**프로젝트명:** LLM 기반 사내 문서 검색 시스템 기획서

**버전:** v0.9

**작성일:** 2024년 12월 18일

**작성자:** 김문수

**1. 개요**

**1.1 프로젝트 배경**

기업 내 생성·축적되는 문서(매뉴얼, 연구 보고서, 회의록, 기술 자료)는 방대한 양에 달하며, 이러한 정보 자산을 효율적으로 탐색하고 활용하는 것은 조직의 경쟁력 강화에 필수적이다. 기존의 키워드 기반 검색은 정확한 키워드 매칭을 필요로 하고, 문서의 맥락(Context) 파악이 어려워 신뢰도 높은 결과를 도출하기 어렵다. 이에 대규모 언어 모델(LLM, Large Language Model)을 활용한 문맥 기반 검색·질의응답 시스템을 도입함으로써, 사용자의 질의 의도를 파악하고 관련 문서를 정교하게 추천하며 요약·정리하는 지능형 검색 환경을 구축하고자 한다.

**1.2 프로젝트 목적**

* 사내 문서에 대한 효율적이고 맥락적인 검색 및 질의응답 제공
* 키워드 중심에서 벗어난 의미 기반 검색(semantic search) 구현
* LLM을 통한 빠르고 정확한 정보 접속 시간 단축 및 생산성 향상
* LangChain, Agent, LangGraph 등 기술 활용을 통한 외부 자원 연동
* 사내 보안 규정을 만족시키는 안전하고 확장 가능한 시스템 구축

**2. 시장 조사**

**2.1 글로벌 시장 동향**

최근 OpenAI(챗GPT) 및 Google(Bard), Anthropic(Claude), Meta(Llama 2) 등 대규모 언어 모델 기술 진전으로 자연어 이해 능력이 현격히 개선되었다. 이로 인해 기업들은 내부 정보 관리에도 LLM 기반 기술을 적극 도입하기 시작하였다.

* **엔터프라이즈 검색 시장 성장:** 기존 엔터프라이즈 서치(Enterprise Search) 솔루션에서 LLM을 결합해 검색 결과 품질 향상을 추구하는 움직임이 활발하다. Google Cloud, Microsoft Azure OpenAI, AWS Bedrock 등 클라우드 서비스 사업자들도 LLM 기반 검색 기능 지원을 확대 중이다.
* **컨텍스트 기반 Q&A 수요 증가:** 단순 문서 검색에서 나아가, 문서 내용을 요약하고 질의응답하는 능력에 대한 기업 니즈가 급증하고 있다.

**2.2 경쟁 솔루션 사례**

1. **Microsoft Viva Topics:** Microsoft 365 내 문서, 지식 관리에 AI를 도입한 솔루션으로, 조직 내 지식 흐름 개선을 목표로 한다.
2. **Elasticsearch + LLM 결합 솔루션:** 오픈소스 검색엔진 Elasticsearch에 LLM 기반 플러그인(Elastic AI Assistant)을 연동하여 의미 기반 검색 및 Q&A 기능을 제공한다.
3. **OpenAI + Vector DB 연동 패턴:** Pinecone, Weaviate, Chroma 등 벡터 DB를 OpenAI API와 결합해 문서 임베딩과 의미 기반 검색을 구현하는 베스트 프랙티스가 확립되고 있다.

**2.3 기술 성숙도**

* LLM 모델의 API 제공은 안정화 단계 진입, Prompt Engineering을 통한 커스터마이징 가능
* 벡터 DB 기술(Pinecone, Weaviate 등)의 발전으로 대규모 임베딩 관리 용이
* LangChain, Agent, LangGraph 등 프레임워크를 통한 워크플로우 자동화 및 외부 자원 연동은 초창기 단계지만 빠른 속도로 성숙 중

**2.4 시장 타당성 분석**

* **수요 적합성:** 사내 지식 활용 극대화를 원하는 기업 수요에 부합
* **경쟁 우위 가능성:** LLM + 벡터 DB + 외부 자원 연동을 통한 고도화된 검색 경험은 경쟁사 대비 차별화 포인트
* **비용/효율성:** 클라우드 기반 LLM API를 활용하면 초기 인프라 투자 부담 완화 가능. 단, API 사용량 증가 시 비용 관리 필요.

**3. 구축 타당성 검토**

**3.1 비즈니스 타당성**

* **비즈니스 가치:**
  + 사내 정보 접근 시간 단축 및 업무 생산성 향상
  + 의사결정 속도 개선
  + 신규 입사자 온보딩 기간 단축(지식 접근성 강화)
* **ROI 측면:**
  + 단순한 시간 절약 효과에서 나아가 정확한 정보 활용으로 오류 감소, 재작업 최소화
  + 장기적으로 지식 관리 인프라 고도화를 통해 경쟁력 확보

**3.2 기술 타당성**

* **구현 가능성:**
  + 사내 문서 임베딩 후 벡터 DB 인덱싱 기술 성숙
  + LLM API 활용 및 LangChain 연동 사례 다수 존재
  + Agent/LangGraph 통한 외부 자원 연동 시나리오 충분히 구현 가능
* **인프라 준비성:**
  + 클라우드 기반 서비스 이용 시 초기 구축 용이
  + 기업 내부망 문서 인덱싱 시 보안 요구사항 충족 가능

**3.3 리스크 분석**

* **보안 이슈:**
  + 사내 문서(민감 정보) 외부 전송 시 안전성 문제(LLM API 호출 시 데이터 마스킹, 프라이버시 보호 필요)
  + 인증/접근제어 강화 필요(JWT/OAuth2, RBAC)
* **비용 관리:**
  + LLM API 호출 빈도 증가 시 비용 상승 가능, 캐싱 전략 및 사내 로컬 모델 배포 검토 필요
* **품질 관리:**
  + LLM 응답 품질 편차, Hallucination 문제(LLM이 확신 없이 틀린 정보 생성)
  + 내부 문서 기반 Context 제공 및 시스템 설계를 통한 품질 개선 필요

**4. 시스템 개요 및 요구사항 요약**

**4.1 시스템 개요**

* 내부 문서(여러 포맷) 벡터화 후 벡터 DB에 저장
* 사용자 질의 시 벡터 DB 검색 결과를 LLM에 Context로 제공, 의미 기반 Q&A 수행
* Agent, LangChain, LangGraph를 통해 외부 API(사내 ERP, 외부 뉴스/지식베이스) 연계 가능
* 토큰 기반 인증/권한 관리, 사용자 관리(등록, 조회, 수정, 삭제), 동적 메뉴 구성 및 보안 통제

**4.2 주요 요구사항**

* **인증/권한:** 토큰 기반 로그인/로그아웃, RBAC(Role-Based Access Control)
* **문서 관리:** 문서 업로드, 수정, 삭제, 벡터 DB 인덱싱
* **검색/QA:** 자연어 질의를 통한 문맥 기반 문서 검색 및 요약 응답
* **외부 자원 연동:** Agent 활용해 외부 API/DB 연계, LangGraph로 복합 워크플로우 처리
* **보안/감사:** HTTPS, CSRF/CORS 방어, 감사 로그, 에러 로그 및 모니터링

**5. 기대 효과 및 성공 지표**

**5.1 기대 효과**

* **업무 효율성 향상:** 원하는 정보를 수초 내에 확보 가능
* **지식 활용 극대화:** 사내 누적 지식을 LLM 기반으로 재해석, 고품질 정보 전달
* **의사결정 지원:** 실시간 지식 접근으로 의사결정 프로세스 단축

**5.2 성공 지표(KPI)**

* **검색 응답시간:** 평균 1~2초 이내
* **사용자 만족도 설문:** 80% 이상 '개선' 응답
* **LLM 응답 정확도 지표:** 내부 평가 문항 기준 90% 이상 정확성 달성
* **시스템 가용성:** 월 가용률 99.9% 이상

**6. 결론 및 제언**

LLM 기반 사내 문서 검색 시스템은 현재 시장에서 주목받는 AI 기술 트렌드와 기업의 지식 관리 니즈를 균형 있게 반영하는 솔루션이다. 기술 성숙도 및 구현 사례가 풍부해졌으며, 벡터 DB와 LangChain, Agent, LangGraph 등 생태계를 활용하면 의미 기반 검색 및 외부 자원 연동이 현실적 수준으로 실현 가능하다. 다만, 보안과 비용, LLM 품질 관리 측면의 리스크 대응 전략을 마련해야 한다.

향후 단계별 PoC(Proof of Concept) 진행을 통해 모델 선택, 벡터 DB 성능 검증, 보안 정책 수립 및 사용자 피드백 수렴을 권장한다. 이를 토대로 단계적 서비스 확장 및 최적화를 통해 사내 지식 관리 혁신을 달성할 수 있을 것이다.