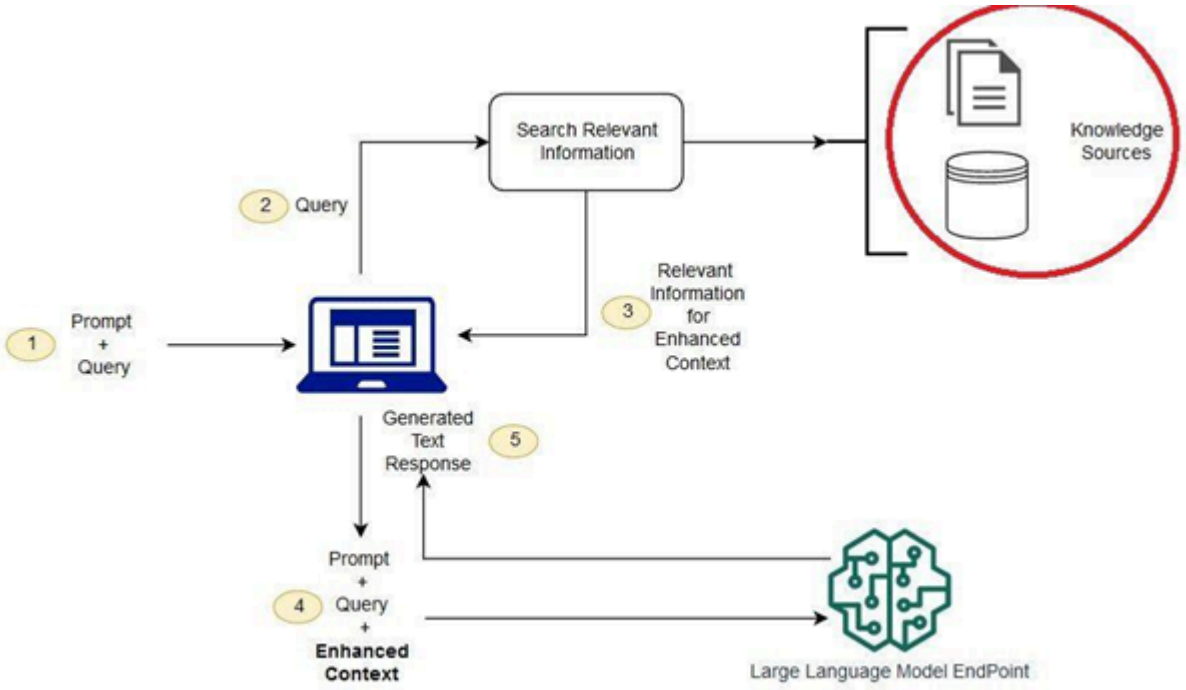
이때 벡터 임베딩 모델(예: SBERT 등)과 벡터 DB(예: Pinecone, Milvus 등), LLM(API 형태) 등의 기술이 사용됩니다.

검색 단계에서는 주로 코사인 유사도 등 유사도 함수를 통해 관련 문서를 찾으며, 필요 시 추가 도구(웹 검색, 계산기 등)를 연동할 수도 있습니다.



Q. Local LLM을 내부적으로 사용 중이며, RAG에도 로컬 LLM 연동이 가능한가?

A. LLM을 파싱 및 설명 보조 용도로 활용 시 정확도 향상의 효과가 있습니다. .토큰 비용 및 GPU 머신 리소스 문제가 있지만, 최근 고객들도 멀티모달 LLM 연동을 시도 중이며. 고객 요구 시 로컬 LLM(Llama 등) 연동도 가능합니다. 프로젝트 범위 내에서 지원 가능 및 지원 의사 또한 있습니다.그리고 해당 케이스에 대한 구축 경험도 있습니다

Q. RAG 관련 성능 평가를 받은 지표가 있는가?

A. 정확률(Precision), 재현율(Recall), 사용자 피드백 기반 정성 평가 등으로 관리하며, 하이브리드 검색 + ReRanking 등을 적용합니다.