RAG 시스템 평가 보고서

생성일: 2025년 06월 30일 02:23

1. 평가 개요

- 평가 방법: RAGAS (RAG Assessment) 프레임워크
- 평가 데이터: Synthetic 평가 세트 (10개 샘플)
- 평가 모델: klue/roberta-base 토크나이저
- 평가 시간: 약 203초

2. 평가 지표 설명

지표명	설명	점수 범위
정합성 (Faithfulness)	생성된 답변이 제공된 컨텍스트에 얼마나 충실한지 측정	0~1 (높을수록 좋음)
정답 관련성 (Answer Relevancy)	생성된 답변이 사용자 질문과 얼마나 관련있는지 측정	0~1 (높을수록 좋음)
컨텍스트 재현율 (Context Recall)	관련된 모든 정보가 컨텍스트에 포함되었는지 측정	0~1 (높을수록 좋음)
컨텍스트 정밀도 (Context Precision)	컨텍스트에 포함된 정보가 얼마나 관련있는지 측정	0~1 (높을수록 좋음)
의미 유사도 (Semantic Similarity)	생성된 답변과 참조 답변의 의미적 유사도 측정	0~1 (높을수록 좋음)

3. 평가 결과

평가 지표	점수	등급	해석
정합성 (Faithfulness)	0.5156	보통	개선 필요
정답 관련성 (Answer Relevancy)	0.6622	양호	적당한 수준
컨텍스트 재현율 (Context Recall)	1.0000	우수	완벽한 정보 포함
컨텍스트 정밀도 (Context Precision)	1.0000	우수	완벽한 관련성
의미 유사도 (Semantic Similarity)	0.9026	우수	매우 높은 유사도

4. 성능 분석

- 우수한 부분:
- 컨텍스트 재현율과 정밀도가 1.0으로 완벽함
- 의미 유사도가 0.90으로 매우 높음

- 정보 검색 및 임베딩 시스템이 효과적으로 작동
• 개선이 필요한 부분:
- 정합성(0.52)이 보통 수준으로, 답변 생성 품질 향상 필요
- 정답 관련성(0.66)이 양호하지만 더 개선 가능
• 전체 평가:
- 정보 검색 및 임베딩: 우수 (검색 엔진이 관련 정보를 잘 찾음)
- 답변 생성 품질: 보통~양호 (생성 모델의 정확성 개선 필요)
- 시스템 안정성: 우수 (일관된 성능 제공)

5. 개선 방안

• 답변 생성 품질 향상:
- 프롬프트 엔지니어링 개선
- 컨텍스트 길이 최적화
- 답변 검증 로직 추가
• 모델 튜닝:
- 더 정확한 답변 생성을 위한 파인튜닝
- 도메인 특화 학습 데이터 추가
• 평가 지속:
- 정기적인 성능 모니터링
- 사용자 피드백 반영