RAG 테스트 문서 (가짜 데이터) - 러너(Reranker) 개요 (Fictional)

Generated: 2025-12-25 07:05:05

목적: RAG 파이프라인에서 '러너(Reranker)' 개념을 테스트하기 위한 가짜 문서입니다.
주의: 아래 수치/모델명/정책은 모두 예시이며 실제 서비스/논문/제품 사실이 아닙니다.

1) 러너(Reranker)란?
- 정의: 1차 검색(Retriever)이 뽑아온 후보 문서(top_k 후보)를 다시 평가하여 '정확한 순서'로 재정렬하는 단계.
- 위치: Embed/Search 이후, LLM 생성 전에 배치되는 것이 일반적.
- 효과: 관련도(precision) 상승, 노이즈 청크 감소, 근거 기반 답변 안정화.

2) 가짜 러너 모델/정책(Fictional)
- 가짜 러너 모델명: R-Runner-Small v1 (예시)
- 점수 스케일: 0.0 ~ 1.0 (1.0에 가까울수록 질문과 더 관련)
- 가짜 운영 정책:
  * Retriever에서 top_k=12 후보를 뽑는다.
  * Reranker가 상위 top_n=4만 남긴다.
  * 최종 컨텍스트는 top_n 청크만 사용한다.

3) 가짜 성능 지표(Fictional Metrics)
- baseline(러너 없음) 정답률: 62% (예시)
- 러너 적용 후 정답률: 74% (예시)
- 개선 요약: '불필요한 문서'가 컨텍스트에 들어오는 비율이 줄어 LLM 회피/환각이 감소했다고 가정(예시).

4) 실무 팁(일반론)
- 러너는 '정확도'를 올리지만 '지연 시간(latency)'을 늘릴 수 있다.
- 문서가 섞여 있는 경우(source 혼합)에는, 러너보다 먼저 '필터링(where/source)'가 필요하다.

RAG 테스트 문서 - 러너 샘플 QA (정답 명시)

Generated: 2025-12-25 07:05:05

A) 샘플 QA (문서에 정답이 '명시'되도록 구성)

Q1) 이 문서에서 말하는 러너(Reranker)의 정의는 무엇인가?
A1) 1차 검색이 뽑아온 후보 문서를 다시 평가해 '재정렬'하는 단계.

Q2) 가짜 러너 모델명은 무엇인가?
A2) R-Runner-Small v1

Q3) 가짜 운영 정책에서 Retriever top_k는 얼마인가?
A3) 12

Q4) 가짜 운영 정책에서 Reranker top_n은 얼마인가?
A4) 4

B) 테스트 팁
- 질문은 문서의 키워드(모델명/숫자/top_k/top_n)를 그대로 묻는 형태가 가장 안정적입니다.
- RAG 컬렉션에 여러 문서가 섞여 있으면, source 필터링(또는 컬렉션 분리)이 먼저입니다.