- 라이선스 비용, 컴퓨팅 자원 고려: 오픈소스 여부, 상용 서비스 비용, 클라우드 비용이나 하드웨어 구축 비용 등을 검토하여 제약 조건에 맞게 선택합니다.
- 성능 검증: 후보 모델을 실제 데이터로 벤치마크 테스트하여 유사도 정확도, 응답 시간, 메모리 사용량 등을 평가합니다

Q. 벡터 및 LLM 모델은 어디서 확인하나요?

A. 관리 도구에서 벡터 모델, LLM 연동 관리 기능을 통해 설정 및 변경이 가능하며, GUI 기반 관리 대시보드도 제공합니다.

3. 성능 및 유지보수 이슈

#PoC #유지보수 #기술역량

3.1 정확도 및 성능관리

#문서버전관리 #검색정확도 #속도최적화 #ReRanking

Q. 과거 문서와 최신 문서 간 정확도 불일치 문제는 해결 가능한가?

A. 문서 수집 및 버전관리 기능과 최신 검색 순위 보정 기술(HyDE, ReRanking)을 통해 해결 가능합니다.
최신 수정 일자와 버전기준으로 최신 문서 기반으로 답변을 구성 하게 구현 할 수 있습니다. 문서의 권한 및 수정 등은 수시로 일어나서 DB에 실시간으로 관리되는것이 중요하며 이부분에 대한 권한 및 검색 하는 방식이 매우 중요합니다.

Q. 대규모 데이터 응답 속도 문제는 해결 가능한가?

A. HNSW 기반 인덱싱, 유사도 기반 탐색 등으로 빠른 응답이 가능하며, 벡터DB 최적화가 적용됩니다.
이부분은 LLM의 컨텍스트 지원 사이즈와 연관이 있으며 AI Agent 기반으로 DB에서 많은 데이터를 끌어 와서라도 LLM 이 이를 처리 할 수 있는 범위가 넘어간다면 성능 상에 문제가 발생 할 수 있습니다.

그래서 다량의 데이터 처리에 대한 아키텍처 구성이 중요하며 특정 단위로 데이터를 잘라서 처리하거나 외부에서 처리한 요약 된정보를 LLM 기반으로 활용하는게 좋습니다