Q. 파인튜닝(모델 재학습)과 기본 모델 활용 방식의 차이는 무엇인가?

1. **모델 신뢰성과 변경 추적의 어려움**
   파인튜닝을 통해 기본 모델에 추가 학습을 진행할 경우, 어떤 데이터가 어떤 방식으로 모델의 응답에 영향을 미쳤는지를 명확히 추적하기 어렵습니다. 이로 인해 응답 품질 변화의 원인을 파악하거나 문제 발생 시 수정 포인트를 특정하는 데 어려움이 발생할 수 있습니다.

2. **운영 인프라 요구사항 증가 및 서비스 중단 리스크**
   파인튜닝에는 전용 GPU 장비와 고성능 인프라가 필요하며, 학습 완료 후 모델을 적용하는 과정에서 일시적인 서비스 중단(Downtime)이 요구됩니다. 이는 실시간 응답이 중요한 서비스 환경에서 운영 리스크로 작용합니다.

3. **실시간 학습 불가 및 고비용 구조**
   파인튜닝은 정제된 데이터를 바탕으로 수 시간 이상의 학습 시간이 필요하고, 실시간으로 유입되는 데이터를 즉시 반영하는 구조는 현실적으로 어렵습니다. 또한 장비, 시간, 인력 등 운영 비용이 크게 증가하게 됩니다.

2. RAG 관련 질문

2.1 RAG 구성 및 운영

#RAG #ReRanking #하이브리드검색 #Text2SQL #멀티에이전트

Q. RAG란 무엇인가? 구성과 기술 요소는?

RAG(Retrieval-Augmented Generation)는 외부 지식 기반 검색을 LLM 생성과 결합하는 기법입니다.

일반적인 RAG 구성은

(1) 데이터 색인: 문서를 임베딩해 벡터 DB에 저장,

(2) 검색 단계: 사용자 쿼리를 임베딩하여 벡터 DB에서 유사한 문서를 검색,

(3) 생성 단계: 검색된 문서를 LLM에 컨텍스트로 제공하여 답변을 생성하는 구조로 이루어집니다.