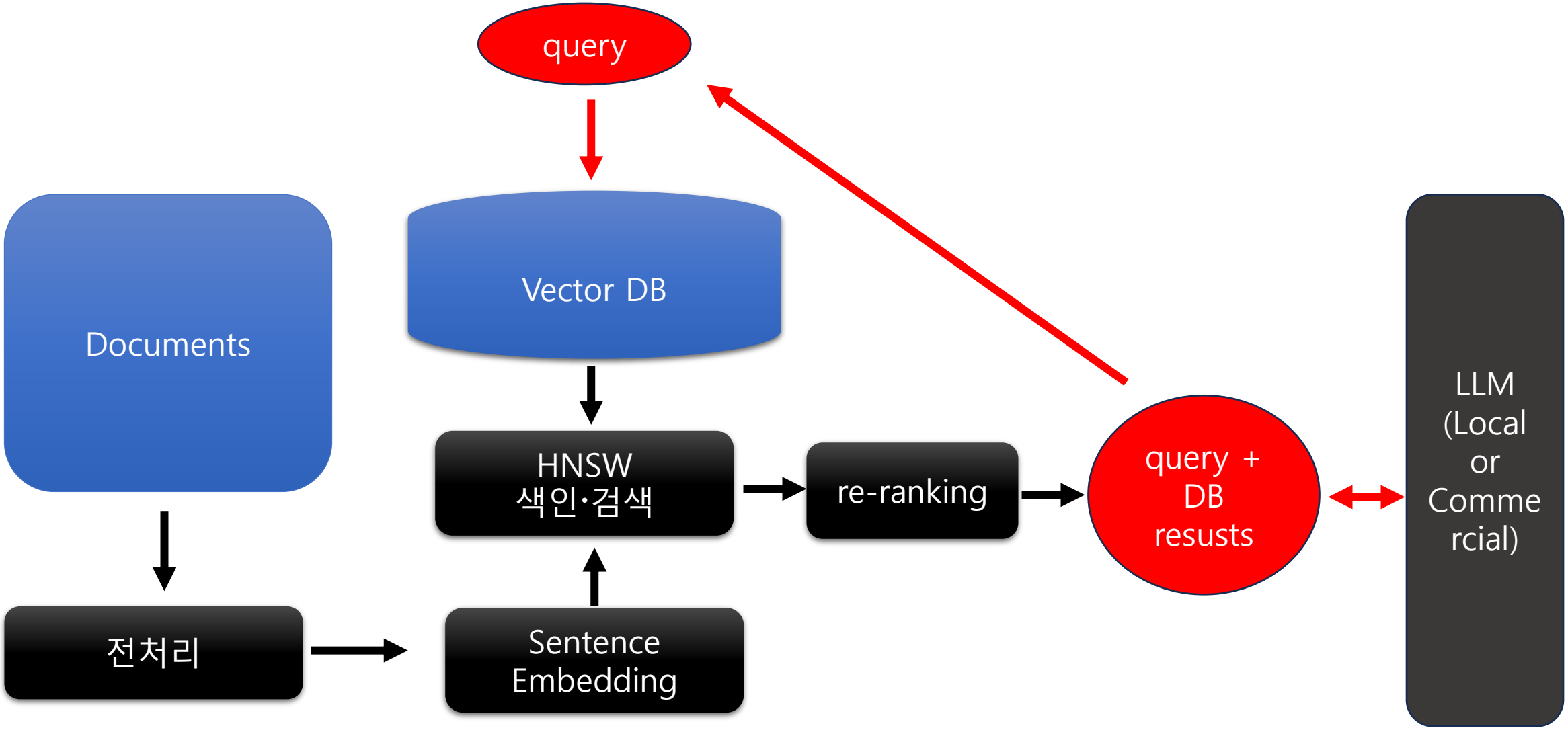
RAG System

Langchain, VectorDB, LLM

RAG 시스템 이란?

- Agent가 사용자 질의 받아 먼저 DB 자료를 검색하고 그 결과를 반영하여 다시 Query를 만들어 LLM에 질의하여 결과를 얻어내는 시스템
  - 멀티 모달(텍스트, 이미지, 오디오, 동영상 등) 검색 지원
- DB는 자연어 검색이 가능한 Vector DB를 사용하여 기업 내부의 문서를 자연어 검색 지원
  - 전통적인 키워드 검색 및 메타 데이터 필터링 지원
- LLM은 기업 내부 자료 보안을 위하여 로컬에 기업용 LLM 설치
  - 대기업의 경우 자체 개발한 LLM을 사용
  - 기업의 규모가 작은 경우 Open Source 모델을 사용

RAG 시스템 구조



RAG 시스템 핵심 요소 기술

① Document 전처리

- 전통적인 SQL DB에 문서 메타 정보 기록
- 문서를 읽고 색인하기 쉽도록 분할하는 작업

② 키워드 검색(Sparse Retrieval) 지원

- 전통적인 Keyword 기반 검색 시스템 지원 (BM25)
- Elastic Search 시스템

③ Sentence Embedding – Bi-Encoder (두 개의 네트워크를 학습시킨다는 의미)

- S-BERT는 지도학습
- 2020년 이후 대용량 embedding 학습 – Contrastive Learning

④ ANN(Approximate Nearest Neighbor) 검색

- Hierarchical Navigable Small World 기반 ANN이 대세

⑤ Re-ranking

- Embedding 검색에서 나온 결과를 다시 순위를 매김 – Cross Encoder (원래 질의, 검색 결과) → 유사도
- Cross Encoder는 Bi-Encoder 보다는 가벼운 네트워크로 학습

⑥ LLM(Large Language Model)

Bi-Encoder vs. Cross Encoder

구분	Bi-Encoder(1차 검색용)	Cross-Encoder(2차 Re-Ranking용)
입력 방식	쿼리와 문서를 따로 임베딩(독립적)	쿼리와 문서를 함께 입력(하나의 쌍)
계산 방식	쿼리 벡터와 문서 벡터를 내적(코사인 유사도)	[쿼리+문서] 쌍이 모델을 통과하여 하나의 점수 산출
핵심 장점	매우 빠름(미리 문서 벡터 계산 가능)	매우 정확함(쿼리와 문서 단어 간 상호작용을 모두 고려)
핵심 단점	정확도 상대적 낮음	쿼리마다 새로 계산 필요
비유	경기장에서 사진으로 사람 찾기: 미리 찍어둔 전광판 사진(벡터)과 내 사진(쿼리 벡터)을 빠르게 비교 (1차 검색)	100명의 후보와 직접 면접: 후보(문서)와 내가(쿼리) 함께 방에 들어가 대화(Cross-Encoder)하며 점수 매김 (Re-ranking)

Hybrid Retrieval 시스템

- 전통적인 Keyword 검색 + Embedding 방식을 결합한 검색 시스템
- Sparse Retrieval 장단점
  - 비용 절약
  - 고유 명사, 주요 키워드 탁월
  - 의미는 같으나 다르게 표현된 것 찾아내지 못함
- Dense Retrieval 장단점
  - 의미·문맥 검색 가능
  - 비용 높음
  - 고유명사 정확도 낮을 수 있음
- Hybrid Retrieval
  - Sparse Retrieval 점수와 Dense Retrieval 점수를 합쳐서 최종 순위 결정
  - 정확도 매우 높아짐
  - Hybrid Retrieval 결과를 LLM에게 보내서 최종적인 답변 반환하는 강력한 RAG 시스템