JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 2, pp. 51-64, February 2026
https://doi.org/10.9708/jksci.2026.31.02.051

RAR-Agent: A Rationale-Augmented Retrieval Framework for Legal Question Answering

Gyuhyeong Kim*, Yunhyeok Do*, Joonhyeon Song*, Ziyang Liu**

*AI Research, QI, Seoul, Korea
**Professor, Dept. of Global Business, Kyonggi University, Gyeonggi, Korea

[Abstract]

Hallucination and outdated knowledge in large language models critically undermine their reliability and applicability in specialized domains such as law and medicine, where factual accuracy is essential. While Retrieval-Augmented Generation (RAG) has been proposed as a mitigation strategy, its effectiveness in the legal domain is often hindered by lexical mismatches, which impede the accurate retrieval of highly relevant external knowledge. Although several studies have explored query formulation–based approaches to address this issue, additional training costs and hallucination during the retrieval phase remain persistent challenges. In this paper, we propose RAR-Agent (Rationale-Augmented Retrieval Agent) to overcome these limitations. RAR-Agent employs a Chain-of-Thought and Rationale-based query formulation technique, combined with a Reciprocal Rank Fusion and Reranker-based filtering mechanism, to alleviate lexical mismatch problems and effectively suppress hallucination during retrieval. Furthermore, to precisely evaluate the agent's factual accuracy, we constructed the KL-BQA (Korean Legal Binary Question-Answering) benchmark. The proposed model achieved superior performance on both the KL-BQA and KL-RQA benchmarks.

▸ **Key words**: AI Agent, Legal Agent, RAG, Legal QA, Query Formulation, Hybrid Retrieval

[요 약]

대규모 언어 모델의 환각 및 지식 노후화 문제는 법률, 의료와 같은 사실적 정확성이 필수적인 전문 도메인에서의 신뢰도와 활용성을 저해한다. 이러한 문제를 완화하기 위한 접근으로, RAG(Retrieval-Augmented Generation)가 제안되었지만, 법률 도메인의 어휘적 불일치로 인하여 높은 관련성을 지닌 외부 지식을 정확히 탐색하지 못하는 한계가 존재한다. 이를 보완하기 위한 Query Formulation 기반의 연구들이 다수 등장하였으나, 추가적인 학습 비용과 검색 단계에서 발생하는 환각 문제는 여전한 과제로 남아있다. 본 연구에서는 기존 연구들의 한계를 극복하기 위한 RAR-Agent를 제안한다. RAR-Agent는 Chain-of-Thought와 Rationale 기반의 Query Formulation 기법, 그리고 Reciprocal Rank Fusion 및 Reranker 기반 필터링 메커니즘을 활용하여 어휘적 불일치 문제를 완화하고 검색 단계의 환각을 효과적으로 억제한다. 또한, 에이전트의 사실적 정확성을 정밀하게 측정하기 위해, KL-BQA 벤치마크를 구축하였고, KL-BQA 및 KL-RQA 벤치마크 모두에서 우수한 성능을 달성하였다.

▸ **주제어**: AI 에이전트, 법률 에이전트, 검색 증강 생성, 법률 질의응답, 쿼리 구성, 하이브리드 검색

• First Author: Gyuhyeong Kim, Corresponding Author: Ziyang Liu
*Gyuhyeong Kim (gyuhyeong@attenote.com), QI
*Yunhyeok Do (yunhyeokd@attenote.com), QI
*Joonhyeon Song (nemosong@attenote.com), QI
**Ziyang Liu (victor@kgu.ac.kr), Dept. of Global Business, Kyonggi University
• Received: 2025. 11. 11, Revised: 2025. 12. 28, Accepted: 2026. 02. 02.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

최근 대규모 언어 모델(Large Language Models, LLMs) 기반 에이전트들이 다양한 과제에서 탁월한 성능을 보이며 법률, 의료 등 전문 도메인으로까지 그 활용 범위를 빠르게 확장하고 있다[1-3]. 그러나 이러한 성능적 우수성에도 불구하고, LLM이 가지고 있는 환각(hallucination)과 지식 노후화(outdated knowledge) 문제는[4, 5], 사실적 정확성(factual accuracy)과 최신 지식 반영이 필수적인 전문 도메인에서 에이전트의 신뢰도를 저해하는 치명적인 결함으로 작용하며 그 활용성을 제한한다(Fig. 1).

이러한 한계를 극복하기 위한 접근법으로, RAG(Retrieval-Augmented Generation)가 주목받고 있다. RAG[6]는 외부 지식(external knowledge)을 탐색하여 사용자 질문과 관련된 정보를 LLM에 제공하고, 이를 바탕으로 맥락에 맞는 답변을 생성하는 접근법이다(Fig. 2). 이 방식은 LLM의 환각을 억제하고 더 신뢰도 높은 추론을 가능하게 함으로써, 다양한 과제에서 높은 잠재력과 가능성을 보여주었다[7-9].

그러나 RAG를 도입했음에도 불구하고, 법률과 같은 전문 도메인에서는 여전히 다음과 같은 한계가 존재한다. 사용자 질문(query)과 판례 원문 사이에 존재하는 어휘적 불일치(lexical mismatch)로 인해 사용자 질문과 높은 관련성을 지닌 외부 지식을 정확히 탐색하지 못하는 검색 실패(retrieval failure) 문제에 직면한다[10-12].

이러한 문제를 완화하기 위해 검색에 최적화된 쿼리를 생성하는 Query Formulation(QF) 접근법이 등장하였다. 이들은 대부분, (1) 미세조정(fine-tuning)이나 강화학습(Reinforcement Learning)을 통해 특정 도메인에 특화된 최적의 검색 쿼리(Search Query)를 생성하는 방식이나[10][13-19], (2) prompting LLM을 통해 검색 성능을 극대화하는 방식이 주를 이룬다[20-22].

그러나 이러한 QF 접근법들은 다음과 같은 한계가 존재한다. 미세조정 및 강화학습 방식은 학습 데이터셋을 구축하기 위한 레이블링 작업과 학습 과정에서 추가적인 비용이 요구되며, prompting LLM 방식은 LLM의 환각 문제가 검색 단계에서 치명적인 오류를 야기한다.

본 연구는 이러한 접근법들의 한계를 극복하기 위해 RAR-Agent(Rationale-Augmented Retrieval Agent)를 제안한다(Fig. 3). 별도의 학습 없이 Chain-of-Thought (CoT) prompting[26]을 통해 LLM의 추론(reasoning) 능력을 체계적으로 활용하였고, Rationale-Based Query Formulation 기법과 Reciprocal Rank Fusion(RRF) 및 Reranker 기반의 Filtering Mechanism을 사용하여 환각 문제를 효과적으로 개선하였다.

또한, 제안하는 방법론의 유효성을 검증하고 법률 에이전트의 신뢰성을 확보하기 위해서는, 한국어 법률 도메인에 특화된 벤치마크와 이를 객관적으로 평가할 수 있는 평가지표가 필수적이다. 기존에 이를 평가하기 위한 QA 벤치마크가 다수 존재하나[23, 24], 이들은 모두 영문으로 되어있어 한국어 법률 에이전트의 성능을 정량적으로 평가하기에는 적합하지 않다. 이외에도, RAGAS[25]와 같이 LLM을 평가자로 활용하는 지표는, 생성된 답변의 사실적 정확성을 직접 측정하지 않고 LLM의 판단에 의존하기에, 높은 신뢰성이 요구되는 법률 에이전트의 주 평가지표로 사용하기에 한계를 지닌다.

이에 본 연구는 법률 에이전트의 사실적 정확성을 객관적으로 평가하기 위해, AI-Hub의 법률 QA 데이터셋 중 '예/아니요(Yes/No)'로 명확히 판별 가능한 337개 QA 쌍을 선별하고, 수동으로 라벨링 하여 KL-BQA(Korean Legal Binary Question-Answering) 벤치마크를 구축하였다. 본 벤치마크를 활용하여 RAR-Agent의 정확도를 Accuracy와 F1-Score 지표를 통해 정량적으로 평가하였다.

우리의 RAR-Agent는 검색 성능, RAGAS, 정확도 모든 면에서 비교 모델 대비 가장 우수한 성능을 보였다. KL-RQA 벤치마크에서 약 86% 성능(NDCG@20)을 보였으며, KL-BQA 벤치마크에서는 GPT-5 기준 97.92% 정확도(Accuracy)를 달성하였다.

본 연구의 주요 기여점은 다음과 같다.

첫째, 한국어 법률 도메인에 최적화된 Agentic RAG 프레임워크인 RAR-Agent를 제안하였다.

둘째, 별도의 학습 없이 CoT prompting을 활용한 QF 기법을 제안하여, 어휘적 불일치와 환각 문제를 완화하였고 기존 RAG의 법률 문서 검색 성능을 크게 향상시켰다.

셋째, 검색 성능뿐만 아니라, RAR-Agent의 정확도까지 정량적으로 측정할 수 있는 KL-BQA 벤치마크를 구축하였고, 이를 통해 제안하는 에이전트의 뛰어난 사실적 정확성을 검증하였다.

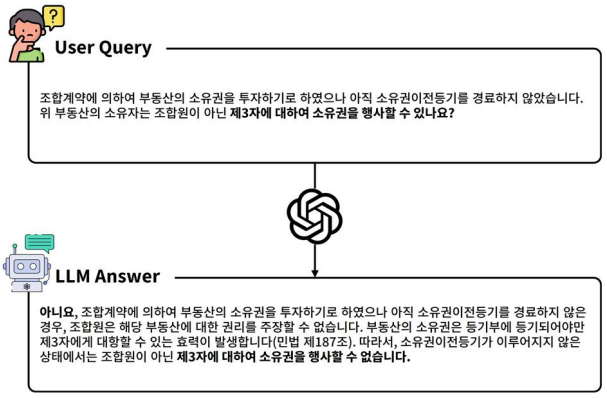

Fig. 1. Failure Case of LLM Only in Legal QA

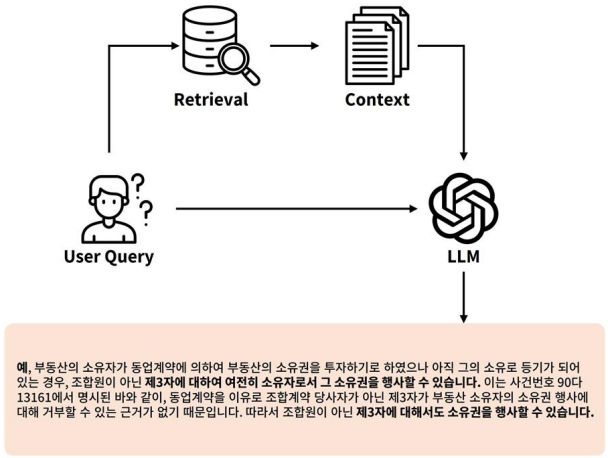

Fig. 2. Success Case of Standard RAG in Legal QA

II. Related Work

2.1 RAG

최근 법률 도메인에서 LLM의 잠재력을 활용하기 위해, LLM이 내재적으로 가지는 환각과 지식 노후화 문제를 보완하기 위한 다양한 연구가 활발히 진행되고 있다. RAG는 이러한 한계를 극복하기 위한 대표적인 접근법이다. 외부의 신뢰할 수 있는 지식을 결합해 LLM의 응답을 보강함으로써, 답변의 신뢰도와 정확도를 향상시킨다[6][28, 29]. 일반적인 RAG 파이프라인은 Fig. 2와 같이 사용자 질문(User Query)에 기반한 근거 문서(context) 검색(Retrieval)과 LLM의 답변 생성 단계를 결합한 형태로 구성된다.

RAG를 고도화하려는 연구 역시 활발히 진행되고 있다. 예를 들어, Self-RAG[7]는 검색된 문서를 스스로 비판하고 교정하는 과정을 통해 답변 품질을 개선하였으며, Adaptive-RAG[30]는 질문 복잡도(Query Complexity)를 판단하여 상황에 맞는 검색 전략을 동적으로 선택함으로써, 효율성과 성능을 동시에 향상시킨다.

그러나 RAG의 전반적인 성능은 검색 구성 요소(Retrieval Component)의 품질에 크게 의존한다. 특히, 법률과 같은 전문 도메인에서는 사용자 질문과 판례 원문 간의 어휘적 불일치(lexical mismatch)가 빈번하게 발생하여, 질문과 관련성이 높은 chunk를 찾지 못하는 검색 실패 문제[10-12]가 나타난다. 이러한 검색 단계에서의 한계를 보완하기 위해, 검색에 사용되는 검색 쿼리 자체를 개선하여, 질문과 문서 간의 어휘적 불일치를 줄이고 검색 성능을 향상시키는 Query Formulation 연구가 주목받고 있다.

2.2 Query Formulation

Query Formulation은 사용자 질문을 검색에 더 효과적이고 최적화된 질문으로 변환하는 방식을 말한다. 일반적으로 'Query Rewriting'과 'Query Expansion'으로 나눌 수 있다.

2.2.1 Query Rewriting

Query Rewriting은 사용자의 원본 질문(Original Query)을 검색에 적합한 형태로 재구성하는 방법이다. 사용자의 쿼리가 모호하거나 불완전할 때, 의미를 명확히 하거나 누락된 정보를 보완하여 검색 의도를 정확히 반영하도록 돕는다. 기존 연구에서는 언어 모델(Language Model, LM)을 Query Rewriter[10][13-15][31]로 학습시키거나, 강화학습[16-19]을 활용해 쿼리 재작성의 품질을 향상시키는 접근법이 주로 사용되었다.

그러나 이러한 연구들은 맥락의 모호성 해소에 초점을 맞추었기 때문에, 전문 도메인에서 발생하는 어휘적 불일치 문제를 근본적으로 해결하지는 못하는 한계가 있다.

2.2.2 Query Expansion

Query Expansion은 원본 질문을 확장함으로써 사용자 질문과 외부 지식 간의 어휘적 차이를 줄이고, 관련성이 높은 문서를 보다 효과적으로 검색하도록 하는 접근법이다.

최근에는 LLM이 가지고 있는 방대한 지식과 연관 개념 추론 능력을 활용하는 연구가 활발히 이루어지고 있다[27][32]. 예를 들어, Query2doc[21]은 few-shot prompting을 통해 LLM이 가상의 문서(pseudo-document)를 생성하도록 하고, 이를 원본 쿼리와 결합해 확장된 검색 쿼리를 구성함으로써 검색 성능을 향상시켰다. 또한 ConvGQR[15]은 미세조정(fine-tuning) 된 언어 모델을

활용해 잠재적 답변(potential answer)을 생성하여, 이를 재작성된 쿼리와 결합해 최종 검색 쿼리로 사용함으로써 검색 정확도를 개선시킬 수 있음을 보여주었다.

2.3 Reasoning in RAG

LLM은 추론 능력에서 강점을 보이지만 환각과 지식 노후화에 취약하고, RAG는 외부 지식을 통해 이를 보완하지만, 검색 품질에 의존한다. 이에 따라, LLM의 추론 능력을 RAG에 결합하려는 접근법들이 최근 주목받고 있다[34]. 특히 여러 접근법 중에서도 LLM의 추론 능력을 Query Formulation 단계에 직접 활용하여 검색 성능을 개선하려는 시도가 활발히 이루어지고 있다.

예시로, Rationale-Guided RAG[33]와 ConvGQR[15]은 LLM이 원본 질문을 기반으로 가상의 잠재적 답변(Rationale)을 먼저 생성하도록 유도한다. 이렇게 생성된 답변은 원본 질문에 포함되지 않았던 풍부한 관련 개념과 문맥 정보를 제공하여, 기존 쿼리만으로는 해결하기 어려웠던 검색 실패 문제를 완화하고 검색 성능을 향상시킨다.

그러나 이러한 LLM 추론 기반 접근법은 여전히 두 가지 한계를 지닌다. 첫째, 가상의 답변을 생성하는 과정에서 환각이 발생해 부정확한 검색을 유도할 수 있다. 둘째, 단일 쿼리에 의존하기 때문에 다양한 검색 관점을 충분히 반영하기 어렵다.

이에 본 연구는 앞서 언급한 문제를 해결할 수 있는 RAR-Agent를 제안한다. RAR-Agent는 Multi-Query 생성과 2중 Filtering Mechanism을 결합함으로써, 앞서 서술한 기존 접근법의 한계를 보완하고 보다 안정적이며 정밀한 검색 성능을 달성하고자 하였다. 이를 통해 LLM의 추론 능력과 RAG의 검색 효율성을 효과적으로 결합하는 새로운 방향을 제안하였다.

III. Methodology

본 연구에서 제안하는 RAR-Agent의 전체 파이프라인은 다음과 같다. 먼저, (1) Rationale-Based Query Formulation은 원본 질문을 입력받아, 법률적 관점에서 분석한 분석적 쿼리(Analytical Query)와 가상 판례 쿼리(Rationale Query)를 생성한다. 원본 쿼리를 포함한 3개의 쿼리 집합은 Multi-Query Hybrid Retrieval로 전달된다. (2) Multi-Query Hybrid Retrieval은 이 쿼리 집합을 활용하여 3개의 독립적인 후보 chunk 목록을 도출한다. 이후 (3) Filtering Mechanism에서 Reciprocal Rank Fusion(RRF)과 Reranker를 통해 이 목록들을 융합하고 재검증하여, 원본 질문과 가장 관련성이 높은 상위 K개의 chunk를 최종 선별한다. 마지막으로, (4) Answer Generation에서 선별된 K개의 chunk를 근거 문서(context)로 사용하여 최종 답변을 생성한다. 구체적인 RAR-Agent의 아키텍처는 Fig. 3에서 확인할 수 있다.

3.1 Rationale-Based Query Formulation

검색 성능 향상과 다양한 검색 관점을 반영하기 위해 우리는 단일 쿼리에 의존하는 대신, LLM의 추론 능력을 활용하여 법률 도메인에 특화된 3개의 쿼리를 생성한다.

3.1.1 Key Issue Extraction

첫 번째 단계에서는 어휘적 불일치를 완화하기 위해, LLM을 법률 분석가로 호출한다. 이 LLM은 원본 질문을 법률적 관점에서 분석하여 질문에 내재된 핵심 법률 용어(key legal terms)와 주요 법적 쟁점(primary legal issue)을 반영하는 분석적 쿼리(Analytical Query)를 생성한다. 이 생성 과정은 수식 (1)과 같다.

$$Q_{analy} \sim LLM_{analy}(Q_{orig}) \tag{1}$$

여기서 Q_{orig}는 사용자의 원본 질문이며, LLM_{analy}은 법률 분석을 진행하는 LLM이다. 그 결과인 Q_{analy}는 원본 질문의 핵심 의도를 법률 용어와 법적 쟁점 중심으로 명료화한 쿼리이다.

이렇게 생성된 쿼리는 원본 질문에는 명시적으로 등장하지 않지만, 판례 원문에 포함될 가능성이 높은 핵심 키워드를 포함함으로써 희소 검색(Sparse Retrieval)의 성능을 실질적으로 향상시킨다. 또한, LLM(법률 분석가)이 원본 질문에 내재된 의미를 논리적으로 해석하는 이 과정은, 다음 단계인 Hypothetical Rationale Generation에서 발생할 수 있는 환각을 효과적으로 억제하는 역할을 수행한다.

3.1.2 Hypothetical Rationale Generation

최근 다수의 RAG 연구에서 LLM이 생성한 가상의 답변을 검색 쿼리로 활용할 경우, 원본 쿼리보다 더 높은 검색 성능을 달성할 수 있음을 보여주었다. 이에 기반하여, 본 연구에서는 LLM을 판사로 활용하여 가상 판례(Rationale)를 생성한다.

본 연구는 가상 판례 생성 시 환각을 완화하기 위해 Contextual Scaffolding 방식을 제안한다. 이는 원본 질

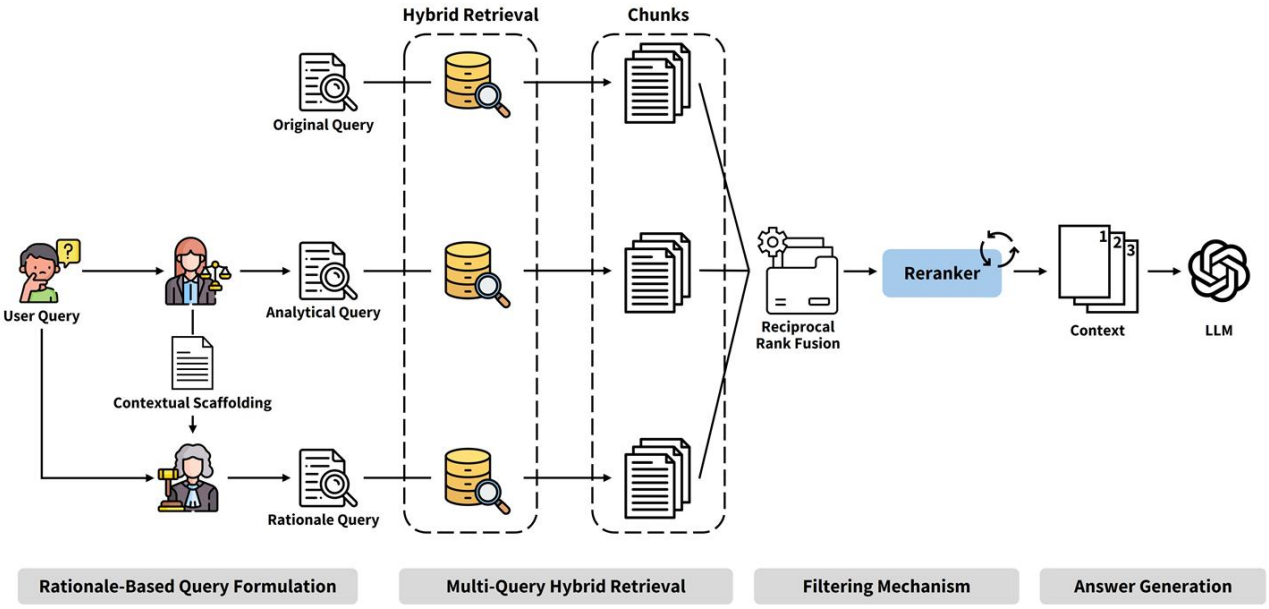

Fig. 3. The Rationale-Augmented Retrieval (RAR) Agent Framework

문뿐만 아니라, 3.1.1에서 도출된 분석적 쿼리를 LLM(판사)에 입력으로 함께 제공하는 방식이다. 이 접근법은 LLM이 가상 판례를 생성할 때, 추론 범위를 문맥적으로 제약하고 생성 결과의 일관성을 높인다. 아울러, LLM의 추론 능력을 향상시키기 위해 CoT를 사용한다.

LLM이 Contextual Scaffolding을 기반으로 가상 판례 쿼리(Rationale Query)를 생성하는 과정은 다음 수식 (2)와 같다.

$$Q_{rati} \sim LLM_{judge}^{CoT}(Q_{orig}, Q_{analy}) \tag{2}$$

이처럼 Contextual Scaffolding과 CoT prompting을 통합한 방식은, 원본 쿼리와의 의미적 일관성을 유지하면서도 내용적으로 풍부한 Q_{rati}를 생성하게 하여, 이후의 검색 단계에서 보다 높은 정확도와 관련성을 확보하도록 한다.

3.2 Multi-Query Hybrid Retrieval

우리는 3.1에서 생성된 분석적 쿼리와 가상 판례 쿼리뿐만 아니라, 사용자의 질문 의도를 온전히 담고 있는 원본 쿼리까지 포함한 3개의 쿼리를 모두 검색에 사용한다.

$$Q_{set} = \{Q_{orig}, Q_{analy}, Q_{rati}\} \tag{3}$$

이는 사용자의 실제 질문 의도와 LLM의 추론을 통해 법률적 맥락으로 풍부해진 쿼리들을 동시에 활용해 검색 실패를 최소화하기 위함이다.

우리는 '의미'와 '키워드'를 모두 고려하기 위해, Dense Retrieval과 Sparse Retrieval을 결합한 Hybrid Retrieval을 사용한다. 각 chunk의 최종 하이브리드 점수는 수식 (4)와 같이 두 검색 방식의 점수에 각각 가중치를 부여하여 합산하는 방식으로 계산된다.

$$R_{hyb} = w_d R_{dense} + w_b R_{bm25} \tag{4}$$

우리는 Q_{set}에 속하는 각 쿼리 q에 대해 R_{hyb}를 독립적으로 적용하여, 상위 N개의 후보 chunk 목록(L_q)을 얻는다.

$$\forall_q \in Q_{set} \qquad L_q = R_{hyb}(q, N) \tag{5}$$

이 과정을 통해 최종적으로 원본 쿼리, 분석적 쿼리, 가상 판례 쿼리에 대한 각각의 검색 결과에 해당하는 3개의 후보 chunk 목록 L_{set}이 생성된다.

$$L_{set} = \{L_{orig}, L_{analy}, L_{rati}\} \tag{6}$$

3.3 Filtering Mechanism

추가로 일부 환각으로 인한 부정확한 검색 결과가 도출될 경우를 방지하기 위해, 이를 최종 단계에서 효과적으로 걸러내기 위한 2중 Filtering Mechanism을 제안한다.

3.3.1 Reciprocal Rank Fusion

L_{set}을 단순히 통합할 경우, 하나의 환각 쿼리가 도출한 관련성 낮은 chunk가 최종 후보군의 상위 순위를 오염시킬 수 있다. 이를 해결하기 위해, 우리는 순위 가중 투표 방식인 Reciprocal Rank Fusion(RRF)을 사용하여 3개의 후보 목록을 하나의 목록으로 융합한다.

RRF는 하나의 쿼리에서만 1위를 한 chunk에는 낮은 점수를 부여하고, 여러 쿼리에서 공통으로 높은 순위에 있는 chunk에는 높은 점수를 부여한다. 각 chunk d의 최종 RRF 점수(S_{RRF})는 수식 (7)을 통해 계산된다.

$$S_{RRF}(d) = \sum_{L \in L_{set}} \frac{1}{k_{rrf} + rank_L(d)} \tag{7}$$

모든 후보 chunk의 점수를 계산한 뒤, 내림차순으로 정렬하여 하나의 융합된 목록(L_{fused})을 생성한다. 이 목록에서 상위 M개의 chunk만 선별하여 Reranker에 전달할 후보군(C_{cand})을 구성한다. 선별 과정을 수식으로 나타내면 (8)과 같다.

$$C_{cand} = Top_M(L_{fused}) \tag{8}$$

3.3.2 Reranker

RRF 알고리즘을 적용한 후에도 혹시 남아있을 수 있는 관련성 낮은 chunk를 최종적으로 필터링하기 위해, Reranker(RR)를 사용한다.

선별된 M개의 후보 chunk $d \in C_{cand}$ 각각을 원본 질문과 1:1로 비교하여, 원본 질문과의 의미적 부합성을 나타내는 관련성 점수를 계산한다. 이 과정은 (9)와 같다.

$$S_{rr}(d) = RR(Q_{orig}, d) \tag{9}$$

이후, 후보군(C_{cand})을 계산된 관련성 점수(S_{rr})가 높은 순으로 정렬하고, 최종적으로 상위 k개의 chunk만 선별하여 C_{final}을 구성한다. k개의 chunk는 하나로 통합되어 근거 문서(context)로서 LLM에 제공된다. 이 과정은 (10)과 같다.

$$C_{final} = Top_k(sort_{desc}(C_{cand}, S_{rr})) \tag{10}$$

3.4 Answer Generation

근거 문서와 원본 질문을 LLM에 함께 제공하며, (11)과 같이 최종적으로 에이전트의 답변을 생성한다.

$$A_{final} \sim LLM_{answer}(Q_{orig}, C_{final}) \tag{11}$$

IV. Experiments

4.1 Datasets

RAR-Agent를 평가하기 위해, 우리는 민사 판례 Corpus와 KL-RQA, KL-BQA 벤치마크를 구축했다. 모든 데이터는 AI-Hub에서 제공하는 '법률/규정 텍스트 분석 데이터(고도화)-상황에 따른 판례 데이터'를 기반으로 한다.

4.1.1 Corpus Construction

우리의 Corpus는 AI-Hub의 '상황에 따른 판례 데이터' 중, 민사 판례로 한정했다. 민사 판례는 법적 분쟁의 가장 기본적이고 광범위한 영역을 다루므로, RAG 시스템의 일반적인 법률 추론 성능을 검증하기에 가장 적합하다고 판단했다. 최종적으로 총 13,582개의 민사 판례 원문을 Corpus 구축 대상으로 선정했다.

판례 원문은 '판시사항', '판결 요지', '판례 내용' 등 명확히 Section 별로 구성되어 있다. 우리는 이러한 판례 원문의 내재적 구조를 보존하는 것이 검색 품질에 중요하다고 판단해 다음과 같은 'Structure-Aware Chunking' 전략을 설계했다.

첫째, 먼저 각 판례 원문을 '판시사항', '판결 요지', '참조 조문', '참조 판례', '판례 내용'으로 분할한다.

둘째, 각 Section의 텍스트가 설정한 글자 수 임계값(chunk_size)을 초과할 때만, Recursive Character Text Splitter를 적용해 구분자(separators)를 기준으로, 재귀적으로 분할한다.

셋째, 각 chunk에는 case_no(사건 번호), decision_date(선고 일자)와 같은 판례 원문의 메타데이터와 함께, section(판례 원문 section 명칭), section_order(해당 section 내 순서)를 추가한다.

이 전략을 통해 13,582개의 판례 원문으로부터 총 117,769개의 chunk를 생성해 ChromaDB를 구축했다.

4.1.2 Benchmark Construction

RAR-Agent의 성능을 다각도로 측정하기 위해, AI-Hub의 QA 데이터 20,160개를 기반으로 2개의 독립적인 벤치마크를 구축했다.

KL-RQA

검색 성능을 측정하기 위해 다음과 같이 엄격한 세 단계 필터링을 통해, KL-RQA(Korean Legal Retrieval Question-Answering) 벤치마크를 구축했다.

첫째, 전체 QA 데이터 중 민사 도메인에 해당하는 1,672개의 QA를 선별한다.

둘째, 검색 성능 평가의 정답(Ground Truth) 기준이 되는 사건 번호 값이 누락된(null) QA를 제외해 634개를 선별한다.

셋째, 정답 기준이 되는 사건 번호가 구축한 ChromaDB 내에 실제로 존재하는지 확인하여, 검색 가능한 질문만을 대상으로 하는 공정한 평가 환경을 구축한다.

이 모든 과정을 통과한 392개의 QA 쌍을 검색 성능 평가에 사용했다.

KL-BQA

RAGAS와 같이 LLM을 평가자로 활용하는 방식은 생성된 최종 답변의 정확도를 직접적으로 측정하기에는 한계가 있다. 이에 우리는 RAR-Agent가 도출한 최종 답변의 사실적 정확성을 엄밀하게 평가하기 위해, 다음과 같이 KL-BQA(Korean Legal Binary Question-Answering) 벤치마크를 구축했다.

Table 1. Distribution of Positive and Negative Classes in the KL-BQA Benchmark

Class	Count	Percentage (%)
Positive (예)	173	51.3
Negative (아니요)	164	48.7
Total	337	100.0

첫째, KL-RQA 벤치마크의 정답 답변(Ground Truth Answer)을 분석해 '예' 또는 '아니요'로 답변이 시작하고 명확히 판별할 수 있는 337개의 QA를 선별한다.

둘째, 선별된 337개의 QA 쌍에 대해 긍정('예')일 경우 '1', 부정('아니요')일 경우 '0'으로 라벨링을 수동으로 진행한다.

KL-BQA 벤치마크의 상세한 클래스 분포는 Table 1과 같다.

4.2 Evaluation Metrics

4.2.1 Retrieval Performance Metrics

KL-RQA 벤치마크를 사용해, Reranker가 최종 선별한 Top-k 목록을 2가지 지표를 사용해 평가한다.

Recall@k는 k개의 검색 결과 목록에 정답 chunk가 하나라도 포함되어 있는지를 측정한다. k개 안에 정답 chunk가 하나라도 포함되어 있다면 '1'을, 없다면 '0'을 부여한다.

NDCG@k는 여러 개의 정답 chunk가 상위권에 잘 배치되어 있는지를 종합적으로 측정하는 순위 품질 지표이다. 여러 개의 정답 chunk가 모두 상위권에 배치되어 있다면, 1에 가까운 높은 점수를 부여한다.

4.2.2 RAGAS

KL-BQA 벤치마크를 대상으로 RAGAS를 사용해 RAR-Agent가 생성한 최종 답변의 품질을 Faithfulness, Answer Relevancy, Answer Correctness 지표로 평가한다. 이 과정에서 RAGAS는 LLM을 평가자로 활용해 해당 지표들을 측정한다.

Faithfulness

생성된 답변이 제공된 근거 문서에 얼마나 충실한지를 측정한다. LLM이 답변을 여러 개의 주장으로 분해하고, 각 주장이 근거 문서에 의해 명시적으로 뒷받침되는지를 검증하는 방식으로 이루어진다.

Answer Relevancy

생성된 답변이 사용자 질문의 의도에 얼마나 부합하는지를 측정한다. LLM이 생성된 답변을 바탕으로 여러 개의 질문을 생성한 뒤, 이 질문들과 사용자 질문 간의 의미론적 유사도(semantic similarity)를 측정하여 계산된다.

Answer Correctness

생성된 답변이 정답 답변(Ground Truth Answer)과 의미론적으로 얼마나 일치하는지를 측정한다. LLM이 생성된 답변과 정답 답변을 비교하여 사실관계가 일치하는지를 판단하는 방식으로 점수가 부여된다.

4.2.3 Answer Accuracy Metrics

RAGAS는 최종 답변 품질만 평가할 뿐, 최종 답변의 사실적 정확성은 측정하지 못한다. 이에 본 연구에서는 보다 정확하고 신뢰성 있는 평가를 위해 KL-BQA 벤치마크를 사용해 최종 답변의 정확도를 측정한다.

Accuracy & F1-Score

Accuracy는 전체 질문 중 '예'를 '예'로, '아니요'를 '아니요'로 올바르게 예측한 비율이고, F1-Score는 정밀도와

재현율의 조화 평균을 의미한다. 이들은 모델의 실질적인 분류 성능과 신뢰도를 평가하는 대표적인 지표이다.

4.3 Baseline Model

제안하는 RAR-Agent의 성능을 비교하기 위해, 우리는 LLM 단독 모델부터 Hybrid Retrieval과 Reranker로 조합한 모델에 이르기까지, 네 가지 베이스라인 모델을 설정하였다.

LLM Only

가장 기본적인 LLM의 성능을 측정하기 위해 오직 LLM의 추론 능력과 내부 지식에만 의존해 측정한다.

Standard RAG

가장 표준적인 RAG 구조로, Dense 검색기를 사용하여 LLM에 근거 문서를 제공한다. 이를 통해 외부 지식 결합이 LLM의 답변 품질에 미치는 효과를 평가한다.

Dense + Reranker

Standard RAG에 Reranker 모델을 추가하여 법률 도메인과 같이 문서 간 의미 차이가 미묘한 환경에서, 관련성 판단을 정밀하게 수행할 수 있는지를 검증한다.

Hybrid + Reranker

법률 텍스트와 같이 의미적 유사성과 어휘적 일치가 모두 중요한 도메인에서는 Dense Retrieval만으로는 한계가 존재한다. 이에 Sparse Retrieval(BM25)을 결합한 Hybrid Retrieval을 베이스라인으로 추가한다.

4.4 Implementation Details

본 연구는 Ubuntu 22.04 LTS 및 NVIDIA RTX 3090 GPU 환경에서 진행되었다.

4.4.1 Architecture Components

본 실험에서는 LLM으로 OpenAI의 'GPT-4o-mini-2-024-07-18'를, 임베딩 모델로는 'text-embedding-3-large'를 사용했다. Vector Store는 ChromaDB를 활용했으며, Hybrid Retrieval은 text-embedding-3-large[36]와 BM25[37]를 결합하여 구현했다. Reranker로는 다국어 성능이 검증된 'jina-reranker-v2-base-multilingual'을 채택했다.

4.4.2 Hyperparameters

실험 결과의 일관성과 재현성을 보장하기 위해, 모든 LLM의 temperature는 0으로 설정했다. RAG 파이프라인의 검색 파라미터는 공정한 비교를 위해 다음과 같이 구성했다. Dense+Reranker 및 Hybrid+Reranker는 Retrieval을 통해 상위 30개의 후보 chunk를 검색하고, Reranker가 상위 30개 중에 최종 20개를 선별했다. 반면 제안하는 RAR-Agent는 3개의 Multi-Query가 각각 30개씩 chunk를 검색한 뒤, RRF를 통해 하나로 융합된 목록의 상위 30개를 Reranker에 전달해 동일하게 최종 20개를 선별했다.

RRF의 k 값은 원본 논문[35]에서 제안한 60을 사용했으며, RAR-Agent의 Hybrid Retrieval 최적 가중치를 찾기 위한 실험을 통해, Dense 40%와 Sparse(BM25) 60%의 조합이 가장 높은 성능을 보여 이를 Hybrid 가중치로 채택했다(Fig. 4 참조).

V. Results and Analysis

본 연구에서 제안하는 RAR-Agent의 우수성을 입증하기 위해, 4개의 베이스라인 모델과 성능을 종합적으로 비교 검증한다. 검증을 위해, (1) 검색 성능, (2) 최종 답변 품질, (3) 최종 답변 정확도를 측정한다.

다음으로 Ablation Study를 통해, 쿼리 생성의 핵심 기법인 Contextual Scaffolding이 환각을 억제하고 풍부한 문맥을 생성하여 성능 향상에 기여하는 정도를 검증한다. 또한, 최적의 하이브리드 검색 가중치와 LLM의 추론 능력이 최종 답변 정확도에 미치는 영향도 함께 분석한다.

마지막으로 Case Study에서는 실제 복잡한 법률 QA를 통해, 베이스라인 모델의 답변 실패 원인을 분석하고, RAR-Agent가 어떻게 정확한 법적 근거를 제시하며 Ground Truth보다 더 상세하고 신뢰도 높은 답변을 생성하는지 정성적으로 입증한다.

5.1 Main Results

본 연구가 제안하는 RAR-Agent의 성능을 베이스라인 모델들과 다각도로 비교 및 검증하기 위해 3가지 평가를 진행했다.

5.1.1 Retrieval Performance

Table 2는 KL-RQA 벤치마크에 대한 4개의 모델 검색 성능을 보여준다.

Standard RAG는 NDCG@20이 0.6153, Recall@20은 0.8648로 가장 낮은 성능을 보였다. 이는 Reranker가 없는 Standard RAG(Dense Retrieval)가 질문과 관련된 chunk를 상위권으로 가져오는 능력이 취약함을 명확히 보여준다.

이에 Reranker를 도입한 Dense+Reranker 모델은 NDCG@20이 0.7895로, Standard RAG 대비 28.3% 대폭 향상되었다. 이는 Reranker가 순위 정밀성(Precision) 복원에 핵심적인 역할을 했다는 것을 보여준다. 그러나 Recall@20은 0.9005에 그쳤다.

Hybrid Retrieval을 적용한 Hybrid+Reranker 모델은 Recall@20을 0.9770까지 크게 끌어올렸으며, NDCG@20 역시 0.8389로 Dense+Reranker 대비 6.3% 향상되었다.

Table 2. Retrieval Performance on the KL-RQA Benchmark

	Recall @10	Recall @15	Recall @20	NDCG @10	NDCG @15	NDCG @20
Standard RAG	0.7832	0.8240	0.8648	0.6066	0.6100	0.6153
Dense+ Reranker	0.8929	0.9005	0.9005	0.7922	0.7907	0.7895
Hybrid+ Reranker	0.9566	0.9719	0.9770	0.8496	0.8433	0.8389
Ours	**0.9617**	**0.9770**	**0.9847**	**0.8566**	**0.8572**	**0.8581**

최종적으로, 본 연구가 제안하는 RAR-Agent(Ours)는 Recall@20이 0.9847, NDCG@20은 0.8581을 달성하여, 모든 지표($k=10, 15, 20$)에서 가장 강력한 베이스라인 모델(Hybrid+Reranker)을 일관되게 능가했다. 특히, 다른 베이스라인 모델들은 k가 10에서 20으로 증가할 때 NDCG 점수가 하락하는 경향을 보이지만, RAR-Agent는 0.8566에서 0.8581로 유일하게 성능이 상승했다. 이는 우리가 제안하는 접근법이 가장 우수한 성능을 달성함과 동시에, 검색 결과의 최상위권 및 그 외 순위에서도 정답 chunk를 효과적으로 찾아내었음을 보여준다.

5.1.2 Answer Quality

검색 성능이 LLM의 최종 답변 품질에 미치는 영향을 분석하기 위해 KL-BQA 벤치마크에서 RAGAS를 사용해 Faithfulness, Answer Relevancy, Answer Correctness 총 3가지 지표를 측정했다. 해당 결과는 Table 3에 제시하였다.

Table 3. Evaluation of Answer Quality using RAGAS

	Faithfulness	Answer Relevancy	Answer Correctness
Standard RAG	0.8840	0.4561	0.6488
Dense+ Reranker	0.9167	0.4692	0.6699
Hybrid+ Reranker	0.9328	0.4579	0.6806
Ours	**0.9337**	**0.4831**	**0.6917**

RAG가 고도화될수록, 생성된 답변 품질 또한 전반적으로 일관되게 향상되는 경향이 나타났다. Standard RAG는 모든 지표에서 가장 낮은 점수를 기록했다.

반면, RAR-Agent는 모든 지표에서 가장 높은 점수를 달성했다. 이는 제안하는 접근법이 더 정확하고, 근거 문서에 충실하며, 질문 의도에 부합하는 최종 답변을 생성했다는 것을 정량적으로 보여준다.

5.1.3 Answer Accuracy

최종 답변 정확도를 평가하기 위해 KL-BQA 벤치마크를 사용해 Accuracy와 F1-Score를 측정했다.

Table 4. Answer Accuracy Performance on the KL-BQA Benchmark

	Accuracy	F1-Score
LLM Only	0.6053	0.6295
Standard RAG	0.8665	0.8703
Dense+Reranker	0.8991	0.9012
Hybrid+Reranker	0.9347	0.9371
Ours	**0.9466**	**0.9486**

Table 4에서 볼 수 있듯이, RAG의 도입은 최종 답변 정확도에 결정적인 영향을 미쳤다. LLM Only 모델은 F1-Score가 0.6295로 가장 낮은 성능을 보였다. 이는 LLM의 내부 지식만으로는 법률 도메인의 질문에 정확히 답변할 수 없음을 명확히 보여준다.

RAG의 검색 성능이 향상됨에 따라, 최종 답변 정확도(F1-Score) 또한 0.8703(Standard RAG), 0.9012(Dense+Reranker), 0.9371(Hybrid+Reranker)로 일관되게 상승하였다.

제안하는 RAR-Agent는 F1-Score가 0.9486으로 가장 뛰어난 성능을 달성했다. 이는 베이스라인 모델 중 가장 우수한 성능을 보이는 Hybrid+Reranker 대비 1.23% 향상된 성능이다. Table 4의 결과를 통해, 우리가 제안하는 Agentic RAG Framework인 RAR-Agent가 정확한 최종 답변을 생성할 수 있음을 입증한다.

5.2 Ablation Study

5.2.1 Analysis of Hybrid Retrieval Weights

제안하는 RAR-Agent의 Dense와 Sparse(BM25) 가중치에 따른 성능 변화를 측정했다. Fig. 4는 x축을 Dense:Sparse 가중치 조합으로, y축을 검색 성능(Recall/NDCG)으로 설정하여 9가지 조합의 결과를 시각화한 것이다.

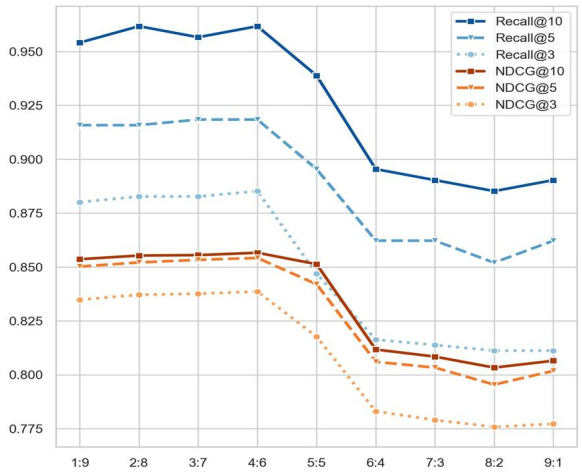

Fig. 4. Impact of Hybrid Retrieval Weights on Performance

결과에서 볼 수 있듯이, Dense의 가중치가 10%(1:9)에서 40%(4:6)로 증가할수록 Recall과 NDCG 점수가 일관되게 상승하다가 40%(4:6)에서 정점을 기록했다. 하지만 Dense의 가중치가 50%(5:5)가 되면서, Recall과 NDCG 점수가 하락세로 돌아섰으며, 60%(6:4) 때 급격한 성능 저하가 관찰되었다.

이는 사실적 정확성이 중요한 법률 도메인에서, Dense Retrieval뿐만 아니라, Sparse Retrieval(BM25)이 여전히 결정적인 역할을 하고 있음을 보여준다. 따라서 본 연구는 가장 안정적이고 우수한 성능을 보인 Dense 40% : BM25 60%를 실험에 사용했다.

5.2.2 Impact of Contextual Scaffolding

3.1.2에서 설명한 바와 같이, RAR-Agent는 Rationale Query 생성 시 CoT를 활용하여 추론 능력을 강화하고, Contextual Scaffolding(CS)을 통해 환각 억제 및 검색에 유효한 풍부한 문맥 생성을 목표로 한다. 이에 따라, 두 가지 핵심 구성 요소가 각각 실제 검색 성능과 Rationale Query 품질에 미치는 영향을 검증하고자 한다.

Retrieval Performance Analysis

Table 5는 CS 및 CoT의 적용 유무에 따른 검색 성능 지표를 비교한다.

CS와 CoT가 모두 적용된 RAR-Agent는 각 구성 요소를 하나씩 제외한 모델들 대비 모든 지표에서 일관되게 가장 우수한 검색 성능을 기록했다. 그리고 CoT를 제외한 'Without CoT' 모델이 CS를 제외한 'Without CS' 모델보다 전반적으로 더 높은 검색 성능을 보인 것을 통해, 우리가 제안하는 CS 접근법이 CoT prompting보다 법률 도메인의 복잡한 쿼리를 검색 친화적인 형태로 변환하는 것에 더 결정적인 역할을 수행했음을 보여준다. 특히 RAR-Agent가 'Without CS' 모델 대비 Recall@15에서 1.86%, NDCG@15에서 0.67% 향상된 결과는, CoT 단독 사용 시의 한계를 CS가 효과적으로 보완하여 최종 검색 성능을 극대화했음을 입증한다.

Table 5. Impact of Contextual Scaffolding on Retrieval Performance

	Recall @10	Recall @15	Recall @20	NDCG @10	NDCG @15	NDCG @20
Without CoT	0.9464	0.9617	0.9745	0.8541	0.8545	0.8562
Without CS	0.9490	0.9592	0.9719	0.8513	0.8515	0.8533
Ours	**0.9617**	**0.9770**	**0.9847**	**0.8566**	**0.8572**	**0.8581**

Rationale Query Analysis

이러한 검색 성능의 차이는 Rationale의 품질, 특히 환각 제어 능력과 원본 질문과 연관된 문맥 생성 능력에서 비롯된다. Fig. 5는 '부동산 점유권'에 대한 동일한 질문에 대해 'Without CS'와 RAR-Agent(With CS)가 생성한 Rationale Query 결과를 보여준다.

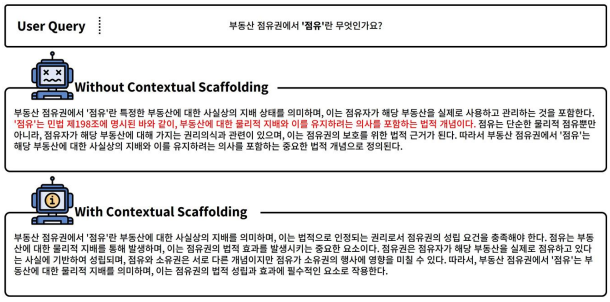

Fig. 5. Efficacy of Contextual Scaffolding in Mitigating Hallucinations during Rationale Generation

Without CS의 경우, Rationale Query 생성 과정에서 '점유'의 정의를 설명하며 '민법 제198조에 명시된 바와 같이...'라는 구체적이지만 부정확한 법 조항을 인용하는 환각(Fig. 5에서 빨간색)을 확인할 수 있다. 이는 검색 단계에서 '민법 제198조'라는 잘못된 키워드를 포함하는 chunk를 우선적으로 탐색하도록 유도하며, 질문의 핵심 의도와 관련 없는 chunk를 상위권에 배치시킨다.

반면, CS를 적용한 RAR-Agent는 환각을 억제하여 '점유'의 핵심 법률 용어인 '사실상의 지배'에 집중한다. 더 나아가, Rationale을 생성하는 과정에서 '점유권의 성립 요건', '법적 효과', 그리고 '점유와 소유권의 구별'과 같이 법률적으로 중요하고 연관성이 높은 핵심 개념들을 Rationale Query에 포함시켰다.

이는 RAR-Agent의 쿼리가 특정 조항(잘못된 조항)에 매몰된 검색이 아닌, '점유'와 관련된 법적 맥락 전반을 탐색할 수 있는 '개념적으로 풍부한 쿼리'의 역할을 수행할 수 있음을 의미한다.

결론적으로, Contextual Scaffolding은 LLM의 환각을 효과적으로 억제하는 동시에, 검색에 필수적인 관련 개념과 풍부한 문맥 정보를 Rationale Query에 포함시켜 최종 검색 성능을 향상시키는 RAR-Agent의 핵심적인 구성 요소임을 보여준다.

5.2.3 Impact of LLM Reasoning Capability on Accuracy

본 연구에서는 최종 답변을 생성하는 LLM의 차이가 정확도에 미치는 영향을 분석하였다. 이를 위해 동일한 조건에서, 'GPT-5-mini'와 'GPT-5'를 사용해 KL-BQA 벤치마크에서 Accuracy와 F1-Score를 측정했다. Table 6은 LLM의 변화에 따른 정확도의 차이를 보여준다.

Table 6. Performance Comparison of Different LLM on the KL-BQA Benchmark

	Accuracy	F1-Score
GPT-4o-mini	0.9466	0.9486
GPT-5-mini	0.9585	0.9598
GPT-5	**0.9792**	**0.9798**

Table 6에 따르면, 상위 LLM을 사용할수록 에이전트의 정확도 또한 비례하여 상승하는 경향을 보였다. Accuracy와 F1-Score가 일관되게 개선되었으며, 이는 RAR-Agent 방법론을 법률 특화형 협업 에이전트(Collaborative Agent)와 결합할 경우, 더욱 뛰어난 성능을 기대할 수 있음을 시사한다.

5.3 Case Study

사용자 질문(User Query)에 대한 실제 사례 분석을 통해, 제안하는 RAR-Agent가 기존 베이스라인 모델(Hybrid+Reranker) 대비 어떻게 더 정확하고 신뢰할 수 있는 답변을 생성하는지 분석한다.

분석에 사용된 질문은 '토지 소유자가 사용·수익권을 채권적으로 포기한 경우, 토지에 대한 반환청구가 배척되려면 그 포기는 영구적인 것이어야 하나요?'이다. 이는 명확한 법률적 근거를 요구하는 '예/아니요' 질문이다. 해당 질문으로부터 RAR-Agent가 생성한 Analytical Query와 Rationale Query는 Fig. 6에 명시되어 있다.

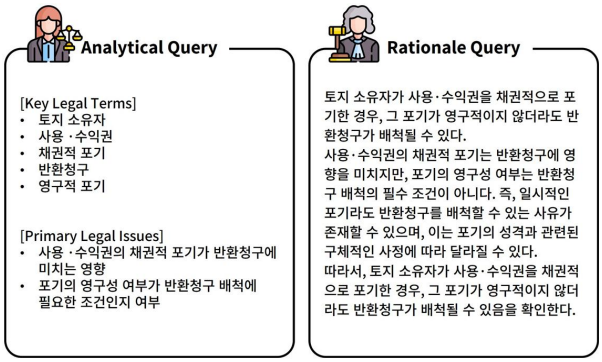

Fig. 6. Example of an Analytical Query and a Rationale Query Generated by RAR-Agent

Fig. 7을 보면, Hybrid+Reranker 모델은 '아니요. ... 영구적일 필요는 없다.'라고 답변했다. 이는 GT와 정반대되는, 사실적으로 부정확한 답변이다. 이를 통해 관련성이 낮거나 잘못된 chunk를 검색하면, LLM이 부정확한 답변을 생성한다는 것을 확인할 수 있다.

반면, RAR-Agent는 '예. ... 영구적인 것이어야 반환청구가 배척될 수 있습니다.'라고 정확하게 답변했다.

더 나아가, RAR-Agent의 답변은 GT와 비교했을 때 단순한 긍정/부정 판단을 넘어, '판례 1(사건 번호: 2009다228)'이라는 구체적인 근거를 제시하며 답변의 신뢰도를 높였다. GT가 '그렇습니다'라는 결론과 부연 설명만 제시하는 것과 달리, RAR-Agent는 정확한 결론과 핵심적인 법적 근거를 함께 제공하여 GT보다 더 상세하고 유용한 정보를 생성함을 확인했다.

결론적으로, 이 사례는 제안한 RAR-Agent의 Query Expansion 및 2중 Filtering Mechanism이 복잡한 법률 QA에 대해 정확한 근거 문서를 검색하고, 최종적으로 사실에 입각한 상세한 답변을 생성함을 명확히 보여준다.

VI. Conclusion

본 연구는 사실적 정확성이 필수적인 법률 도메인에서, 기존 RAG가 어휘적 불일치로 인해 겪는 검색 실패 문제를 해결하고자 했다. 이를 위해, 우리는 LLM의 추론 능력을 검색 단계에서 활용하는 Rationale-Based Query Formulation에 기반한 Agentic RAG Framework인 RAR-Agent를 제안했다.

제안한 RAR-Agent의 성능을 검증하기 위해, 4개의 베이스라인 모델과 검색 성능, 답변 품질, 답변 정확도라는 3가지 범주로 나누어 포괄적인 비교 평가를 수행했다. 그

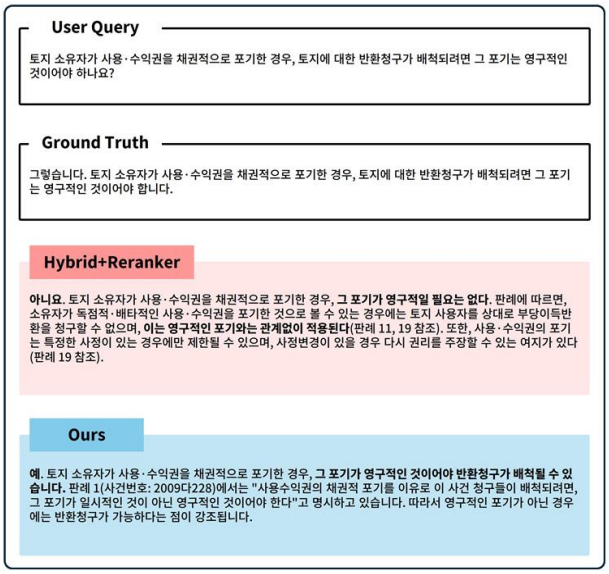

Fig. 7. Comparison of Answers Generated by Hybrid+Reranker and RAR-Agent

결과, RAR-Agent는 모든 평가지표에서 가장 우수한 성능을 달성했다.

본 연구의 핵심 기여는 다음과 같다. 첫째, 별도의 학습 없이 CoT prompting으로 Query Expansion을 수행하여 법률 도메인에서의 어휘적 불일치 문제를 해결하고 우수한 검색 성능을 달성하였다. 둘째, RAGAS가 측정하지 못하는 사실적 정확성을 평가하기 위한 벤치마크(KL-BQA)를 구축하여 RAR-Agent의 실질적인 우수성을 입증하였다. 마지막으로, 환각을 억제하기 위해 두 단계로 이루어진 Rationale 생성과 2중 Filtering Mechanism을 결합한 RAR-Agent를 제안했다.

VII. Limitations

RAR-Agent는 4개의 베이스라인 모델 대비 우수한 성능을 보여주었지만, 다음과 같은 한계점을 갖는다.

첫째, RAG 파이프라인의 핵심 구성 요소가 특정 모델로 고정된다. 본 연구는 임베딩 모델로 text-embedding-3-large를, Reranker로 jina-reranker-v2-base-multilingual을 사용하여 실험을 진행한다. 제안하는 RAR-Agent는 이들 구성 요소의 성능에 의존하는 특성을 갖는다. 따라서 향후 더 강력한 성능의 임베딩 모델 또는 Reranker 모델을 적용할 경우, RAR-Agent는 추가적인 성능 향상의 잠재력을 갖는다.

둘째, 일반 도메인에 대한 일반화 가능성이 검증되지 않았다. 본 연구의 실험은 '한국어 법률'이라는 매우 특화되고 전문화된 도메인에 한정되어 수행된다. RAR-Agent가 보여준 성능 향상이 일반 도메인의 태스크에서도 동일하게 재현될 수 있는지에 대해서는 추가적인 검증이 필요하다.

셋째, 단일 질의응답(single-turn) 환경에서만 성능을 검증한다. 본 연구의 평가는 정적인 QA 벤치마크를 기반으로 하며, 이는 실제 법률 상담 등에서 발생할 수 있는 연속적인 대화(multi-turn conversation)의 동적인 특성을 반영하지 못한다. 사용자의 후속 질의나 대화의 맥락을 고려해야 하는 multi-turn 환경에서도 제안하는 에이전트가 우수한 성능을 유지할 수 있는지 확인이 필요하다.

VIII. Future Work

본 연구는 오정보 방지와 정확성이 필수적인 법률 도메인의 특성을 고려하여 다중 쿼리 생성 및 Reranking을 통해 환각을 효과적으로 억제하여 정확도를 제고하였다.

그러나 이 과정에서 수반되는 연산 비용 및 지연 시간의 증가는 여전히 극복해야 할 과제로 남아있다. 이에 추후 연구에서는, 정확성과 효율성 간의 trade-off를 정량적으로 분석하고, 이를 바탕으로 성능 저하를 최소화하는 비용 효율적 최적화 방법론을 규명하고자 한다.

또한 본 연구에서 사용하였던 337개의 QA 쌍으로 구성된 KL-BQA 벤치마크의 규모를 대폭 확장하고, 추가로 기존 영문 법률 벤치마크와의 비교 실험을 통하여 해당 방법론의 강건성과 일반화 가능성을 검증하고자 한다.

REFERENCES

[1] Achiam, Josh, et al, "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023. DOI: 10.48550/arXiv.2303.08774

[2] Wang, Shansong, et al, "Capabilities of gpt-5 on multimodal medical reasoning," arXiv preprint arXiv:2508.08224, 2025. DOI: 10.48550/arXiv.2508.08224

[3] Savelka, Jaromir, "Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts," Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, 2023. DOI: 10.1145/3594536.3595161

[4] Huang, Lei, et al, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," ACM Transactions on Information Systems 43.2, 1-55, 2025. DOI: 10.1145/3703155

[5] Vu, Tu, et al, "Freshllms: Refreshing large language models with

search engine augmentation," arXiv preprint arXiv:2310.03214, 2023. DOI: 10.48550/arXiv.2310.03214

[6] Lewis, Patrick, et al, "Retrieval-augmented generation for knowledge-intensive nlp tasks," Advances in neural information processing systems 33, 9459-9474, 2020. DOI: 10.48550/arXiv.2005.11401

[7] Asai, Akari, et al, "Self-rag: Learning to retrieve, generate, and critique through self-reflection," 2024. DOI: 10.48550/arXiv.2310.11511

[8] Lazaridou, Angeliki, et al, "Internet-augmented language models through few-shot prompting for open-domain question answering," arXiv preprint arXiv:2203.05115, 2022. DOI: 10.48550/arXiv.2203.05115

[9] Yan, Shi-Qi, et al, "Corrective retrieval augmented generation," 2024. DOI: 10.48550/arXiv.2401.15884

[10] Kim, Daehee, et al, "GuRE: Generative Query REwriter for Legal Passage Retrieval," arXiv preprint arXiv:2505.12950, 2025. DOI: 10.48550/arXiv.2505.12950

[11] Li, Haitao, et al, "Delta: Pre-train a discriminative encoder for legal case retrieval via structural word alignment," Proceedings of the AAAI Conference on Artificial Intelligence, 2025. DOI: 10.48550/arXiv.2403.18435

[12] Li, Haitao, et al, "SAILER: structure-aware pre-trained language model for legal case retrieval," Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023. DOI: 10.48550/arXiv.2304.11370

[13] Mo, Fengran, et al, "CHIQ: Contextual history enhancement for improving query rewriting in conversational search," arXiv preprint arXiv:2406.05013, 2024. DOI: 10.48550/arXiv.2406.05013

[14] Qian, Hongjin, and Zhicheng Dou, "Explicit query rewriting for conversational dense retrieval," Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022. DOI: 10.18653/v1/2022.emnlp-main.311

[15] Mo, Fengran, et al, "Convgqr: Generative query reformulation for conversational search," arXiv preprint arXiv:2305.15645, 2023. DOI: 10.48550/arXiv.2305.15645

[16] Wang, Yujing, et al, "MaFeRw: Query rewriting with multi-aspect feedbacks for retrieval-augmented large language models," Proceedings of the AAAI Conference on Artificial Intelligence, 2025. DOI: 10.1609/aaai.v39i24.34732

[17] Chen, Zhiyu, et al, "Reinforced question rewriting for conversational question answering," arXiv preprint arXiv:2210.15777, 2022. DOI: 10.48550/arXiv.2210.15777

[18] Wu, Zeqiu, et al, "Conqrr: Conversational query rewriting for retrieval with reinforcement learning," arXiv preprint arXiv:2112.08558, 2021. DOI: 10.48550/arXiv.2112.08558

[19] Ma, Xinbei, et al, "Query rewriting in retrieval-augmented large language models," The 2023 Conference on Empirical Methods in Natural Language Processing, 2023. DOI: 10.48550/arXiv.2305.14283

[20] Ye, Fanghua, et al, "Enhancing Conversational Search: Large Language Model-Aided Informative Query Rewriting," arXiv preprint arXiv:2310.09716, 2023. DOI: 10.48550/arXiv.2310.09716

[21] Wang, Liang, et al, "Query2doc: Query Expansion with Large Language Models," Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 9414-9423, 2023. DOI: 10.18653/v1/2023.emnlp-main.585

[22] Mao, Kelong, et al, "Large Language Models Know Your Contextual Search Intent: A Prompting Framework for Conversational Search," Findings of the Association for Computational Linguistics: EMNLP 2023, 1211-1225, 2023. DOI: 10.18653/v1/2023.findings-emnlp.86

[23] Guha, Neel, et al, "LEGALBENCH: A COLLABORATIVELY BUILT BENCHMARK FOR MEASURING LEGAL REASONING IN LARGE LANGUAGE MODELS," arXiv preprint arXiv:2308.11462, 2023. DOI: 10.48550/arXiv.2308.11462

[24] Pipitone, Nicholas, and Ghita Houir Alami, "LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain," arXiv preprint arXiv:2408.10343, 2024. DOI: 10.48550/arXiv.2408.10343

[25] Es, Shahul, et al, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 150-158, 2024. DOI: 10.18653/v1/2024.eacl-demo.16

[26] Wei, Jason, et al, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," Advances in Neural Information Processing Systems 35, 24824-24837, 2022. DOI: 10.48550/arXiv.2201.11903

[27] Jagerman, Rolf, et al, "Query Expansion by Prompting Large Language Models," arXiv preprint arXiv:2305.03653, 2023. DOI: 10.48550/arXiv.2305.03653

[28] Gao, Yunfan, et al, "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv preprint arXiv:2312.10997, 2023. DOI: 10.48550/arXiv.2312.10997

[29] Guu, Kelvin, et al, "REALM: Retrieval-Augmented Language Model Pre-Training," Proceedings of the 37th International Conference on Machine Learning (PMLR 119), 3929-3940, 2020. DOI: 10.48550/arXiv.2002.08909

[30] Jeong, Soyeong, et al, "Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity," arXiv preprint arXiv:2403.14403, 2024. DOI: 10.48550/arXiv.2403.14403

[31] Kim, Gangwoo, et al, "Learn to Resolve Conversational

Dependency: A Consistency Training Framework for Conversational Question Answering," arXiv preprint arXiv:2106.11575, 2021. DOI: 10.48550/arXiv.2106.11575

[32] Li, Minghan, et al, "Can Query Expansion Improve Generalization of Strong Cross-Encoder Rankers?," arXiv preprint arXiv:2311.09175, 2023. DOI: 10.48550/arXiv.2311.09175

[33] Sohn, Jiwoong, et al, "Rationale-Guided Retrieval Augmented Generation for Medical Question Answering," arXiv preprint arXiv:2411.00300, 2024. DOI: 10.48550/arXiv.2411.00300

[34] Li, Yangning, et al, "Towards Agentic RAG with Deep Reasoning: A Survey of RAG-Reasoning Systems in LLMs," arXiv preprint arXiv:2507.09477, 2025. DOI: 10.48550/arXiv.2507.09477

[35] Cormack, Gordon V., Charles L. A. Clarke, and Stefan Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 758-759, 2009. DOI: 10.1145/1571941.1572114

[36] Wang, Liang, et al, "Improving Text Embeddings with Large Language Models," arXiv preprint arXiv:2401.00368, 2024. DOI: 10.48550/arXiv.2401.00368

[37] Robertson, S. E., et al, "Okapi at TREC-3," Overview of the Third Text REtrieval Conference (TREC-3), 109-126, 1995.

Authors

Gyuhyeong Kim received the B.S. degree in Computer Engineering from Baekseok University, Cheonan, Korea, in 2025. He is currently working as an AI Researcher at QI.

Yunhyeok Do reveived the B.S. degree in Business Administration and Computer Engineering from Sogang University, Seoul, Korea, in 2025. He is co-founder of QI.

Joonhyeon Song received the B.S. degree in Data Science from Kwangwoon University, Seoul, Korea, in 2025. He is co-founder of QI.

Ziyang Liu received the B.A. degree in Management from the Army Command College of Shijiazhuang, China PLA, China, in 2006, and the M.A. and Ph.D. degrees in Management from Kyonggi University Korea, in 2010 and 2013, respectively. Dr. Liu joined the faculty of Global Business at Kyonggi University, Korea, in 2015. He is currently a Professor of Global Business at Kyonggi University. His research interests include quality management, management information systems, international economics, and e-business.