FAQ 챗봇 MVP 개발

기술 명세서

안성준 (전략사업부)

2025/05/20

# 1. 배경 및 목적

퍼플아카데미의 전문적인 교육 서비스에 대한 정확한 정보를 제공하는 챗봇 개발

# 2. 필수 구현 사항

* 퍼플아카데미 학습 시스템에 대한 정확한 정보 제공

# 3. 고려 사항

* 비슷한 단어 구분이 가능해야함
* 이용자들이 반드시 학습 내용에 대한 질문만 하는 것이 아니므로 이에 대한 대응이 준비되어있어야함
  + 환불, 결제, 이용 문의, 클레임, 학습 관련 문의
  + 장난 질문

# 4. 구현 방식

* RAG 시스템 : LLM 응답이 외부 FAQ DB를 참조하도록 하여 오정보 생성(Hallucination)을 최소화
* 가능한 간단한 MVP 아키텍처 구현
  + 텍스트 데이터 등 비정형 데이터에 최적화되어 있으며 별도 검색 엔진을 붙이지 않아도 되는 간단한 아키텍처가 가능해야함.
  + 이를 위해서 MongoDB보다는 Elasticsearch와 Qdrant 를 고려. 정확도가 중요한 목표를 고려했을 때 처리 속도 손실이 있더라도 Elasticsearch가 더 낫다고 판단
* Elasticsearch + 메타 데이터 레이어 :
  + Elasticsearch 만으로는 사실관계가 중요한 복잡한 정보에 대한 정교한 처리가 어려울 수 있음.
  + 특히, 조건문이 여러 개 중첩되는 정보처리가 필요한 질문에도 대응할 수 있어야 함
  + 지식 베이스 구축 단계에서 메타 데이터 태깅 + 레이어 연동을 통해 하이브리드 서치를 구현할 필요
* 계층화된 필터 :
  + Elasticsearch는 정확하지만 응답속도가 빠르지 않아 1~5밀리초 단위의 응답은 어려울 수 있음.
  + LLM API의 생성을 통과하므로 쿼리와 생성 시간이 합쳐지면 체감 가능한 성능 저하가 있을 수 있음.
  + 반복되는 질문에 대한 API 코스트 절감도 고려해야함.
  + Redis 등 캐싱 전략을 활용하여 즉각적인 처리가 가능한 질문과 그렇지 않은 질문에 대한 폴백 처리를 통해 더 깊은 계층에 있는 벡터화 DB까지 질문을 전달하는 로직을 분리할 필요성

# 5. MVP 아키텍처

Redis + Elasticsearch + LLM

5.