규정 문서 기반 LLM 응답의 인용 품질 평가 지표

김동현, 이수린, 이흥노
광주과학기술원

dearkimdh@gm.gist.ac.kr, leesurin@gm.gist.ac.kr, heungno@gist.ac.kr

Citation Quality Metrics for LLM Responses on Regulatory Documents

Kim Dong Hyeon, Lee Su Rin, Lee Heung No
Gwangju Institute of Science and Technology

요 약

본 논문은 규정 기반 질의응답 시스템에서 기존 RAG 평가지표가 인용 오류를 식별하고 이를 평가에 온전히 반영하지 못하는 문제를 해결하기 위해 인용 정밀도, 인용 재현율, 인용 충실도의 3 가지 지표를 정의하고, 37 건의 학칙 질의응답 데이터셋에서 3 종의 추론 모델을 대상으로 지표의 평가 능력을 검증하였다. 제안하는 성능평가를 이용한 실험 결과, 동일한 답변에서 과잉 인용, 조항 누락, 귀속 오류 등 인용 오류를 세부 지표가 평가 가능함을 보였다.

Ⅰ. 서 론

규정 문서를 다루는 질의응답 시스템에서 답변의 정확성뿐 아니라 근거 조항의 정확한 인용이 중요하다. 잘못된 인용은 규정 해석의 오해를 초래할 수 있다. 사용자가 제시된 조항을 통해 답변을 검증하는 것을 어렵게 하기 때문이다. 그러나 기존 RAG 평가지표인 충실도(faithfulness)나 답변 관련성(answer relevance)[3]은 답변이 맥락에 충실한지, 질문 의도에 부합하는지를 평가하는 데 한정되어 인용 자체의 오류를 직접 측정하지 못한다.

본 연구에서는 동일한 답변을 생성한 모델들 사이에서도 조항 누락, 잘못된 조항 귀속, 과잉 인용 등 인용 방식에 따라 신뢰도가 달라지는 현상을 관찰하였다. 본 연구의 기여는 다음과 같다. 첫째, 과잉 인용, 조항 누락, 귀속 오류를 각각 탐지하는 3 가지 인용 품질 지표를 정의하였다. 둘째, 각 LLM 기반의 평가 지표가 언어 모델에 있는 인용 오류를 식별하고 점수에 반영할 수 있는지 실험적으로 검증하였다.

Ⅱ. 본론

본 논문은 인용 오류 유형을 구분하기 위해 인용 정밀도(citation precision), 인용 재현율(citation recall), 인용 충실도(citation faithfulness)의 3 가지 지표를 다음과 같이 정의한다.

1. 인용 품질 지표 정의

본 연구는 인용을 "제A조 제B항"과 같은 조항 번호로 식별한다. 이를 바탕으로 인용 오류 유형을 구분하는 3 가지 지표를 정의한다.

인용 정밀도(Citation Precision)는 모델이 인용한 조항 중 정답 조항의 비율로, 과잉 인용을 탐지한다.

$$\text{Precision} = \frac{|\text{인용 조항} \cap \text{정답 조항}|}{|\text{인용 조항}|}$$

인용 재현율(Citation Recall)은 정답 조항 중 모델이 인용한 비율로, 조항 누락을 탐지한다.

$$\text{Recall} = \frac{|\text{인용 조항} \cap \text{정답 조항}|}{|\text{정답 조항}|}$$

인용 충실도(Citation Faithfulness)는 정답 일치 여부와 무관하게, 각 인용에 대해 다음 세 가지를 LLM 이 평가한다: (1) 존재성은 인용된 조항이 원본 문서에 실제로 존재하는가, (2) 내용 타당성은 인용 내용이 원문을 왜곡 없이 반영하는가, (3) 귀속 정확성은 인용 내용이 명시된 조항에서 실제로 유래하는가를 의미한다. 마지막으로 세 가지 기준을 모두 만족하는 인용의 비율을 산출한다. 이와 같이 설계한 목표는 규칙 기반으로 평가하기 어려운 세부 항 인용 오류를 탐지하기 위함이다. 다만, LLM 모델에 의존하므로 평가자가 선택한 모델의 성능에 따라 결과가 달라질 수 있다는 한계가 있다.

2. 실험 설정

본 실험은 지표가 가지는 인용 오류 식별 능력의 평가를 목적으로 한다. 평가 데이터셋으로 광주과학기술원 학칙을 기반으로 복합적인 규정 해석이 요구되는 37 건의 질의응답 쌍을 구축하였다. 생성 모델로는 Artificial Analysis 플랫폼의 오픈소스 LLM 리더보드를 참고하여 추론 성능이 우수한 DeepSeek-R1, Kimi-k2-thinking, Qwen3-VL-235B 의 3 종 오픈소스 모델을 선정하였고[2], 평가 모델은 Judge-Bench[1] 기준 최고 정확도를 기록한 Gemini 3 Flash 를 사용하였다.

3. 결과 및 사례 분석

선정된 3 가지 오픈소스 모델을 대상으로 37 건의 데이터셋 전체에 대해 정의한 3 가지 인용 지표를 평가하였다. 실험 결과는 표 1 과 같다.

모델	인용 정밀도	인용 재현율	인용 충실도
DeepSeek R1	0.4244	0.5878	0.9894
Kimi K2 Thinking	0.4461	0.7523	0.9775
Qwen3 VL 235B A22B Thinking	0.4550	0.6937	0.9842

표 1. 모델별 성능 평가 결과

인용 정밀도가 낮은 것은 과잉 인용을, 인용 재현율의 차이는 조항 누락 여부를, 인용 충실도는 귀속 오류를 각각 반영한다. Kimi K2 Thinking 이 인용 재현율 0.75 로 가장 높았으며, 인용 충실도는 모든 모델이 0.97 이상으로 대체로 원문을 정확히 인용하였다. 반면 인용 정밀도는 0.42~0.46 으로, 불필요한 조항을 함께 인용하는 경향이 있었다.

3.1 대표 사례 분석

정의한 지표의 구분 능력을 시연하기 위해 부전공 인정 요건에 관한 질의를 대표 사례로 선정하였다. 이 질의는 "주전공 이외에 15 학점 이상을 취득해야 한다"는 기본 요건을 규정한 제 71 조의 2 제 2 항과, "세부사항은 총장이 따로 정한다"는 위임 규정인 제 4 항을 모두 참조해야 완전한 답변이다.

세 모델 모두 "15 학점 이상"이라는 동일한 답변을 제시했으나 조항 인용 방식에서 차이를 보였다. 표 2 는 각 모델의 인용 지표 평가 결과이다.

모델	인용 정밀도	인용 재현율	인용 충실도
DeepSeek R1	1.00	0.50	1.00
Kimi K2 Thinking	1.00	0.50	0.00
Qwen3 VL 235B a22B Thinking	1.00	1.00	1.00

표 2. 대표 사례에 대한 인용 지표 평가 결과

DeepSeek-R1 은 제 71 조의 2 제 2 항만 명시적으로 인용하고, 세부사항은 "총장이 별도로 정하므로"라고 언급하였으나 제 4 항을 인용하지 않아 인용 재현율 0.50 을 기록하였다. 인용한 조항의 내용은 정확히 서술하여 인용 충실도 1.00 을 받았다.

Kimi-k2-thinking 은 제 71 조의 2 제 2 항을 인용하면서 "세부사항은 총장이 따로 정한다고 명시되어 있으므로"라고 서술하였다. 이 표현은 세부사항 규정이 제 2 항에 포함된 것처럼 해석되나, 실제로는 제 4 항에 별도로 규정되어 있다. 이러한 귀속 오류로 인해 인용 충실도 0.00 을 기록하였다.

Qwen3-VL-235B 는 "제 71 조의 2 제 2 항에 따르면"과 "제 71 조의 2 제 4 항에 따라"를 명확히 구분하여 두 조항을 각각 인용하였다. 이에 따라 모든 지표에서 만점을 받았다.

이 사례는 답변 내용이 동일하더라도 인용 재현율이 조항 누락을, 인용 충실도가 귀속 오류를 각각 포착함을 보여준다.

Ⅲ. 결론

본 연구는 규정 기반 질의응답 시스템에서 기존 RAG 평가 지표의 한계인 인용 오류 유형을 구분하기 위해 3 가지 지표를 정의하고, 평가 지표를 이용한 실험 결과를 보였다.. 인용 정밀도는 과잉 인용을, 인용 재현율은 조항 누락을, 인용 충실도는 귀속 오류를 각각 탐지하여, 기존 RAG 평가지표가 놓치는 인용 품질의 차이를 구분할 수 있음을 보였다.

본 연구의 한계는 다음과 같다. 첫째, 37 건의 소규모 데이터로 통계적 일반화에 제약이 있다. 둘째, 인용 정밀도와 재현율이 문자열 매칭에 의존하여 표기 변형에 취약하며, 인용 충실도가 LLM 평가자에 의존하여 평가 모델의 성능에 따라 결과가 달라질 수 있다. 향후 연구에서는 지표의 통계적 유의성과 범용성을 검증하기 위해 대규모 데이터셋을 구축하고 다양한 규정 도메인으로 적용 범위를 확장할 필요가 있다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터지원사업의 연구결과로 수행되었음 (IITP-2026-RS-2021-II211835) 그리고 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임. (RS-2025-22932973)

참 고 문 헌

[1] S. Tan et al., "JudgeBench: A Benchmark for Evaluating LLM-Based Judges," in Proc. The Thirteenth International Conference on Learning Representations (ICLR), 2025.

[2] Artificial Analysis, "LLM Leaderboard: Analysis of Open-weights and Proprietary Models," 2026. [Online].

[3] S. Es et al., "RAGAS: Automated Evaluation of Retrieval Augmented Generation," in Proc. EACL 2024 System Demonstrations, 2024, pp. 150-163.