Q. LLM 품질 및 정확도는 어떤 방식으로 측정하는지?

A. LLM 품질 및 정확도는 100개이상의 질문과 정답 데이터셋을 활용하여 정량적 지표와 정성적 평가를 병행하여 다각도로 측정합니다. 주요 방식은 다음과 같습니다.

1. 의미적 유사도(Cosine Similarity)를 통한 정량 평가

모델의 응답과 원 쿼리(질문) 간 임베딩을 기반으로 Cosine Similarity를 계산합니다.

이 값은 응답이 쿼리의 의도를 얼마나 잘 반영하고 있는지를 수치로 나타냅니다.

일반적으로 유사도가 높을수록 의미적 일치도가 높다고 해석할 수 있습니다.

2. 도메인 전문가 또는 기준 데이터 기반의 정성 평가

사람이 직접 판단하여 모델 응답이 정확하고 유용한지를 Success Flag 형태로 기록합니다.

이진 값(성공/실패)으로 표현되며, 모델 응답이 실제 사용 목적에 적합했는지를 나타냅니다.

3. 정답(Ground Truth)과의 비교

각 쿼리에는 대응되는 근거자료(예: 실제 논문 제목, 학술지 정보 등)가 존재합니다.

모델이 생성한 응답 내에 이러한 사실 기반 정보가 정확히 포함되어 있는지를 비교하여 정확도를 측정합니다.

4. 정보 검색 정확도

쿼리에 대한 응답이 실제 인덱싱된 정보(예: 논문 DB)와 일치하는지를 확인합니다.

이를 통해 LLM이 단순 생성이 아니라 정확한 정보 검색 기능까지 수행하는지 여부를 판단할 수 있습니다.

5. 모델 간 비교 분석

동일한 쿼리에 대해 여러 모델이 생성한 응답을 비교함으로써, 상대적인 품질과 일관성을 분석합니다.

예를 들어 각 모델의 Cosine Similarity 평균값, Success Rate 등을 비교할 수 있습니다.