.**

DB 선정 이유

비교 항목	**pinecone**	**Faiss**
비용	**Cloud 기반 비용 발생**	**오픈 소스**
특징	**- API key를 통해 간단히 사용** **- 구축 편의성** **- 서버리스, Pod 옵션 제공** **- 다국어 검색 지원**	**- 고성능 유사도 검색 지원** **- 클러스터링 라이브러리 지원** **- GPU 가속 알고리즘 지원** **- 대규모 Dataset 처리 최적화**
성능	**0.8초 (문서 309개)**	**2.1초 (문서 309개)**
비교 항목	**Chroma**	**PGVector**
비용	**오픈 소스**	**PostgreSQL 호스팅 비용**
특징	**- 임베딩, 벡터 검색, 문서 저장 통합 제공** **- 멀티모달 검색 지원**	**- PostgreSQL 기능 사용 가능** **- SQL과 통합된 벡터검색기능**
성능	**1.7초 (문서 309개)**	**29.3초 (문서 309개)**

결론

촉박한 프로젝트 일정 상, 구축이 편리하고, 제일 문서 검색 속도가 빠른 Pinecone Vector DB를 사용하기로 결정함.

II. Data Set

2.1 Data Set 구성

Data Set 명	쿠팡 고객센터 자주 묻는 질문 Q&A
개요	쿠팡 고객센터의 자주 묻는 질문의 취소/교환/반품, 배송문의, 주문/결제, 환불 카테고리의 Q&A쌍 데이터를 가져옴.
출처	https://mc.coupang.com/ssr/desktop/contact/faq?categoryCode
사용목적	벡터DB Embedding 이후 Retriever로 관련 문서 검색
데이터 구조	CSV
수집 방식	크롤링

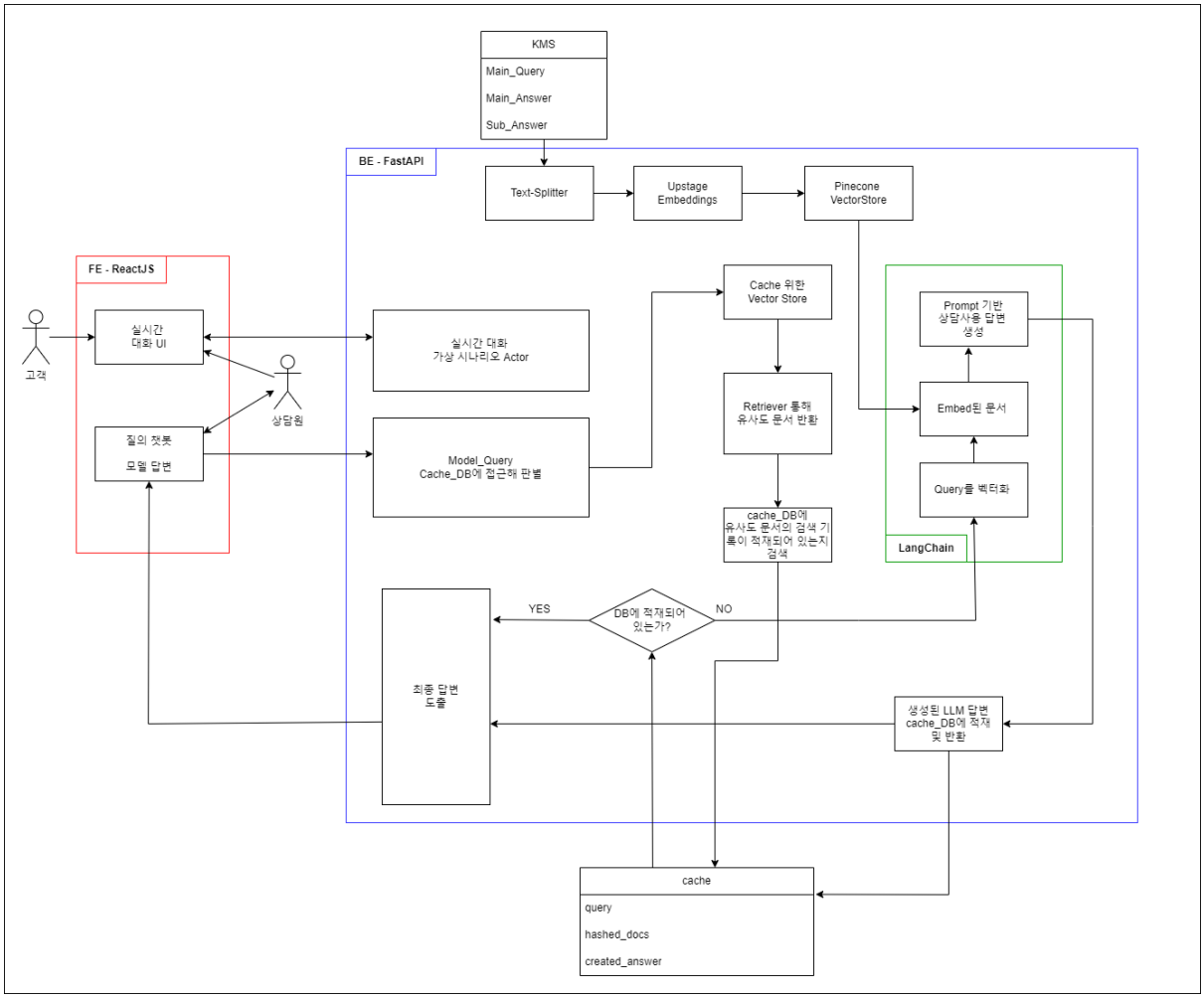
2.2 데이터 전처리

- 한 질문에 있는 Main Question과 파생되는 Sub Question을 분리.
- 취소/교환/반품, 배송 문의, 주문/결제, 환불, 서브 카테고리 5개로 나눠서 데이터 적재
- 메인 카테고리 271개, 서브 카테고리 38개

III. 모델 및 모델구조

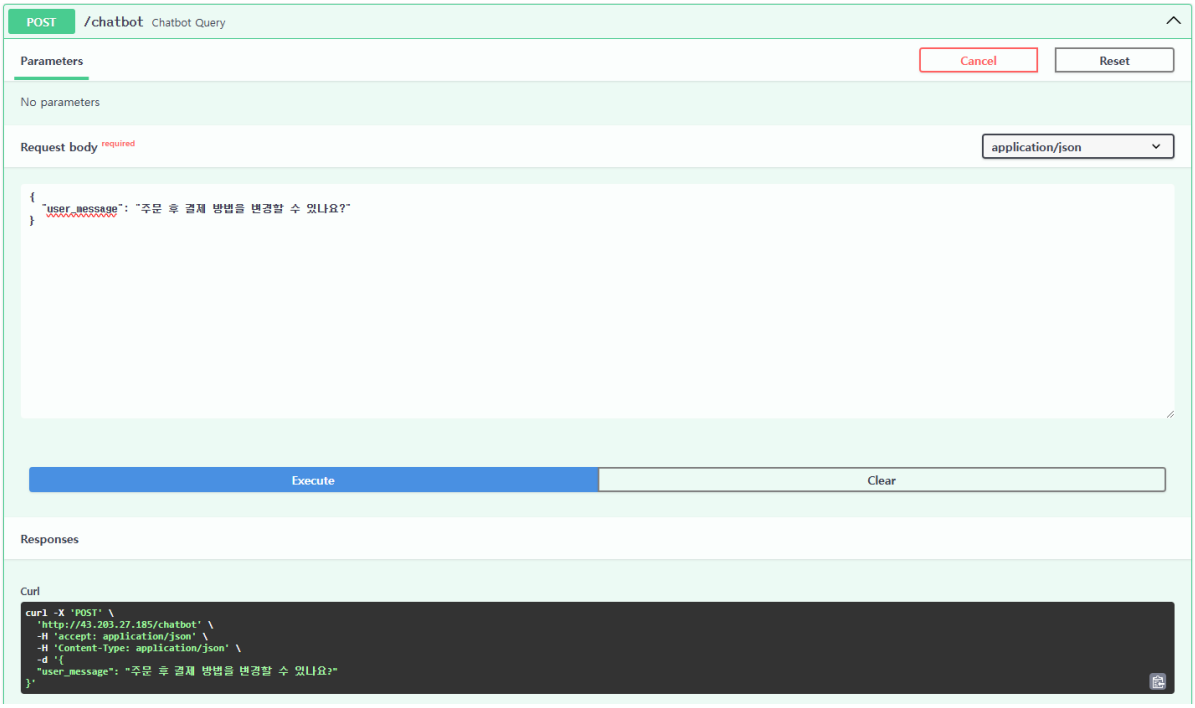
3.1 Work flow 및 모델

전체 flow



3.2 모델

3.2.1 Model에 Query 질의



3.2.2 Model 응답 결과