JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 11, pp. 131-149, November 2026
https://doi.org/10.9708/jksci.2026.31.11.131

A Staged Hybrid Retrieval and Evidence-Aware Extractive QA Framework for Fair Trade Commission Decisions

Tae-Kyu Hwang*, Namgyu Kim**

*Ph.D. Candidate, Graduate School of Business IT, Kookmin University, Seoul, Korea
**Professor, Graduate School of Business IT, Kookmin University, Seoul, Korea

[Abstract]

Fair Trade Commission decisions contain complex legal and administrative structures, including case summaries, respondents, market definitions, violations, applicable provisions, legal determinations, and sanctions. This study proposes a staged hybrid retrieval and evidence-aware extractive QA framework that combines M1 Baseline Hybrid Retrieval, M2 Section-Aware Retrieval, M3 Legal Metadata Retrieval, M4 Graph-Enhanced Retrieval, and M5 Agentic Retrieval Control. Answers are constructed deterministically by selecting relevant passages from retrieved evidence, without a generation LLM. M5 is a rule-based controller for query analysis, route selection, evidence evaluation, and limited fallback retrieval rather than a learning-based multi-agent system. With evaluation-only annotations excluded from runtime retrieval and parameters frozen after development, the framework was evaluated on 144 Internal-test questions and 720 document-disjoint External-test questions. Recall@10 improved from 0.7380 to 0.8708 on Internal-test and from 0.6861 to 0.8493 on External-test. M4 did not improve general strict ranking over M3, but evaluation on similar-case questions showed significant gains in target recall and path coverage. These results support relationship and path evidence recovery rather than general ranking gains or multi-hop reasoning accuracy. Post-hoc LLM-as-a-Judge evaluation was complemented by claim-evidence entailment for all 864 M5 answers, with a mean score of 0.9290.

▸ **Key words**: Legal Extractive QA, Graph-Enhanced Retrieval, Hybrid Retrieval, Agentic Retrieval Control, Leakage-safe Evaluation

• First Author: Tae-Kyu Hwang, Corresponding Author: Namgyu Kim
*Tae-Kyu Hwang (andy4627@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
**Namgyu Kim (ngkim@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
• Received: 2026. 08. 06, Revised: 2026. 10. 20, Accepted: 2026. 10. 23.

Copyright © 2026 The Korea Society of Computer and Information
http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945
www.kci.go.kr

[요 약]

공정거래위원회 의결서는 사건개요, 피심인, 시장구조, 위반행위, 적용 법조항, 위법성 판단 및 제재 등 복합적인 법률·행정 구조를 포함한다. 본 연구는 이러한 문서에서 근거 회수 성능과 답변의 근거 대응성을 높이기 위해 M1 Baseline Hybrid Retrieval, M2 Section-aware Retrieval, M3 Legal Metadata Retrieval, M4 Graph-enhanced Retrieval, M5 Agentic Retrieval Control로 구성된 단계적 하이브리드 검색 및 근거 기반 추출형 질의응답 프레임워크를 제안한다. 답변은 검색된 근거에서 관련 구문을 선택하여 구성하며 별도의 생성형 LLM을 사용하지 않는다. M5는 학습형 다중 에이전트가 아니라 질문 분석, 경로 선택, 근거 평가 및 제한적 재검색을 수행하는 규칙 기반 제어기이다. 평가정보를 runtime 검색에서 제외하고 개발세트에서 정한 파라미터를 고정한 뒤 Internal-test 144문항과 External-test 720문항을 평가하였다. 외부 평가 문서는 내부 corpus와 중복되지 않는다. Recall@10은 Internal-test에서 0.7380→0.8708, External-test에서 0.6861→0.8493으로 향상되었다. 일반 strict 검색에서 M4는 M3와 동일한 순위지표를 보였으나 유사사건 질문군의 Graph Gold 평가에서는 target recall과 path coverage가 유의하게 증가하였다. 이는 관계, 경로 근거 회수의 개선이며 일반 검색 순위 향상이나 multi-hop 추론 정확도의 증거로 확대하지 않는다. 전체 M5 답변 864건의 claim-evidence entailment 평균은 0.9290이었다.

▸ **주제어**: 법률 추출형 질의응답, Graph-enhanced Retrieval, Hybrid Retrieval, Agentic Retrieval Control, Leakage-safe 평가

I. Introduction

대규모 언어모델은 자연어 질의응답, 문서 요약, 정보추출 및 의사결정 지원 등 지식집약적 업무에 활용되고 있다. 법률·행정 분야에서도 방대한 문서에서 사실관계, 적용 법령, 판단 근거 및 처분결과를 찾아 설명하는 데 관심이 높다[1]. 그러나 학습에 포함되지 않은 정보나 기관 내부 전문 문서를 다룰 때 부정확하거나 근거 없는 내용을 생성하는 환각(hallucination)이 발생할 수 있다[2]. 따라서 법률 답변은 문장의 자연스러움뿐 아니라 사실과 판단이 원문에 의해 뒷받침되는지, 사건 시점과 적용 조문이 일치하는지, 인용 근거가 실제 주장과 연결되는지를 확인해야 한다[3].

공정거래위원회 의결서는 사건 배경, 피심인, 거래구조, 관련시장, 위반행위, 적용 법령, 위법성 판단 및 처분 등으로 구성된다[4]. 하나의 질문에 답하려면 여러 영역의 근거를 함께 검토해야 하는 경우가 많다. 예를 들어 거래상지위 남용의 판단 이유를 설명하려면 피심인의 지위, 상대방의 거래의존성, 불이익 제공행위와 적용 조문을 연결해야 한다. 반면 과징금액을 묻는 질문에서는 주문과 과징금 산정 근거를 우선 확인해야 한다. 이처럼 질문에 따라 필요한 섹션과 근거가 달라지므로 이를 반영한 검색 전략이 필요하다.

검색증강생성(Retrieval-Augmented Generation, RAG)은 외부 저장소에서 관련 자료를 검색하고 그 문맥을 생성모델에 제공하는 방식이다[5]. 실제 문서를 참조하여 답변의 사실성과 출처 추적성을 높일 수 있지만, 필요한 근거를 정확하고 충분하게 검색하는 것이 전제되어야 한다. 일반적으로 사용하는 고정 길이 청킹과 BM25,Dense Retrieval[6]만으로는 의결서의 구조와 논리적 연결을 충분히 반영하기 어렵다. 하나의 판단 근거가 여러 Chunk로 나뉘거나 같은 피심인명,법령명이 문서 전반에 반복되면 질문에 필요한 구간을 구별하기 어렵기 때문이다. 사건번호, 조문번호, 위반유형 등의 정보가 본문에만 포함되어 있을 때도 검색 범위 제한과 재순위화에 한계가 있다.

이러한 문제를 보완하기 위해 Graph RAG[7]와 Multi-Agent RAG[8, 9]가 제안되었다. Graph RAG는 사건, 기업, 법조항 등의 관계를 활용하여 문서 간 근거를 연결하고, Multi-Agent RAG는 질의 분해와 반복 검색으로 부족한 근거를 보완한다. 다만 그래프 관계가 원문 근거와 연결되지 않거나 청킹,법률 메타데이터가 부정확하면 효과가 제한될 수 있다. 따라서 의결서 질의응답에서는 문서 구조, 법률 메타데이터, 그래프 관계와 검색 제어를 함께 고려해야 한다.

본 연구는 이들 요소를 결합한 단계적 하이브리드 검색 및 근거 기반 추출형 질의응답 프레임워크를 제안한다. Baseline Hybrid Retrieval(M1)을 시작으로

www.kci.go.kr

Section-aware Retrieval(M2), Legal Metadata Retrieval(M3), Graph-enhanced Retrieval(M4), Agentic Retrieval Control(M5)을 순차적으로 적용한다. 답변은 검색된 근거의 관련 구문을 선택하여 구성하며 별도의 생성형 LLM을 사용하지 않는다. M5는 검색 실패를 탐지하고 제한적으로 재검색하는 단일 규칙 기반 제어기이며, 단계별 검색 성능과 근거 확보 수준을 구분하여 평가한다.

이하 2장에서는 관련 연구를, 3장에서는 제안 방법과 단계별 구성을 설명한다. 4장에서는 실험 환경, 평가 방법 및 결과를 제시하고, 5장에서는 결론과 후속 연구 과제를 정리한다.

II. Preliminaries

1. Retrieval-Augmented Generation and Hybrid Retrieval

검색증강생성(Retrieval-Augmented Generation, RAG)은 외부 문서 저장소에서 질의와 관련된 정보를 검색하고, 검색된 문맥을 대규모 언어모델의 답변 생성에 활용하는 방법이다[5]. 외부 문서를 생성의 근거로 활용함으로써 모델 내부의 매개변수 지식에만 의존하는 방식보다 답변의 사실성과 출처 추적 가능성을 높일 수 있다[5]. 그러나 RAG의 최종 성능은 검색 단계에서 질의에 적합한 근거를 얼마나 정확하고 충분하게 확보하는지에 크게 의존한다[6].

RAG의 검색 단계에는 희소 검색(Sparse Retrieval), 밀집 검색(Dense Retrieval) 또는 둘을 결합한 하이브리드 검색(Hybrid Retrieval)이 사용된다. BM25와 같은 희소 검색은 단어 일치로 관련성을 계산하여 사건번호, 법률명, 조문번호 검색에 효과적이다[10]. 밀집 검색은 질의와 문서의 임베딩을 비교하므로 표현이 달라도 의미적으로 유사한 문서를 찾을 수 있다[11]. 하이브리드 검색은 점수 융합이나 순위 통합으로 두 방식의 장점을 결합한다[6, 12]. 다만 문서를 나누는 단위와 검색에 사용하는 정보에 따라 성능이 달라진다.

법률정보 검색에서는 단순한 어휘 일치뿐만 아니라 사건의 사실관계, 법적 쟁점 및 적용 법조항 사이의 의미적 관련성까지 함께 고려해야 한다[13]. 기존 연구는 사건 관련 행위정보나 답변 근거 구간을 활용하여 일반적인 검색보다 향상된 검색 성능을 보고하였다[3, 14]. 그러나 일반적인 하이브리드 검색만으로는 법률문서의 계층구조와 논증 구조를 충분히 반영하기 어렵고, 질문에 필요한 근거가 여러 섹션에 분산된 경우에는 관련 정보를 효과적으로 결합하기 어렵다[15]. 따라서 법률 RAG에서는 검색 방식뿐 아니라 문서 구조와 법률지식을 함께 활용하는 구조 및 지식 기반 검색이 요구된다.

2. Structure- and Legal Knowledge-Aware Retrieval

법률문서는 계층구조와 논증 순서가 명확하며, 각 문장은 사건 개요, 사실관계, 적용 법령, 법적 판단 및 처분과 같은 기능적 역할을 갖는다[15, 16]. 따라서 문장의 의미는 소속 섹션과 문서 전체의 맥락에서 해석해야 한다. 이를 반영하는 구조 인식 청킹은 제목,문단,조문 경계를 고려하고, 구조 인식 검색은 질문 유형에 필요한 섹션을 우선 탐색한다[15, 16]. 이러한 방법은 문장의 의미와 구조적 맥락을 함께 활용한다.

문서 구조와 함께, 법률문서에 포함된 다양한 메타데이터와 법률지식 역시 검색 정확도를 향상시키는 중요한 요소이다. 사건번호, 법률명, 조문번호, 당사자 및 사건유형과 같은 메타데이터와 법률 개체는 검색 범위를 제한하거나 검색 결과를 재순위화하는 데 활용될 수 있으며 [13, 14, 17], 온톨로지와 같은 법률지식은 일상적인 질의를 전문적인 법률 개념으로 연결하고 법률 개체 간의 관계를 활용하여 질의 확장, 개념 정규화 및 관계 기반 검색을 가능하게 한다 [18, 19]. 즉, 구조 정보가 문서의 조직 방식을 반영한다면, 법률지식은 문서 간 의미적 관계를 보완하여 보다 정교한 검색을 지원한다.

구조,지식 기반 검색의 성능은 문서구조 분석, 메타데이터 추출, 개체명 인식 및 온톨로지 구축의 정확도에 의존한다[13, 17]. 근거가 여러 문서나 개체 관계에 분산되면 구조 정보와 법률지식을 개별적으로 활용하는 것만으로는 충분하지 않다[13]. 이를 보완하기 위해 구조화된 지식을 그래프로 표현하는 Graph RAG와 검색 과정을 동적으로 조정하는 Agentic Retrieval이 제안되고 있다[19, 20].

3. Graph RAG and Agentic Retrieval

앞서 살펴본 구조 및 지식 기반 검색을 더욱 발전시킨 대표적인 접근법이 Graph RAG이다. Graph RAG는 검색된 문서를 단순한 텍스트 집합이 아니라 개체와 관계로 구성된 그래프 형태로 표현하여 검색과 추론을 수행한다[19]. 기존 RAG가 개별 문서나 Chunk 단위의 검색에 의존하는 것과 달리, Graph RAG는 개체 간 관계를 따라 다단계(Multi-hop) 탐색을 수행할 수 있어 여러 문서에 분산된 정보를 효과적으로 연결할 수 있다. 법률 분야에서는 법률 개체, 법령, 판례 및 이들 간의 관계를 그래프로 구성하여 복잡한 법률 질의에 필요한 근거를 보다 효과적으로

www.kci.go.kr

검색하는 연구가 이루어지고 있다 [19, 20].

한편, Graph RAG가 구조화된 법률지식을 효과적으로 활용하는 데 초점을 둔다면, Agentic Retrieval은 검색 과정 자체를 동적으로 개선하는 데 초점을 둔다. Agentic Retrieval은 검색을 단일 단계로 수행하는 대신 계획(Planning), 검색(Retrieval), 평가(Evaluation), 질의 수정(Query Refinement) 및 재검색을 반복하여 필요한 근거를 점진적으로 보완하는 접근법이다 [21, 22]. 또한 자기 개선(Self-Reflection)과 다중 에이전트 협업을 통해 검색 결과의 정확성과 완전성을 지속적으로 향상시키는 연구도 활발히 이루어지고 있다 [21-23].

Graph 관계와 반복 검색을 함께 활용하려면 문서 구조, 법률 메타데이터, 근거 평가 및 재검색 절차를 연결해야 한다. 본 연구는 Graph 기반 근거 확장과 규칙 기반 검색 제어를 결합하여 질문에 필요한 근거를 확보하고 부족한 경우 제한적으로 재검색하도록 구성한다.

III. The Proposed Scheme

1. Overall Research Process

본 연구는 의결서의 문서 구조, 법률 메타데이터, 지식 그래프 및 규칙 기반 검색 제어를 단계적으로 결합한다. 이를 통해 여러 섹션에 분산된 사실관계와 법적 판단의 근거를 함께 검색하고, 근거가 부족하면 제한적으로 재검색한다. M1은 BM25와 Dense Retrieval을 결합한 내부 기준선이다. M1~M5의 비교는 같은 시스템에 기능을 순차적으로 추가했을 때의 변화를 평가하며, 외부 알고리즘이나 대표적인 비교방법과의 성능 비교를 대신하지 않는다.

전체 과정은 PDF 파싱,품질관리, 섹션 구조화, 법률 메타데이터 추출,정규화, Hybrid,Graph 검색, 검색 제어, 추출형 답변 구성 및 근거 검증으로 이루어진다(Fig. 1). 단계별 정보는 후속 처리에 활용하며, 문서와 Chunk 식별자를 통해 원문 근거를 추적한다.

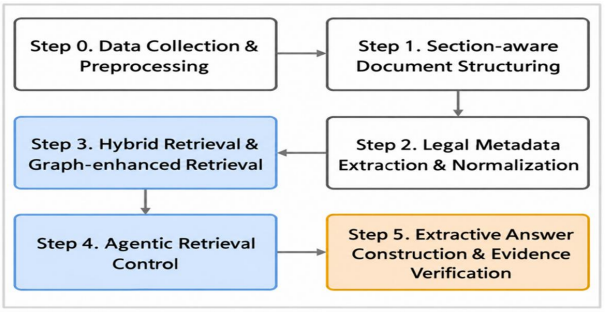

Fig. 1. The Entire Research Procedure of the Proposed Evidence-Aware Extractive QA Framework

전처리 단계(Step 0)에서는 텍스트,스캔 PDF를 구분하고 필요한 경우 OCR을 수행하였다. 이후 반복 머리말,꼬리말을 제거하고 문단을 복원하며 표,별지를 분리하였다(Fig. 2). 품질 검증을 통과한 문서만 인덱스에 저장하고, 오류 문서는 재처리하거나 수동 검수하였다.



Fig. 2. Decision PDF Parsing and Quality Control

정제 문서에 Table 1의 기능을 단계적으로 추가하여 M1~M5를 구성하였다. M2는 질문에 필요한 섹션의 누락을, M3는 법률 개체와 조문의 불일치를 보완한다. M4는 사건 간 관계 근거를 확장하고, M5는 근거 충분성을 평가하여 필요한 경우 검색 전략을 조정한다.

Table 1. Functions by Stage (M1-M5)

Step	Feature	Main Role
M1	Hybrid Retrieval(BM25+Dense)	Baseline hybrid search
M2	Section-aware Chunking	Structure-aware retrieval
M3	Legal Metadata	Legal entity-based retrieval
M4	Knowledge Graph	Relationship-aware evidence retrieval
M5	Agentic Retrieval Control	Evidence verification and adaptive re-retrieval

이러한 단계적 구성은 각 검색 요소의 기여도를 비교할 수 있도록 하며, 다음 절부터 각 처리 단계의 세부 방법을 설명한다.

2. Section-aware Document Structuring

고정 길이 청킹은 서로 다른 법률적 의미를 섞거나 하나의 의미 단위를 분리하여 문맥을 끊을 수 있다. Section-aware Document Structuring은 의결서의 계층구조와 섹션 경계를 우선 보존하여 각 Chunk의 법률적 문맥을 유지한다(Fig. 3).

www.kci.go.kr

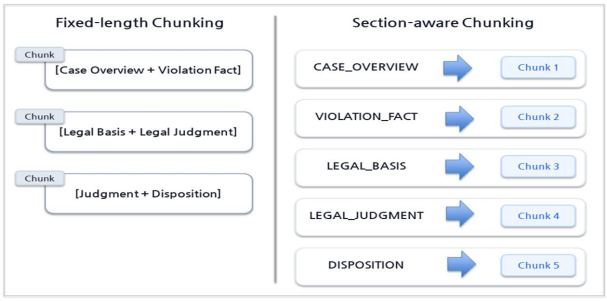

Fig. 3. Comparison of Fixed-length and Section-aware Chunking

제목 패턴, 번호 체계, 핵심어 및 주변 문맥으로 문단을 표준 섹션에 분류하였다(Table 2). 최대 길이를 초과한 섹션만 문단 단위로 나누고 인접 Chunk에 제한적인 중첩을 적용하였다. 각 Chunk에는 섹션 유형,순서, 페이지 및 이전,다음 Chunk 정보를 저장하였다.

Table 2. Standard Section Types for FTC Resolutions

Section Type	Definition
CASE_OVERVIEW	Summarizes the case background and key issues.
RESPONDENT	Identifies the respondent and business information.
MARKET_ DEFINITION	Defines the relevant market and its scope.
FACT_RELATION	Describes the factual background and event sequence.
VIOLATION_FACT	Presents facts supporting the alleged violation.
LEGAL_BASIS	Specifies the applicable legal provisions.
LEGAL_JUDGMENT	Explains the legal reasoning and judgment.
PENALTY_ CALCULATION	Describes the basis for penalty calculation.
DISPOSITION	Final disposition
APPENDIX	Appendix and attachments

M2에서는 (식 1)과 같이 질문 유형과 섹션의 적합도를 산출하여 기존 Hybrid Retrieval 점수에 추가한다. 여기서 q는 사용자 질의(query), d는 검색 대상 Chunk, $S_{M1}(q,d)$는 M1 단계에서 계산된 Hybrid Retrieval 점수, $S_{section(q,d)}$는 질의 유형과 문서 섹션 간의 적합도, γ는 섹션 적합도가 최종 검색 점수에 미치는 영향을 조절하는 가중치이다.

$$S_{M2}(q,d)=S_{M1}(q,d)+\gamma S_{section}(q,d)$$
$$S_{section}(q,d)\in[0,1] \qquad (식\ 1)$$

$S_{section}(q,d)$는 질문 유형과 표준 섹션의 대응에 따라 계산한 0~1 범위의 적합도이다. 예를 들어 적용 법령을 묻는 질문에서는 LEGAL_BASIS를 우선 검색한다. 질문 유형별 대응 섹션은 Table 3과 같다. M2에서 생성한 섹션 정보는 후속 법률 메타데이터 추출과 그래프 구축에도 공통으로 사용한다.

Table 3. Question Types and Preferred Sections

Question Type	Preferred Sections
Case and Respondent Identification	CASE_OVERVIEW, RESPONDENT, TITLE_PAGE, DISPOSITION
Violation Fact Identification	VIOLATION_FACT, FACT_RELATION, RESPONDENT, CASE_OVERVIEW
Applicable Law Identification	LEGAL_BASIS, LEGAL_JUDGMENT, VIOLATION_FACT, TITLE_PAGE, DISPOSITION
Legal Judgment and Reasoning	LEGAL_JUDGMENT, LEGAL_BASIS, VIOLATION_FACT
Sanction and Penalty Identification	PENALTY_CALCULATION, SANCTION_BASIS, CORRECTIVE_ORDER, DISPOSITION
Similar Case Retrieval	VIOLATION_FACT, LEGAL_JUDGMENT, MARKET_DEFINITION
Temporal and Statutory Amendment Comparison	LEGAL_BASIS, CASE_OVERVIEW, FACT_RELATION, LEGAL_JUDGMENT
Complex Legal Query	VIOLATION_FACT, LEGAL_BASIS, LEGAL_JUDGMENT, PENALTY_CALCULATION, DISPOSITION

3. Legal Metadata Extraction and Normalization

법률 질의에 필요한 사건번호, 피심인, 법령명 및 조문을 문서 Chunk에서 추출하고 정규화하였다. Fig. 4는 이 메타데이터 처리 과정을 나타낸다.

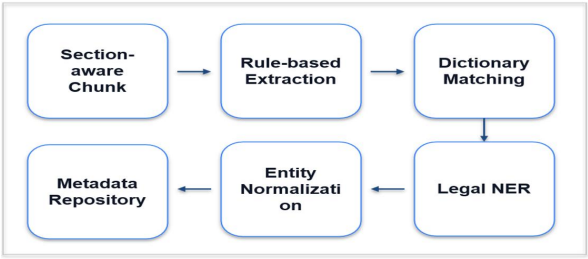

Fig. 4. Legal Metadata Extraction Process

사건번호,조문번호,금액처럼 표기 형식이 일정한 정보는 규칙으로 추출하고, 법령명과 위반유형은 공식 사전과 매칭하였다. 기업명, 시장명 및 복합 위반행위에는 문서 표제,당사자 섹션,사전,일반화 규칙을 사용하였다. Runtime에서는 LLM NER을 사용하지 않았다.

기업명, 법령명 및 조문번호를 표준 식별자로 정규화하여 표현이 달라도 같은 개체와 근거를 연결하도록 하였다. 메타데이터 스키마는 Table 4와 같다.

www.kci.go.kr

Table 4. Legal Metadata Schema

Category	Metadata	Description
Case	Case ID, Decision ID	Unique identifiers for the case and decision.
Respondent	Company Name	Name of the respondent company or organization.
Law	Law Name, Article	Applicable law and relevant article(s).
Violation	Violation Type	Type of legal violation identified in the case.
Market	Market	Relevant market or business sector.
Sanction	Sanction Type, Penalty	Sanctions, including corrective measures and penalties.

M3에서는 M2의 검색 점수에 질의와 Chunk 사이의 엔터티, 법령 일치도를 더한다(식 2).

$$S_{M3}(q,d)=S_{M2}(q,d)+\delta S_{entity}(q,d)+\varepsilon S_{law}(q,d) \quad (식\ 2)$$

$S_{M2}(q,d)$는 M2의 검색 점수이며 S_{entity}와 S_{law}는 엔터티 및 법령, 조문, 버전 정보의 일치도를 나타낸다. Internal의 고정 설정을 두 test에 공통 적용하였다. entity_metadata_boost와 law_metadata_boost는 각각 0.045이며, (식 2)는 엔터티와 법령 정보가 검색 점수에 반영되는 구조를 나타낸다. 실제 계산에는 필드별 가산점과 상한을 적용하며, 세부 설정은 Table 8에 제시하였다.

법률형 질문에는 엔터티,법령 가중치를 적용하고 비법률형 질문에는 M2의 섹션 점수를 우선한다. 후보가 섹션 적합도나 근거밀도를 낮추면 M2 결과를 보존한다. 정규화한 메타데이터는 재순위화뿐 아니라 그래프 노드와 원문 Chunk를 연결하는 공통 식별자로도 활용한다.

4. Hybrid Retrieval and Graph-enhanced Retrieval

M4는 OpenSearch 기반 BM25와 FAISS 기반 Dense Retrieval을 결합한 검색에 Legal Knowledge Graph를 추가한다. BM25는 사건번호,법령명과 같은 정확한 문자열을, Dense Retrieval은 의미적으로 유사한 Chunk를 탐색한다.

Fig. 5와 같이 두 검색 결과를 RRF로 통합하고 섹션,법률 메타데이터 점수로 재순위화한 뒤, 그래프 관계로 사건간 후보를 확장한다. 순위 통합은 (식 3)과 같다.

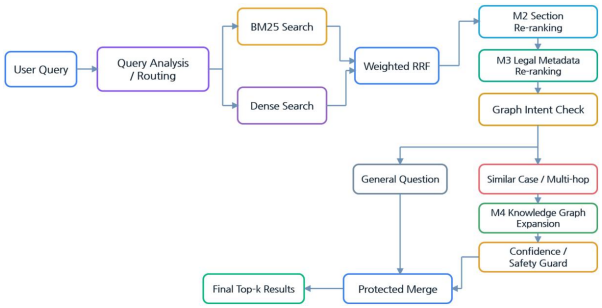

Fig. 5. Graph-enhanced Hybrid Retrieval Process

$$RRF(d) = \sum_{r\in R} w_r/(k + rank_r(d)) \quad (식\ 3)$$

R은 결합할 검색기의 집합이며 $rank_r(d)$는 검색기 r에서 대상 d의 순위이다. 검색기별 가중치 w_r는 BM25에 0.45, Dense Retrieval에 0.40을 적용하고 RRF 상수 k=60을 사용하였다. Legal Knowledge Graph는 사건을 중심으로 피심인, 위반유형, 법조항, 시장, 제재 및 근거 Chunk를 연결한다(Fig. 6). 노드와 관계에는 원문 문서와 근거 Chunk의 식별자를 저장하여 출처를 추적한다.

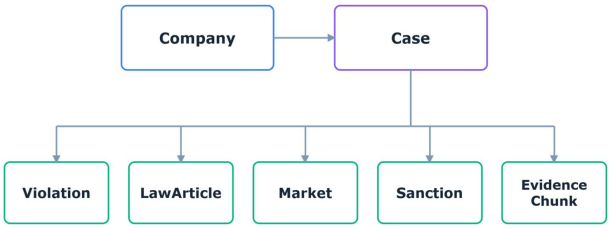

Fig. 6. Legal Knowledge Graph Schema

그래프는 법령, 피심인, 위반유형, 시장을 공유하는 사건을 추가 검색 후보로 확장한다. 노드, 관계, 경로, 원문 근거가 Graph 적합도에 기여하는 구조는 (식 4)와 같이 표현한다.

$$S_{graph}(q,d) = \lambda_1 S_{node}(q,d)+\lambda_2 S_{rel}(q,d)+\lambda_3 S_{path}(q,d) + \lambda_4 E_c(d) \quad (식\ 4)$$

여기서 q는 질의, d는 대상 Chunk이다. S_{node}, S_{rel}, S_{path} 및 E_c는 각각 노드 일치도, 관계 중요도, 경로 적합도, 원문 근거 추적성을 나타낸다. (식 4)와 (식 5)는 이들 요소와 Graph 기여를 나타내는 개념식이며, λ_1~λ_4와 ζ를 독립된 실행 상수로 사용한 것은 아니다. 실제 구현은 후보 생성 방식별 계수를 적용하고 Graph 기본 가중치는 0.030이다(Table 8). 유사사건 후보는 위반유형·법령·조문·시장 일치도와 질의 관련성, 관계·경로 신뢰도를 가중합산하여 계산한다. 탐색 점수≥0.20, 질의 관련성≥0.02, 관계 신뢰도≥0.00, 경로 신뢰도≥0.25를 모두 만족한 후보만 일반 순위 편입 대상으로 선별하였다. 계산 계수와 선별 임계값은 구분한다.

www.kci.go.kr

(식 5)는 일반 순위 편입 대상으로 선별한 Graph 후보의 기여를 M3 점수에 결합하는 개념식이다. ζ는 그 기여의 크기를 나타내며, 점수 결합과 최종 검색결과의 채택은 구분한다.

$$S_{M4}(q,d)=S_{M3}(q,d)+\zeta S_{graph}(q,d) \quad (식\ 5)$$

추가 실험에서는 Graph 근거가 증가한 후보에 대해 M3 결과 대비 섹션 적합도, 법률 근거 Coverage 및 근거 밀도를 비교하였다. 각 지표의 감소폭이 각각 0.10, 0.05, 0.03을 넘지 않을 때만 결합 결과를 채택하였다. 채택 조건을 충족하지 못하면 M3 기반 일반 검색 순위를 유지한다. 이 판정은 검색결과의 proxy를 사용하며, 평가용 정답 정보에 따른 성능 비교나 일반 검색 성능의 비저하를 보장하는 절차가 아니다.

M4는 M3의 검색결과에 그래프 관계를 활용한 보조 근거를 추가한다. Internal-test 144문항에서 M3와 M4의 상위 10개 검색결과 순서와 Recall@10, nDCG@10, MRR@10은 모두 동일하였다. 따라서 Graph의 기여는 일반 순위 성능과 구분하여 관계 근거와 경로 회수 성능으로 평가하였다. Graph 통제 실험은 유사사건 질문군을 대상으로 하였으며 별도의 multi-hop 질문군은 평가하지 않았다.

Graph의 추가 효과는 검색 대상 문서,Chunk와 경로의 회수 성능으로 검증하며, 개별 추출 관계의 의미적 정확도와 구분한다. 원문 근거 식별자의 연결이나 경로 형식의 유효성이 관계 내용의 정확성을 보장하지는 않는다. 관계 추출의 precision,recall 또는 전문가 판정 정확도는 별도 검증이 필요하다.

Graph 재실험의 검색에는 SQLite에 저장된 문서 및 Chunk 메타데이터를 사용하였다. Internal 검색 데이터는 IV-1의 내부 실험 corpus와 같은 의결서 27건과 Chunk 545개이며, External 검색 데이터는 문서 120건과 Chunk 1,825개이다.

Typed graph의 노드는 사건(Case), 기업(Company), 법률(Law), 조문(Article), 위반유형(Violation), 시장(Market), 제재(Sanction)로 구분하였다. 노드 식별자를 기준으로 중복을 제거한 결과, Internal은 752개, External은 2,022개였다. 노드 간 관계는 사건과 기업, 위반행위, 적용 조문, 시장, 제재 및 유사사건 등의 연결을 나타낸다. 출발 노드,관계 유형,도착 노드가 같은 관계를 하나로 계산한 결과, Internal은 1,440개, External은 5,665개로 집계되었다. 이 규모는 추가 재실험에서 실제 검색에 사용한 그래프를 기준으로 산정하였으며, Neo4j 서버에 저장된 전체 그래프나 기존 실험 당시의 규모와는 구분하였다.

Internal과 External 실험에서는 Graph 확장 후보를 최대 10개, 보조 근거 연결을 최대 2개로 제한하였다. 일반 순위에 반영할 후보는 (식 4) 뒤의 임계값으로 선별하며, 관계 신뢰도가 0 이상이면 해당 조건을 통과한다. 보조 근거는 필터링 전 후보에서 연결하므로 일반 순위 반영, 최종 채택과 구분하였다.

후보는 유사사건, 공유 노드 또는 문서 구조의 연결을 이용해 생성하였다. 유사사건 경로는 [출발 문서, 출발 Chunk, 대상 문서, 대상 Chunk]로, typed 경로는 여기에 중간 식별자를 추가하여 기록하였다. 중간 식별자는 공유 노드를 우선하고, 없으면 부모 섹션이나 사건 식별자를 사용하였다. 식별자 목록의 길이는 실제 hop 수와 다르다. OFF/ON 실험에서는 후보 확장만 변경하고 기존 결과의 보호, proxy 채택 조건은 유지하였다.

Internal 3,433건과 External 16,615건의 관계-근거 레코드는 식별자,근거 Chunk 존재, 인용 구간의 유효성, 원문 문자열 일치 점검을 모두 통과하였다. 이는 구조와 근거 연결의 자동 점검이며, 관계의 의미적 정확도나 추출 precision,recall을 뜻하지 않는다.

5. Agentic Retrieval Control

Hybrid Retrieval과 Graph Retrieval을 수행하더라도 일부 질의에서는 적용 법령, 위법성 판단 또는 제재 등 답변 구성에 필요한 핵심 근거가 충분히 검색되지 않을 수 있다. 이러한 문제를 해결하기 위해 본 연구에서는 M5에서 검색 결과의 근거 충분성 여부를 평가하고, 부족한 경우 검색 전략을 동적으로 조정하는 Agentic Retrieval Control을 적용하였다. 본 연구에서 Agent는 독립적으로 법률 판단을 수행하는 생성형 에이전트가 아니라, 검색 품질과 근거의 충분성을 평가하고 필요한 경우 재검색을 수행하는 검색 제어 계층으로 정의한다.

Fig. 7은 단일 규칙 기반 controller의 처리 흐름을 나타낸다. 도식의 Query Analysis, Retrieval Plan, Retrieval 및 Evidence Evaluation은 독립적인 LLM agent가 아니라 동일 controller 내부의 기능적 역할을 의미한다. 질문유형,질의 표현으로 검색 경로를 선택하고, 검색된 Chunk와 그래프 근거 및 proxy 점수를 상태로 관리한다. 근거가 부족한 경우에만 실패 유형에 따른 제한적 Fallback Retrieval을 수행하며, 검색 근거의 충분성은 (식 6)으로 계산한다.

www.kci.go.kr

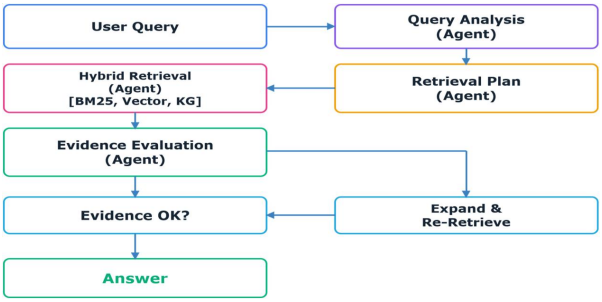

Fig. 7. Agentic Retrieval Control Workflow

$$Evidence_Coverage = N_{re_e_f} / N_{re_e} \quad (식\ 6)$$

여기서 필요한 근거 수 N_{re_e}는 질문 원문에서 규칙 기반으로 추정한 질문유형에 따라 지정한 필수 evidence slot의 수이며, 확보된 근거 수 $N_{re_e_f}$ 는 그중 검색 결과가 충족 조건을 만족한 slot의 수이다. 각 slot은 충족 여부에 따라 한 번만 계산하며, 동일 항목을 뒷받침하는 Chunk가 여러 개 검색되더라도 중복하여 세지 않는다. 필수 항목의 지정 조건과 충족 판정 규칙은 다음과 같다. 별도 명시가 없는 항목은 상위 10개 검색 Chunk를 기준으로 판정한다.

① 문서(DOCUMENT): 모든 질문에 지정하며, 검색 결과가 한 건 이상이면 충족한다.

② 섹션(SECTION): 모든 질문에 지정하며, 추정 질문유형에 대응하는 섹션의 Chunk 비율이 0.10 이상이면 충족한다.

③ 엔터티(ENTITY): 사건,피심인 확인형, 위반행위 사실 확인형, 복합 질의형에 지정한다. 문서명 가산점이 양수이거나 피심인 정보(respondent, respondent_name)가 있는 Chunk가 하나 이상이면 충족한다.

④ 법적 근거(LAW): 적용 법조항, 법률 판단, 시계열,법령 개정 및 복합 질의형에 지정한다. 질문 원문과 추정 유형의 토큰 중 검색 본문,메타데이터에서 확인되는 비율이 0.10 이상이면 충족한다. 이는 법률 관련 근거의 확보 여부를 추정하는 기준이며, 필요한 법률과 조문이 정확히 검색되었는지를 확인하는 판정은 아니다.

⑤ 제재(SANCTION): 제재,과징금 및 복합 질의형에 지정한다. 제재 메타데이터, 시정명령,과징금,경고,과태료,납부명령,교육명령,공표명령 등의 표현이나 과징금 부과,납부,결정에 해당하는 금액이 있으면 충족한다.

⑥ 그래프 관계(GRAPH): 유사사건 탐색형에 지정한다. HAS_CHUNK 이외의 관계, 유사사건 식별자 또는 식별자 목록 길이가 3 이상인 경로가 있으면 충족한다.

⑦ 시점,법령 버전(TEMPORAL): 시계열,법령 개정 비교형에 지정한다. 상위 10개 Chunk 중 시점,버전 정보가 있고, 조문 연결 상태가 exact_at_applicable_date 또는 latest_only이면 충족한다. exact_at_applicable_date는 적용 기준일을 포함하는 유효기간의 조문 버전을 찾은 상태이다. latest_only는 적용 기준일이 없거나 해당 기준일에 맞는 조문 버전을 확보하지 못하여, 저장된 최신 조문 버전을 선택한 상태이다. unresolved는 조문 버전을 연결하지 못한 상태이다. latest_only의 인정은 시점, 법령 버전 관련 근거가 검색되었음을 확인하는 보조 판정이며, 사건 당시 적용 법령과의 일치나 법률적 타당성을 검증한 것은 아니다.

예를 들어 복합 법률질의형에는 DOCUMENT, SECTION, ENTITY, LAW, SANCTION의 5개 항목(slot)을 지정하며, 4개가 충족되면 Evidence Coverage는 0.8이다. 질문유형은 질문 원문에서 추정하고 정답 문서, Chunk, 섹션, 필수 엔터티, 법조항 및 정답 힌트는 검색 입력에서 제외하였다. 최종 Internal-test M5의 평균 N_{re_e}는 3.3750, 평균 $N_{re_e_f}$는 3.1597이며 질문별 Evidence Coverage의 평균은 0.9411이었다.

Table 5. Retrieval Failure Types and Fallback Strategies

Failure Type	Description	Fallback Strategy
MISSING_ DOCUMENT	Relevant document is not retrieved.	Query rewriting and candidate expansion
MISSING_ SECTION	Required document section is missing.	Section relaxation and adjacent chunk expansion
MISSING_ ENTITY	Required respondent or violation entity is missing.	Entity-based re-retrieval
MISSING_ LAW	Applicable law or article is missing.	Law and article expansion
RESPONDENT_ MISMATCH	Retrieved document belongs to a different respondent or case.	Apply document constraints
QUERY_DRIFT	Query rewriting changes the original intent.	Preserve original query entities
GRAPH_OVER_ EXPANSION	Graph traversal retrieves weakly related cases.	Restrict path length and relation scores
INSUFFICIENT_ EVIDENCE	Retrieved evidence is insufficient for answer construction.	Execute multiple fallback strategies

M5는 질문 유형과 표현에 따라 그래프 관계 탐색은 M4, 법률 근거 검색은 M3, 목차 기반 검색은 M2를 우선 선택한다. 검색 근거가 기준에 미달할 때만 질의 재작성, 섹션 제한 완화 또는 후보 확장을 수행한다. 엔터티 누락(MISSING_ENTITY)은 엔터티 중심으로 재검색하고, 질의 의미 이탈(QUERY_DRIFT)은 핵심 엔터티를 유지하여 보완

www.kci.go.kr

한다. 재검색 결과는 근거 충분성의 자동 점수를 개선하고 문서 집중도, 근거밀도, 섹션 적합도, 법률 근거와 그래프 근거의 확보 수준을 낮추지 않을 때만 채택한다. 상위 10개 결과 중 1~8위는 유지하고 9~10위만 교체하며, 조건을 충족하지 못하면 기존 결과를 답변 구성과 근거 검증에 사용한다. 이 과정에는 평가용 정답 정보를 사용하지 않는다.

Table 6. M5 Guard Parameters

Parameter	Value
m5_evidence_density_threshold	0.14
m5_proxy_min_gain	0.02
m5_retry_max_replacements	1
m5_retry_protected_top_n	8
m5_retry_top_k	40

6. Extractive Answer Construction and Evidence Verification

최종 답변은 생성형 LLM을 호출하지 않고 검색된 Evidence Chunk에서 관련 구문을 선택하는 결정적 추출 방식으로 구성한다. 답변을 Claim으로 나누어 근거와의 대응을 점검하고, 뒷받침되지 않는 문장은 제외하거나 표현을 제한한다(Fig. 8).

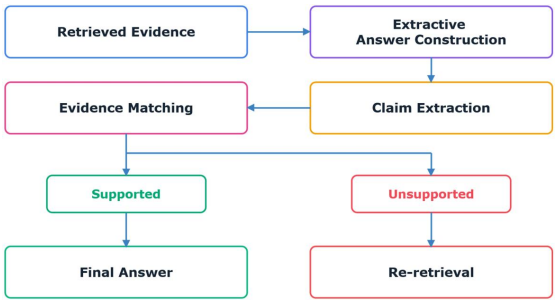

Fig. 8. Claim-based Extractive Answer Construction and Evidence Verification

각 주장(Claim)에 대해 검색된 근거 Chunk와의 최대 의미 유사도를 근거 지지 점수로 사용한다(식 7). 이는 근거가 부족한 주장을 걸러내는 보조 기준으로, 주장의 사실 여부를 독립적으로 판정하는 값은 아니다. 답변 품질은 검색 종료 후 LLM Judge와 별도의 claim-evidence entailment로 평가하였다.

$$Support(c_i)=max_{ej \in E} \; sim(c_i,e_j) \quad (식\ 7)$$

c_i는 답변의 i번째 주장, E는 검색된 근거 Chunk의 집합이며, $sim(c_i,e_j)$ 는 주장과 근거의 의미적 유사도이다. 근거 지지 점수(Support Score)가 기준에 미달하면 해당 주장을 제외하거나 표현을 제한한다.

질문과 무관한 근거가 답변에 포함되지 않도록 조문, 섹션, 사건 및 엔터티의 일치도를 점검하였다. 검토한 전체 후보 중 관련성이 높은 근거의 비율을 Evidence Density로 계산한다.

$$Evidence_Density=N_{h\text{-}r_c}/N_{a_c_c} \quad (식\ 8)$$

$N_{h\text{-}r_c}$는 관련성이 높은 근거 Chunk 수이며, $N_{a_c_c}$는 전체 후보 Chunk 수이다. 이 비율이 임계값보다 낮으면 관련성이 낮은 후보를 제거하거나 재정렬한다(Fig. 9).

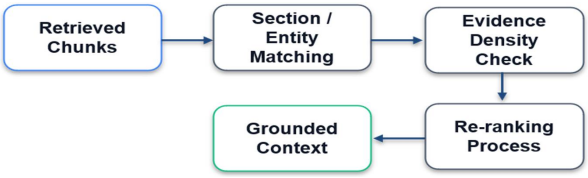

Fig. 9. Evidence Refinement Process based on Precision-aware Evidence Control

검색에 사용하는 정보와 평가용 정답 정보는 Table 7과 같이 분리하였다. LLM Judge와 별도의 entailment 평가는 검색이 끝난 뒤 수행하며 검색 제어에는 사용하지 않는다.

Table 7. Separation of Runtime and Evaluation Information

Information	Runtime Retrieval Control	Evaluation
Question text	Use	Use
Question type/intent	Use	Use
Case name inferable from query	Use	Use
Gold document ID	Do not use	Use
Gold chunk ID	Do not use	Use
Gold section	Do not use	Use
Required entity/law annotation	Do not use	Use
Answer hint	Do not use	Use
Proxy derived from retrieval results	Use	Use

IV. Experiments

1. Experimental Setup

내부 corpus는 의결서 27건, 216문항이며 dev 72문항과 Internal-test 144문항으로 질문 단위 분리하였다. 외부 corpus는 내부와 겹치지 않는 의결서 120건, 960문항으로, External-dev 30건, 240문항과 External-test 90건, 720문항으로 문서 단위 분리하였다. 외부 설정 비교에는 dev 중 5건, 40문항을 사용하였다. 갱신 평가에서는 Internal의 검색 설정을 사전에 고정하여 두 test의

www.kci.go.kr

M1~M5에 공통 적용하였다. 비교 규모는 864문항×5개 모델의 4,320건이다. 정답 문서, Chunk, 섹션, 엔터티, 법조항 및 기준 답변은 검색 종료 후 평가에만 사용하였다.

질문유형 오류의 영향은 기존 864문항을 이용한 별도 통제 재실험으로 확인하였다. 재실험 시점의 소스, 설정, 질문, 인덱스를 스냅샷으로 보존하고, 이 실험 안에서 질문, 인덱스, 가중치, 후보 수를 고정하였다. 자동추정 유형, 정답유형 및 나머지 7개 유형을 각각 적용하여 M2, M5의 총 15,552건을 실행하였다. 유형 변경은 M2의 우선 검색 섹션과 M5의 필수 근거 항목, 검색 경로, 재검색 판단에 일관되게 반영하였다. 정답유형 지정은 평가용 조건에 한정하고 실제 검색 방식은 변경하지 않았다. 이 실험은 유형 변경의 영향을 비교한 것으로, 기존 M1~M5 최종 실행을 그대로 재현한 실험과 구분한다.

평가지표는 Recall@10, nDCG@10, MRR@10 및 필수 근거 항목의 자동 충족률인 슬롯 Coverage proxy이다. 유형을 바꾸더라도 평가 기준과 분모는 정답유형에 맞추어 고정하였다. 질문별로 잘못 지정한 7개 유형의 평균과 정답유형의 차이를 계산하고, paired bootstrap과 양측 paired sign-flip 검정을 각각 10,000회 수행하였다. 신뢰구간은 개별 비교의 95% 구간이며, 검정에는 Monte Carlo plus-one 보정과 전체 16개 비교의 Holm 보정을 적용하였다. 유형 적용 순서는 고정된 난수 시드로 섞었으며 생성 LLM과 사후 Judge는 호출하지 않았다. Fig. 10은 검색 저장소, 벡터 인덱스, 그래프 및 검색 제어 계층의 연결 구조를 나타낸다. 추가 Graph 통제 실험에는 로컬 BM25, FAISS와 SQLite 그래프 메타데이터를 사용하였으며, III-4의 그래프 규모와 경로 설정도 이 검색 스냅샷을 기준으로 제시하였다.

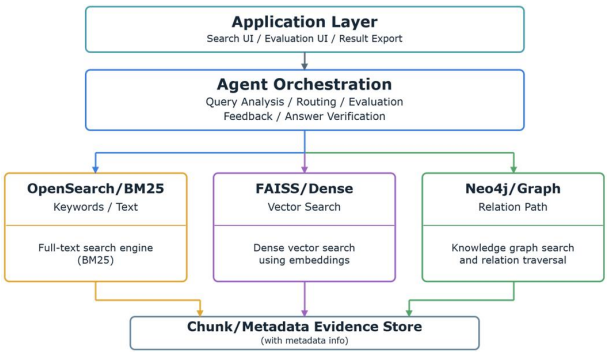

Fig. 10. Experimental Architecture

Fig. 10에서 OpenSearch는 어휘 검색, FAISS는 벡터 검색, Neo4j는 법률 개체와 관계 저장을 담당한다. 각 데이터는 Chunk 식별자로 연결하여 원문 근거를 추적한다. 다만 추가 통제 실험은 로컬 BM25,FAISS와 SQLite 메타데이터를 사용하였으며, 평가 규모와 임베딩,Judge 설정은 Table 8에 제시하였다.

섹션 가중치 γ=0.045와 RRF k=60을 유지하였다. Internal dev 72문항은 단계별 비교와 M4·M5의 경로·근거 점검에 사용하였으며 모든 계수를 개별 최적화한 것은 아니다. 두 test에는 엔터티·법령 가중치 각각 0.045, Graph 기본 가중치 0.030, M5 근거밀도 기준 0.14, 최소 개선폭 0.020을 공통 적용하였다. 이 공통 설정과 Table 8의 External-dev 비교 설정을 대조한 결과, External M3~M5의 세 일반 순위지표는 질문별로 같았으나 M5의 재검색과 일부 답변은 달랐다.

가중치 조합 비교에는 External-dev의 의결서 5건에서 추출한 40문항을 사용하였다. 8개 질문 유형별로 5문항씩 구성하였으며, 최종 External-test와 질문·문서 식별자 및 원문 SHA256 해시의 중복은 각각 0건이었다. 다만 원사건과 후속 사건 사이의 관계까지 배제한 것은 아니다.

Table 8. Fixed Settings and Reproducibility Information for the Evaluation

Item	Value
Evaluation data	Internal dev/test: 72/144 questions; External dev/test: 240/720 questions; tuning subset: 40 questions. M1-M5: 4,320 test cases.
Embedding / index	BAAI/bge-m3; 1,024 dimensions; normalized vectors; FAISS IndexFlatIP. Local BM25 and SQLite metadata in controlled runs.
Answer / NER / rewrite	Deterministic extractive answer; rule/dictionary NER; deterministic query rewrite. No generation, NER or rewrite LLM calls.
Chunking	max_chars=1,800; overlap_chars=200; min_chars=450 (characters).
Retrieval	Final top-k=10; BM25/dense candidate count=max(20, requested top-k × 30); RRF k=60; section boost γ=0.045.
BM25 / fusion / other boosts	Stored BM25Okapi: k1=1.5, b=0.75, epsilon=0.25; BM25 RRF weight=0.45, dense=0.40; document/name boost=0.14; legal text=0.012; temporal=0.035.
Common test profile	entity_metadata_boost=0.045; law_metadata_boost=0.045; graph_metadata_boost=0.030. Both updated test runs use this fixed profile.
M5 guards / retries	evidence density=0.14; minimum proxy gain=0.020; maximum retries=3; protect ranks 1-8 and replace only ranks 9-10 if all guards pass.
Runtime evidence check	All required slots must be fulfilled and density≥0.14 for PASS; coverage is fulfilled/required slots. m5_recall_threshold=0.8 belongs to the legacy gold-dependent branch and is not the leakage-safe runtime coverage threshold. Maximum retry candidates=40; at most one rank 9-10 replacement per attempt.

www.kci.go.kr

Item	Value
Graph equation correspondence	λ1-λ4 and ζ describe conceptual contributions, not separate scalar configuration keys. Implemented coefficients and gates are listed below.
Graph reranker	graph_path_rerank_weight=0.030 (fallback to graph_metadata_boost); raw=max(0, 0.55×relation + 0.25×path + 0.20×overlap − penalty); over-expansion penalty=0.60; maximum path length=4.
Relation weights	SIMILAR_TO=1.00; APPLIED_ARTICLE=0.95; APPLIED_ARTICLE_VERSION=1.10; FOUND_VIOLATION=0.90; ORDERED_SANCTION=0.80; IN_MARKET=0.55; HAS_ADMINISTRATIVE_RULE=0.85; HAS_CHUNK=0.05; other=0.15; evidence bonus=0.20.
Similar candidates	Profile: violation=0.40, law=0.25, article=0.15, market=0.15; query=0.10, relation=0.10, path=0.05, section bonus=0.05; maximum similar cases=5.
Typed candidates	ArticleVersion/Article=0.35; Violation=0.28; LawVersion/Law=0.22; Market=0.14; Sanction=0.12; Company=0.08; Case=0.06; other=0.05. Query=0.25; relation=0.10; section=0.22; same document=0.30; same parent=0.35; adjacent (distance≤2)=0.15.
Graph limits / gates	Maximum expansion candidates=10; supplemental links=2. Ranking gates: traversal≥0.20, query relevance≥0.02, relation confidence≥0, path confidence≥0.25. Supplemental evidence is handled separately.
Post-hoc Judge	gpt-4.1-mini (response model: gpt-4.1-mini-2025-04-14); temperature=0; evidence limit=6,000 characters; judge_v2_strict_hallucination_calibrated.
Post-hoc entailment	Same response model; temperature=0; evidence limit=9,000 characters; claim_evidence_entailment_v1. Labels: entailed / contradicted / unsupported.
Entailment scoring	For every answer: number of entailed claims / total claims; no binary score threshold. Mean over 144 / 720 / 864 answers. Raw model scores retained separately; 42 differ by >0.001; missing raw scores=0.
Citation validation / records	Only evidence indices actually present in the evaluation input are allowed. Thirty invalid-index responses were retried with the same answer, evidence and rubric; the first valid response was retained. Prompts, model, usage, response IDs, input/config/source/index hashes and per-question records are archived.
Runtime separation	Gold annotations and answer hints excluded from search; Judge and entailment run only after answer construction. Query embeddings reused in retrieval rechecks; runtime is descriptive, not a repeated latency benchmark.

Table 8은 평가에 사용한 검색, 답변 구성, 사후 평가 설정이다. 질문, 소스, 설정, 인덱스를 고정하고 실행 기록과 해시를 보존하였다. 답변은 생성형 LLM 없이 추출 방식으로 구성하며, Judge와 entailment는 완성된 M5 답변을 사후 평가한다. 과거의 외부 dev 선택 조합은 비교 이력으로 구분하였다.

2. Evaluation Metrics

성능은 일반 검색, 도메인 특화 검색 및 추출형 답변 품질로 구분하여 평가하였다. 일반 검색에는 Recall@10, nDCG@10, MRR@10을, 문서 구조와 법률 근거의 회수에는 Section Recall, Entity Coverage, Law Article Coverage, Graph Evidence를 사용하였다. 답변 품질은 Grounding, Answer Relevance, Proxy Hallucination과 사후 LLM Judge로 평가하였다. Judge는 근거 일치성, 충실성, 답변 관련성, 법률적 정확성, 인용 정확성, 완전성 및 환각 위험을 평가하였다. 점수가 상위 구간에 집중되는 천장효과를 보완하기 위해 M5 전체 864개 답변에 별도의 주장,근거 일치 평가(entailment)를 추가하였다.

Graph 전용 평가는 승인된 Internal 18문항과 External-test 중 정답 문서, Chunk, 경로가 정의된 80문항을 사용하였다. Internal 전용 질문은 전체 144문항과 일부 표현이 달라 별도 평가셋으로 구분하며, External의 유사사건 질문 90문항 중 정답 정보가 없는 10문항은 제외하였다. 모두 유사사건 탐색형이며 별도 multi-hop 질문군은 평가하지 않았다. 정답 정보는 검색 후 세 Graph 지표 계산에만 사용하였다.

고정된 같은 환경에서 M3, M4 확장 OFF, M4 확장 ON을 실행하였다. Table 11의 M4와 Table 12의 ON에는 동일한 질문별 ON 기록을 사용하였다. 구성요소 비교의 OFF와 BOTH도 이 OFF, ON 기록을 공유하고, SIMILAR와 TYPED만 추가 실행하였다. 각 비교에서 질문, 인덱스, 임베딩, 가중치, 후속 처리, 평가 기준을 고정하고 생성 LLM과 Judge는 호출하지 않았다. 실행 전후 인덱스 해시가 동일함을 확인하였다.

Typed 후보는 공유 노드와 문서,부모 섹션,인접 Chunk의 구조적 연결을 이용한다. 다만 순위 편입에 필요한 질의 관련성,경로 신뢰도 필드가 없는 후보는 기존 필터를 통과하지 못하고 보조 근거로만 연결될 수 있다. 따라서 구성요소 비교는 현재 구현에서 각 후보 생성 기능을 제거하거나 유지했을 때의 효과로 해석하였다.

질문별 차이에 paired bootstrap과 양측 paired sign-flip 검정을 각각 10,000회 적용하였다. 신뢰구간은 개별 비교의 95% 구간이며 p값에는 Monte Carlo plus-one 보정을 적용하였다. M3, M4 비교는 보정 전 p값, OFF/ON 비교는 두 평가셋의 세 지표에 대한 Holm 보정 p값(6개 비교)을 제시하였다. 구성요소 비교는 다섯

www.kci.go.kr

대비×두 평가셋×세 지표의 30개 비교를 별도로 Holm 보정하였다.

전체 단계 비교는 Internal 144문항과 External 720문항의 M1→M2, M2→M3, M3→M4, M4→M5를 대상으로 하였다. 모든 단계에서 새로 산출한 Recall@10, nDCG@10, MRR@10, Grounding, Answer Relevance의 40개 비교를 하나의 Holm 보정 집합으로 정하였다. 이번 사후 Judge, entailment는 M5만 재평가했으므로 단계 간 검정에는 포함하지 않았다. 시간은 단일 실행의 기술통계로 해석하였다.

2.1 Retrieval Metrics

Recall@k는 정답 Chunk 중 상위 k개의 검색 결과에 포함된 정답 Chunk의 비율로, 다음 (식 9)와 같이 계산한다. R_k는 검색 모델이 반환한 상위 k개의 Chunk 집합이고, G는 해당 질문에 연결된 Ground Truth Chunk 집합이다.

$$Recall@k=|R_k \cap G|/|G| \quad (식\ 9)$$

MRR@k는 각 질문에서 상위 k개 검색결과 내에 최초로 나타나는 정답 Chunk 순위의 역수를 전체 질문에 대해 평균한 지표이다. 상위 k개 내에 정답 Chunk가 없으면 해당 질문의 점수는 0으로 처리한다. 본 연구에서는 k=10을 적용하여 MRR@10을 산출하였다. N은 전체 평가 질문 수이며, $rank_i$는 i번째 질문의 상위 10개 검색결과에서 최초로 나타나는 정답 Chunk의 순위이다.

$$MRR@k=(1/N)\Sigma_i\ 1/rank_i \quad (식\ 10)$$

nDCG@k는 정답 Chunk의 관련도와 검색 순위를 함께 평가하는 지표이며 (식 11)과 같이 계산한다. DCG@k는 실제 검색결과의 누적 할인 이득(Discounted Cumulative Gain)이고, IDCG@k는 관련도에 따라 이상적으로 정렬한 결과의 최대 DCG이다.

$$nDCG@k=DCG@k/IDCG@k \quad (식\ 11)$$

2.2 Domain-Specific Retrieval Metrics

도메인 검색 성능은 Section Recall, Entity Coverage, Law Article Coverage 및 Graph Evidence로 평가하였다. 앞의 세 지표는 각각 평가용 정답에 지정된 섹션, 필수 엔터티(피심인,위반유형,시장,제재), 법령명,조문이 검색 결과에서 확보된 비율이다. Graph Evidence는 그래프 탐색에서 원문 근거 Chunk로 추적 가능한 관계나 근거를 확보한 비율이다. 그래프를 사용하지 않는 M1~M3에서는 0으로 산출되며, M4, M5의 그래프 근거 회수 성능을 평가하는 데 사용하였다.

2.3 Extractive Answer Quality Metrics

추출형 답변 품질은 Grounding, Answer Relevance 및 Proxy Hallucination으로 평가하였다. 표의 Grounding은 검색 근거의 정합성, 답변 토큰의 근거 중복률, 인용번호 충족률, 주장 지지 비율을 결합한 사후 자동평가값이며 (식 12)로 계산한다. M5의 검색 제어용 proxy와 구분하며 법률적 정확성을 직접 판정하지 않는다.

$$Grounding = min\{1,\ 0.74G_R + 0.10T + 0.08C + 0.08S\} \quad (식\ 12)$$

G_R은 Recall@10, 문서 회수율, Section Recall, Entity Coverage, Law Article Coverage, 근거 텍스트 중복률, 근거밀도, Graph Evidence, 시점 일치도의 가중합이다. 가중치는 순서대로 0.11, 0.07, 0.34, 0.13, 0.11, 0.12, 0.06, 0.04, 0.02이다.

T는 답변 토큰 중 상위 5개 근거에서 확인되는 비율이다. C는 답변의 인용번호 [1]~[8] 중 나타난 번호 수를 min(검색 Chunk 수, 5)로 나눈 값이며 최대 1로 제한한다. S는 규칙으로 분리한 주장 중 상위 10개 근거와의 토큰 중복률이 0.25 이상인 주장의 비율이다. 근거가 없으면 $S=0$, 근거는 있으나 주장이 검출되지 않으면 $S=1$로 처리한다. 이는 주장 분리 규칙에서 검출된 항목이 없을 때 적용하는 기본값이며, 답변의 사실성이 확인되었다는 의미는 아니다. 토큰은 두 글자 이상의 한글·영문·숫자 연속 문자열의 집합으로 정의하였다. 검색 정답과 기준 답변은 사후 평가에만 사용하고 runtime 검색 제어와 분리한다.

Proxy Hallucination은 (식 13)과 같이 1−Grounding으로 계산한 근거 불일치 대리지표이다. Grounding과 동일한 정보를 역방향으로 나타내므로 독립적인 평가 지표나 실제 환각 발생률로 해석하지 않는다. 본 연구에서는 기존 지표명과의 대응을 위해 이 명칭을 유지하되, 추출형 답변의 근거 대응성을 설명하는 보조값으로 사용한다.

$$Proxy_Hallucination=1-Grounding \quad (식\ 13)$$

Answer Relevance는 기준 답변의 토큰 중 추출형 답변에서 확인되는 비율로 계산한 사후 자동평가값이다. 기준 답변은 평가에만 사용하고 검색 제어에는 제공하지 않았다. M5의 864개 답변은 Judge로 근거 일치성·충실성·답변 관련성·법률적 정확성·인용 정확성·완전성·환각 위험을 평가하고, 별도 지침의 claim-evidence entailment로 주장과 근거의 대응을 점검하였다. 두 평가는 같은 LLM을

www.kci.go.kr

사용하므로 모델 자체가 독립적이라는 의미는 아니며 전문가 검증을 대신하지 않는다.

처리 시간은 검색,재검색,추출형 답변 구성을 포함한 실행 시간과 사후 Judge 평가 시간으로 나누어 각각 pipeline_latency_ms와 judge_latency_ms로 집계하였다. 검색 반복 횟수는 초기 검색과 재검색의 합이며 내부 함수 호출 수와는 다르다. 이 결과는 최종 M5의 단일 실행 기록을 요약한 것으로, 반복 실행에 따른 시간 안정성을 검증한 것은 아니다.

3. Overall Results

3.1 Overall Performance Comparison

Fig. 11은 모든 단계에서 갱신한 Recall@10, nDCG@10, MRR@10, Grounding, Answer Relevance 다섯 지표의 단순 평균이다. Judge는 M5만 새로 평가했으므로 이번 단계별 보조 종합점수에는 포함하지 않았다. M1~M5의 점수는 Internal에서 0.5981, 0.6958, 0.6936, 0.6938, 0.7012이고 External에서 0.5911, 0.6579, 0.6686, 0.6687, 0.6716이다. 이는 종합 경향을 보여주는 보조값이며 개별 지표와 통계검정을 대신하지 않는다.

Recall@10은 M1에서 M5로 확장하면서 Internal 0.7380→0.8708, External 0.6861→0.8493으로 증가하였다. 두 평가셋 모두 가장 큰 증가는 M1→M2에서 나타났으며 M3, M4의 일반 순위지표는 동일하였다.

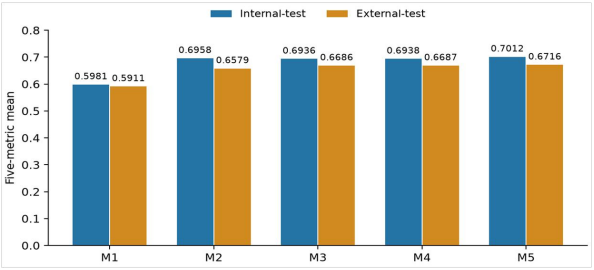

Fig. 11. Overall Performance Comparison

3.2 Retrieval Performance Comparison

Internal-test의 nDCG@10은 M1 0.5809에서 M5 0.6283으로 증가하였다(Fig. 12). MRR@10은 M1 0.5854, M2 0.5212, M5 0.5225였다. External-test의 MRR@10은 M1 0.7652에서 M2 0.5602로 감소했으며 M5는 0.5603이었다. Recall 증가는 상위 10개 내 근거 회수 범위의 개선이며 최초 정답의 순위 상승을 의미하지 않는다.

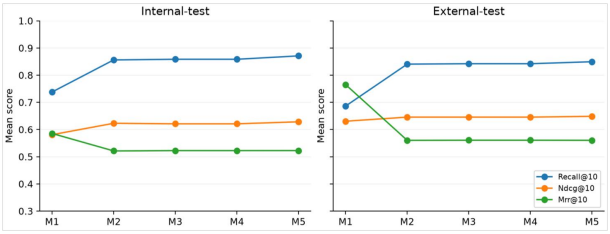

Fig. 12. Retrieval Performance Comparison (Recall@10, MRR@10, nDCG@10)

3.3 Extractive Answer Quality Comparison

Fig. 13은 단계별 자동 답변 품질 지표를 비교한다. Grounding은 Internal에서 0.6053→0.8759, External에서 0.4855→0.7726으로 증가하였다. Answer Relevance도 각각 0.4810→0.6087, 0.3880→0.5272로 증가하였다. Proxy Hallucination은 1−Grounding의 파생값으로 독립적인 환각 검증과 구분한다.

두 지표는 추출형 답변의 근거 대응성과 질문 관련성을 나타낸다.

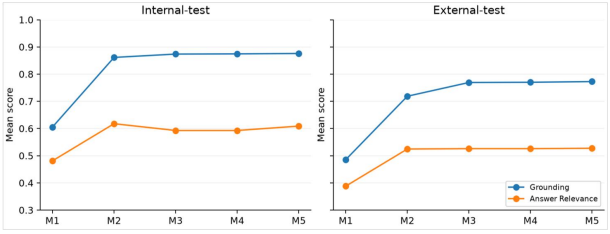

Fig. 13. Extractive Answer Quality Comparison

M5 전체 864개 답변의 claim-evidence entailment 평균은 Internal-test 0.9446, External-test 0.9259, 전체 0.9290이었다(Table 9). 이는 답변의 주장과 검색 근거의 대응 정도를 나타내는 자동평가값이다.

질문 생성과 정확성 평가에는 서로 다른 전문가가 1명씩 참여했지만, 여러 전문가가 동일 답변을 평가하거나 평가자 간 일치도를 분석하지는 않았다. 또한 원문에 근거한 답변이라도 질문과 무관하거나 법적 요건,적용 시점,조문 버전을 잘못 설명할 수 있으므로, 이 자동평가 결과만으로 법률적 정확성을 확인했다고 볼 수는 없다.

Table 9. Claim-Evidence Entailment Results for M5 Answers

Test set	Answers (n)	Mean entailment
Internal-test	144	0.9446
External-test	720	0.9259
Overall	864	0.9290

주: 갱신된 M5 답변 144개와 720개를 평가하였다. 문항별 점수는 근거가 뒷받침하는 주장 수를 전체 주장 수로 나눈 값이며 전체 평균은 864개 문항 점수의 산술평균이다. 소수 넷째 자리까지 표시하였다.

www.kci.go.kr

각 주장을 entailed(근거가 뒷받침함), contradicted(근거와 모순됨), unsupported(근거로 확인되지 않음)로 분류하였다. 문항별 점수는 전체 주장 중 entailed의 비율이며 통과 여부를 나누는 별도 임계값은 두지 않았다. Table 9는 각 평가셋의 문항별 점수를 산술평균한 결과이다.

모델이 별도로 반환한 점수는 검증용으로 보존하고, 본문 결과에는 모든 문항에서 근거가 뒷받침하는 주장의 비율을 일관되게 사용하였다.

평가 모델과 입력 한도 등 세부 설정은 Table 8에 제시하였다. 평가 지침, 입력, 답변 해시, 응답 식별자, 사용 토큰 및 점수 계산 절차를 보존하였다. 입력에 없는 근거번호를 반환한 30건은 동일 답변·근거·평가기준에 허용 번호 제약을 추가하여 다시 평가하고 첫 유효 응답을 채택하였다. 최종 864건 모두 근거번호 검사를 통과하였다.

3.4 Stage-wise and Graph Ablation Analyses

Table 10은 고정된 메타데이터와 공통 검색 설정에서 산출한 Internal-test M1~M5의 결과이다. External-test도 같은 설정으로 720문항을 평가하였다. M5 답변에 대해서는 Judge와 entailment 평가를 별도로 수행하였다.

두 평가셋에서 Recall과 Grounding의 가장 큰 증가는 M1→M2에서 나타났다. M3, M4의 일반 순위지표는 같았으며 M4→M5의 차이는 Holm 보정 후 유의하지 않았다. 단계별 유의성은 갱신된 40개 대응 비교에 근거하였다.

Table 10. Internal-test Results (n=144)

Metric	M1	M2	M3	M4	M5
Recall@10	0.7380	0.8560	0.8583	0.8583	0.8708
nDCG@10	0.5809	0.6229	0.6209	0.6209	0.6283
MRR@10	0.5854	0.5212	0.5224	0.5224	0.5225
Section Recall	0.0000	0.9512	0.9546	0.9546	0.9535
Entity Coverage	0.9202	0.9386	0.9562	0.9562	0.9537
Law Coverage	0.7127	0.7071	0.8104	0.8104	0.8136
Graph Evidence	0.0000	0.0000	0.0000	0.0250	0.0417
Grounding	0.6053	0.8614	0.8737	0.8744	0.8759
Answer Relevance	0.4810	0.6176	0.5927	0.5927	0.6087
Proxy Hallucination	0.3947	0.1386	0.1263	0.1256	0.1241

Table 11은 고정 환경에서 실행한 M3와 M4의 Graph 전용 지표를 비교한다. Table 12는 동일한 M4의 확장 OFF/ON 비교이며 두 표의 M4와 ON은 같은 실행 기록이다. M3, M4 비교의 p값은 다중비교 보정 전 값이다.

Table 11. M3 versus M4 Using the Shared ON Records

Dataset / metric	M3	M4 / ON	Mean diff.	95% CI	p-value
Internal: document recall	0.1389	0.5556	0.4167	[0.2500, 0.5833]	0.00050
Internal: chunk recall	0.0278	0.1389	0.1111	[0.0417, 0.1806]	0.01510
Internal: path coverage	0.0000	0.5556	0.5556	[0.3889, 0.7222]	0.00010
External: document recall	0.0563	0.3250	0.2687	[0.2000, 0.3438]	0.00010
External: chunk recall	0.0344	0.1062	0.0719	[0.0469, 0.1001]	0.00010
External: path coverage	0.0000	0.3063	0.3063	[0.2313, 0.3875]	0.00010

Table 11의 M4와 Table 12의 ON은 동일한 질문별 실행 기록을 공유한다. External 80문항의 문서·Chunk·경로 회수율은 각각 0.3250, 0.1062, 0.3063으로 두 표에서 같다. 두 표는 M3와 OFF라는 비교 조건 및 다중비교 보정 범위를 구분하여 제시하였다.

Table 12. Graph Expansion OFF/ON with Fixed Settings

Dataset / metric	OFF	ON	Mean diff.	95% CI	p-value	Holm p
Internal: document recall	0.1389	0.5556	0.4167	[0.2500, 0.6111]	0.00140	0.00280
Internal: chunk recall	0.0278	0.1389	0.1111	[0.0417, 0.1806]	0.01500	0.01500
Internal: path coverage	0.0000	0.5556	0.5556	[0.3889, 0.7222]	0.00010	0.00060
External: document recall	0.0563	0.3250	0.2687	[0.2000, 0.3438]	0.00010	0.00060
External: chunk recall	0.0344	0.1062	0.0719	[0.0469, 0.1000]	0.00010	0.00060
External: path coverage	0.0000	0.3063	0.3063	[0.2313, 0.3812]	0.00010	0.00060

OFF/ON의 일반 순위지표는 전체 test가 아닌 Graph 전용 Internal 18문항과 External 80문항에서 산출하였다. Table 10은 Internal 전체 144문항의 평균으로 두 평가 대상의 값을 직접 비교하지 않는다. Internal의 Graph 질문 중 일부는 전체 test와 표현도 달라 별도 질문셋으로 구분한다. 전체 단계 대응 분석은 Internal 144문항과 External 720문항을 대상으로 하였으며 40개 비교 중 유

www.kci.go.kr

의한 12개를 Table 13, 14에 제시하였다. ON의 Recall@10, nDCG@10, MRR@10은 Internal 0.9630, 0.6210, 0.4560, External 0.8417, 0.5842, 0.4686이다.

Table 13. Significant Adjacent-stage Comparisons (Internal; Holm family=40)

Stage	Metric	Δ	95% CI	Holm p
M1→M2	Recall@10	0.1181	[0.0718, 0.1653]	0.00400
M1→M2	Grounding	0.2561	[0.2455, 0.2662]	0.00400
M1→M2	Answer Relevance	0.1365	[0.1015, 0.1717]	0.00400
M2→M3	Grounding	0.0123	[0.0091, 0.0161]	0.00400
M2→M3	Answer Relevance	-0.0249	[-0.0414, -0.0100]	0.03190
M3→M4	Grounding	0.0007	[0.0005, 0.0011]	0.00600

Table 14. Significant Adjacent-stage Comparisons (External; Holm family=40)

Stage	Metric	Δ	95% CI	Holm p
M1→M2	Recall@10	0.1544	[0.1318, 0.1772]	0.00400
M1→M2	MRR@10	-0.2051	[-0.2262, -0.1840]	0.00400
M1→M2	Grounding	0.2332	[0.2259, 0.2403]	0.00400
M1→M2	Answer Relevance	0.1365	[0.1177, 0.1554]	0.00400
M2→M3	Grounding	0.0505	[0.0466, 0.0544]	0.00400
M3→M4	Grounding	0.0007	[0.0006, 0.0009]	0.00400

주: 신뢰구간은 개별 비교의 95% bootstrap 구간이며 동시 신뢰구간이 아니다. p값은 40개 단계 비교의 Holm 보정값이다. 표에는 유의한 비교를, 상세 분석 자료에는 비유의 비교를 포함한 전체 결과를 제시하였다.

M1→M2에서 두 평가셋의 Recall@10, Grounding, Answer Relevance는 유의하게 증가했지만 External MRR@10은 감소하였다. M2→M3에서는 두 평가셋의 Grounding 증가와 Internal Answer Relevance 감소가 유의하였다. M3→M4의 Grounding 증가는 약 0.0007로 작았으며 M4→M5의 차이는 유의하지 않았다. 비유의 결과를 효과 부재나 동등성의 증거로 해석하지 않는다.

Grounding의 작은 증가는 법률적 정확성이나 큰 실질적 개선을 뜻하지 않는다. 누적 단계 비교와 각 기능을 단독 제거한 구성요소 실험도 구분하였다. Graph 네 조건의 평균은 Table 15·16에, 30개 대비 중 Holm 보정 후 유의한 14개 결과는 Table 17·18에 제시하였다.

Table 15. Graph Component Means (Internal)

Metric	OFF	SIMILAR	TYPED	BOTH
Target document recall	0.1389	0.4722	0.5556	0.5556
Target chunk recall	0.0278	0.1944	0.1389	0.1389
Path coverage	0.0000	0.4722	0.5556	0.5556

Table 16. Graph Component Means (External)

Metric	OFF	SIMILAR	TYPED	BOTH
Target document recall	0.0563	0.3625	0.3250	0.3250
Target chunk recall	0.0344	0.1844	0.1062	0.1062
Path coverage	0.0000	0.3500	0.2938	0.3063

Table 17. Significant Graph Component Contrasts (Internal; Holm family=30)

Contrast	Metric	Δ	95% CI	Holm p
OFF→SIMILAR	Path coverage	0.4722	[0.2778, 0.6667]	0.01520
OFF→TYPED	Target document recall	0.4167	[0.2500, 0.6111]	0.02520
OFF→TYPED	Path coverage	0.5556	[0.3889, 0.7222]	0.00300
OFF→BOTH	Target document recall	0.4167	[0.2500, 0.6111]	0.02890
OFF→BOTH	Path coverage	0.5556	[0.3889, 0.7222]	0.00400

Table 18. Significant Graph Component Contrasts (External; Holm family=30)

Contrast	Metric	Δ	95% CI	Holm p
OFF→SIMILAR	Target document recall	0.3063	[0.2313, 0.3875]	0.00300
OFF→SIMILAR	Target chunk recall	0.1500	[0.1000, 0.2031]	0.00300
OFF→SIMILAR	Path coverage	0.3500	[0.2750, 0.4313]	0.00300
OFF→TYPED	Target document recall	0.2687	[0.2000, 0.3438]	0.00300
OFF→TYPED	Target chunk recall	0.0719	[0.0469, 0.1000]	0.00300
OFF→TYPED	Path coverage	0.2938	[0.2188, 0.3688]	0.00300
OFF→BOTH	Target document recall	0.2687	[0.1938, 0.3438]	0.00300
OFF→BOTH	Target chunk recall	0.0719	[0.0468, 0.1000]	0.00300
OFF→BOTH	Path coverage	0.3063	[0.2313, 0.3812]	0.00300

Graph의 결합 조건은 단독 조건보다 평균이 항상 높지는 않아 결합의 우월성을 단정할 수 없다. Path coverage는 정답 경로 식별자의 회수율이며 관계 의미나 multi-hop 추론의 정확도가 아니다. 비유의 결과를 포함한 전체 30개 검정은 상세 분석 자료에 제시하였다.

3.5 LLM-as-a-Judge Results

Table 19는 M5 답변의 사후 Judge 결과이다. 최종 Judge 평가의 유효 응답은 Internal 144건과 External

www.kci.go.kr

720건이며, Judge 처리 오류는 없었다. Hallucination Risk≥0.5인 질문은 각각 0개와 9개였다. 근거 일치성·충실성·답변 관련성은 높았지만 완전성의 전체 평균은 0.8457이었다. 전체 평균은 두 평가셋의 문항 수로 가중한 값이다. 일곱 항목은 구분하여 해석하며 단계 간 우열이나 전문가 검증의 근거로 사용하지 않는다.

Table 19. Post-hoc Judge Results for M5 Answers

Metric	Internal M5	External M5
Evaluated answers	144	720
Judge errors	0	0
Unsupported claims	13	111
Questions with unsupported claims	6	37
High-risk questions	0	9
Groundedness	0.9483	0.9445
Faithfulness	0.9638	0.9590
Answer Relevance	0.9510	0.9427
Legal Correctness	0.9608	0.9572
Citation Accuracy	0.9497	0.9415
Completeness	0.8476	0.8453
Hallucination Risk	0.0453	0.0543

주: 품질 점수는 0~1 척도이며 각 평가셋의 문항 평균이다. High-risk questions는 Hallucination Risk≥0.5인 문항 수이다. Unsupported claims는 Judge가 지적한 주장 건수, Questions with unsupported claims는 해당 주장이 있는 문항 수이다.

3.6 Question-Type Analysis

Table 20은 Internal M5의 여덟 질문 유형별 결과이다. 각 유형은 18문항이며, 동일한 질문을 기준으로 검색 지표, 자동 답변 지표 및 Judge 점수를 함께 제시하였다.

Table 20. Internal-test M5 Performance by Question Type

Question type	Retrieval			Auto proxy		LLM judge		
	R@10	MRR@10	nDCG@10	Grd.	Rel.	Grd.	Rel.	Risk
Case / Respondent	0.9815	0.9722	0.9417	0.8917	0.9616	0.9578	0.9667	0.0422
Violation Fact	0.7963	0.4594	0.5411	0.8287	0.6181	0.9361	0.9444	0.0417
Applicable Law	0.8704	0.3919	0.5358	0.8898	0.4353	0.9611	0.9806	0.0406
Legal Reasoning	0.8704	0.3991	0.5437	0.8770	0.5184	0.9517	0.9667	0.0456
Sanction / Penalty	0.9259	0.4306	0.6003	0.8811	0.6575	0.9556	0.9583	0.0444
Similar Case	0.8333	0.4444	0.5555	0.8672	0.5600	0.9306	0.9333	0.0528
Temporal / Amendment	0.8778	0.5463	0.6797	0.8846	0.6030	0.9333	0.9056	0.0556
Complex Legal	0.8111	0.5357	0.6285	0.8875	0.5158	0.9600	0.9528	0.0400

주: R@10 = Recall@10. Grd. = Grounding, Rel. = Relevance. LLM judge 점수는 사후(post-hoc) 평가 결과이다. 각 지표는 해당 유형의 질문별 평가값을 산술 평균한 것으로, 결측값과 Judge 오류에 따른 제외 문항은 없다.

사건 및 피심인 확인형의 Recall@10은 0.9815로 가장 높고 위반행위 사실 확인형은 0.7963으로 가장 낮았다. 적용 법조항 확인형의 Answer Relevance는 0.4353이었다. 시점·법령 버전의 보완 과제는 V장에 제시하였다.

정형 질문에서 자동추정 유형은 정답 유형과 모두 일치하여 두 평가셋의 정확도와 macro-F1은 1.0000이었다. 같은 통제 실험의 자동추정·정답유형 조건에서는 검색 순서와 지표도 같았다. 동일한 메타데이터를 고정한 뒤 Internal M5에 두 조건을 각각 144문항씩 적용하여 검색·재검색·답변·평가지표의 일치를 확인하였다. 이 결과는 정형 질문과 고정 환경에 한정되며 자유로운 표현이나 복합 의도의 일반화 성능을 뜻하지 않는다.

Table 21·22의 Correct는 정답유형 조건이며 Wrong-7은 질문별로 나머지 7개 유형을 적용한 결과의 평균이다. Δ는 Wrong-7−Correct로 계산하였다. M5의 Recall@10, nDCG@10, MRR@10은 Internal 0.8708, 0.6283, 0.5225, External 0.8493, 0.6486, 0.5603이다. 이 세 지표는 단계별 평가와 Table 21·22의 Correct 조건에서 질문별로 일치하였으며, Internal의 단계별 결과는 Table 10에 제시하였다. 유형 오류의 영향은 같은 고정 환경의 Correct와 Wrong-7을 비교하여 해석하며, 16개 유형 통제 비교와 40개 단계 비교의 검정 범위는 구분한다.

Table 21. Question-Type Error Sensitivity (Internal-test)

Model	Metric	Correct	Wrong-7	Δ	95% CI	Holm-adjusted p
M2	Recall@10	0.8560	0.7995	-0.0565	[-0.0874, -0.0271]	0.00200
M2	nDCG@10	0.6229	0.5791	-0.0437	[-0.0633, -0.0244]	0.00160
M2	MRR@10	0.5212	0.4970	-0.0241	[-0.0368, -0.0125]	0.00160
M2	Slot coverage proxy	0.8994	0.8772	-0.0222	[-0.0361, -0.0110]	0.00160
M5	Recall@10	0.8708	0.8087	-0.0622	[-0.0957, -0.0297]	0.00200
M5	nDCG@10	0.6283	0.5826	-0.0457	[-0.0668, -0.0253]	0.00160
M5	MRR@10	0.5225	0.4983	-0.0242	[-0.0366, -0.0126]	0.00160
M5	Slot coverage proxy	0.9428	0.9247	-0.0181	[-0.0282, -0.0089]	0.00160

주: Internal-test n=144, External-test n=720. Correct는 정답유형 조건의 문항 평균이고 Wrong-7은 질문별 오유형 7개 평균의 문항 평균이다. Δ=Wrong-7−Correct. CI는 질문 단위 bootstrap의 개별 95% 신뢰구간이며 다중비교 보정 신뢰구간은 아니다. p값은 16개 주 비교에 대한 Holm 보정값이다. Slot coverage proxy는 정답유형으로 고정한 요구 슬롯의 사후 자동 충족률이다. 평균,차이,신뢰구간은 소수 넷째 자리, p값은 소수 다섯째 자리까지 표시하였다.

www.kci.go.kr

Table 22. Question-Type Error Sensitivity (External-test)

Model	Metric	Correct	Wrong-7	Δ	95% CI	Holm-adjusted p
M2	Recall@10	0.8405	0.8163	-0.0242	[-0.0359, -0.0123]	0.00160
M2	nDCG@10	0.6455	0.6288	-0.0168	[-0.0240, -0.0096]	0.00160
M2	MRR@10	0.5602	0.5559	-0.0042	[-0.0074, -0.0011]	0.01380
M2	Slot coverage proxy	0.8613	0.8565	-0.0048	[-0.0080, -0.0017]	0.00840
M5	Recall@10	0.8493	0.8200	-0.0293	[-0.0421, -0.0171]	0.00160
M5	nDCG@10	0.6486	0.6298	-0.0188	[-0.0260, -0.0116]	0.00160
M5	MRR@10	0.5603	0.5558	-0.0045	[-0.0076, -0.0012]	0.01380
M5	Slot coverage proxy	0.9357	0.9223	-0.0135	[-0.0188, -0.0085]	0.00160

질문유형을 잘못 지정했을 때 두 평가셋의 네 지표는 평균적으로 감소했으며, 16개 비교 모두 Holm 보정 후 통계적으로 유의하였다($p \leq 0.01380$). Recall@10은 M2와 M5에서 각각 Internal 0.0565와 0.0622, External 0.0242와 0.0293만큼 낮아졌다. M5의 필수 근거 항목 충족률인 슬롯 Coverage도 Internal 0.9428→0.9247, External 0.9357→0.9223으로 감소하였다. 이는 질문유형이 검색 경로와 근거 확보에 영향을 준다는 결과이다. 다만 모든 질문에서 성능이 낮아진 것은 아니며, M5가 M2보다 유형 오류에 더 민감하다고 단정할 수는 없다.

이 실험은 유형을 의도적으로 바꾸었을 때의 영향을 확인한 것으로, 실제 사용 중 오분류가 얼마나 발생하고 성능이 얼마나 낮아지는지를 측정한 것은 아니다. 슬롯 Coverage 역시 정답유형에 지정된 근거 항목의 충족 여부를 자동으로 평가한 값이며, 법률적 정확성이나 정답 근거의 완전성을 보장하지 않는다.

3.7 Runtime Efficiency and Evaluation Cost

Table 23은 갱신 M5의 검색, 재검색, 추출형 답변 구성 시간과 별도 Judge, entailment 시간이다. 검색 반복은 초기 검색과 재검색의 합이다. 질의 임베딩을 재사용한 고정 환경의 단일 실행이므로 과거 실행 대비 속도 개선이나 반복 실행의 시간 안정성을 입증하지 않는다.

Table 23. Descriptive Runtime and Separate Post-hoc Evaluation Time for M5

Item	Internal-test	External-test
Questions (n)	144	720
Mean runtime (s)	1.918	3.959
p95 runtime (s)	2.770	6.556
Mean retrieval iterations	2.236	2.761
Mean retries	1.236	1.761
Mean post-hoc Judge time (s)	3.077	3.095
Mean post-hoc entailment time (s)	7.909	8.092
Generation LLM calls	0	0
Generation tokens	0	0
Valid Judge / entailment responses	144 / 144	720 / 720

추출형 답변의 생성 LLM 호출과 토큰은 0이다. 사후 Judge, entailment는 각각 864개 유효 응답과 실제 호출별 토큰, 응답 기록을 보존하였다. 근거번호 오류의 재시도 이력도 별도 남겼다. API 사용량과 runtime 로컬 연산은 구분하며 단가에 따른 금액 환산이나 반복 실행 기반 운영비용은 이번 표에 포함하지 않았다.

V. Conclusions

본 연구는 의결서의 문서 구조, 법률 메타데이터, 그래프 관계, 규칙 기반 재검색을 결합한 Hybrid Retrieval and Evidence-Aware Extractive QA 프레임워크를 제안하였다. 공통 설정의 갱신 평가에서 M1→M5의 Recall@10은 Internal 0.7380→0.8708, External 0.6861→0.8493이었으며 Grounding은 각각 0.6053→0.8759와 0.4855→0.7726으로 증가하였다.

섹션 구조를 반영한 M1→M2에서 가장 큰 개선이 나타났으며, Graph는 유사사건의 관계,경로 근거를 보완하였다. M5는 단일 제어기가 근거 충분성을 판단하고 필요한 경우 재검색하도록 구성하였다. 각 기능의 기여는 Graph OFF/ON, 구성요소별 비교 및 단계별 대응 분석으로 평가하였다. 다만 M3,M4의 일반 순위지표는 동일하였으므로 Graph의 기여는 관계,경로 근거 회수로 한정한다. 그래프 규모 집계와 근거 연결의 자동 점검 역시 구조적 추적성을 확인한 결과이며, 관계 내용의 정확성이나 multi-hop 추론 성능을 검증한 것은 아니다.

본 연구는 단일 기관 문서, 규칙 기반 추출, 정형 질문과 내부 기준선 M1에 한정된다. 동일 문서·질문·정답·평가 기준의 외부 방법 비교, 2인 이상 전문가의 동일 답변 독립 평가와 일치도 분석이 필요하다. 관계 의미의 정확성, 별도 multi-hop·hop별 성능, 표현 변형·복합 의도의 오분

www.kci.go.kr

류 영향도 후속 검증 대상이다. 본 평가의 실제 적용 설정과 점수 계산 방법은 Table 8에 제시하였으며, 재현에 필요한 소스·설정·질문·인덱스의 스냅샷과 해시, 평가 지침, 입력·답변, 실제 응답 모델과 개별 호출 기록을 보존하였다. 질문 생성·검수의 상세 절차와 반복 실행 기반 운영비용은 추가 보완 과제로 남는다.

시계열·법령개정형 질의는 이미 수집한 법령자료에 더해 의결서만으로 확인하기 어려운 행위 시점, 법령·고시의 시행일 및 부칙 근거를 확보하는 것이 우선이다. 이를 바탕으로 Temporal Legal Graph를 갱신하고, 동일 질문군에서 시점 정보의 정확성, 조문 버전 일치도 및 관계 근거 회수 성능을 비교할 계획이다. 추가 자료 수집과 그래프 갱신은 후속 연구에서 수행하며, 성능 향상 여부도 갱신 후 평가로 확인하고자 한다.

REFERENCES

[1] N. Guha, J. Nyarko, D. E. Ho, C. Ré, A. Chilton, A. Narayana, A. Chohlas-Wood, A. Peters, B. Waldon, D. N. Rockmore, D. Zambrano, et al., “LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models,” Advances in Neural Information Processing Systems, Vol. 36, pp. 44123-44279, 2023.

[2] M. Dahl, V. Magesh, M. Suzgun, and D. E. Ho, “Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models,” Journal of Legal Analysis, Vol. 16, No. 1, pp. 64-93, 2024.

[3] N. Pipitone and G. H. Alami, “LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain,” arXiv preprint arXiv:2408.10343, 2024.

[4] Korea Fair Trade Commission, “Rules on the Operation of Meetings and Case Procedures of the Korea Fair Trade Commission,” KFTC Notice No. 2026-17, Sept. 28, 2026. [Online]. Available: https://www.law.go.kr/LSW/admRulInfoP.do?admRulSeq=2100000285468 (in Korean).

[5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Advances in Neural Information Processing Systems, Vol. 33, pp. 9459-9474, 2020.

[6] X. Wang, Z. Wang, X. Gao, F. Zhang, Y. Wu, Z. Xu, T. Shi, Z. Wang, S. Li, Q. Qian, R. Yin, C. Lv, X. Zheng, and X. Huang, “Searching for Best Practices in Retrieval-Augmented Generation,” Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 17716-17736, 2024.

[7] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson, “From Local to Global: A Graph RAG Approach to Query-Focused Summarization,” arXiv preprint arXiv:2404.16130, 2024.

[8] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, “ReAct: Synergizing Reasoning and Acting in Language Models,” Proceedings of the 11th International Conference on Learning Representations, 2023.

[9] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, “Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection,” Proceedings of the 12th International Conference on Learning Representations, 2024.

[10] S. E. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” Foundations and Trends in Information Retrieval, Vol. 3, No. 4, pp. 333-389, 2009. DOI: 10.1561/1500000019.

[11] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, “Dense Passage Retrieval for Open-Domain Question Answering,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pp. 6769–6781, Nov. 2020. DOI: 10.18653/v1/2020.emnlp-main.550.

[12] G. V. Cormack, C. L. A. Clarke, and S. Büttcher, “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, pp. 758- 759, 2009. DOI: 10.1145/1571941.1572114.

[13] Y. Feng, C. Li, and V. Ng, “Legal Case Retrieval: A Survey of the State of the Art,” Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 6472-6485, 2024. DOI: 10.18653/v1/2024.acl-long.350.

[14] A. Joshi, A. Sharma, S. K. Tanikella, and A. Modi, “U-CREAT: Unsupervised Case Retrieval Using Events Extraction,” Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, pp. 13899-13915, 2023. DOI: 10.18653/v1/2023.acl-long.777.

[15] H. Li, Q. Ai, J. Chen, Q. Dong, Y. Wu, Y. Liu, C. Chen, and Q. Tian, “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, pp. 1035-1044, 2023. DOI: 10.1145/3539618.3591761.

[16] V. Malik, R. Sanjay, S. K. Guha, A. Hazarika, S. K. Nigam, A. Bhattacharya, and A. Modi, “Semantic Segmentation of Legal Documents via Rhetorical Roles,” Proceedings of the Natural Legal Language Processing Workshop 2022, pp. 153-171, 2022. DOI: 10.18653/v1/2022.nllp-1.13.

[17] P. Kalamkar, A. Agarwal, A. Tiwari, S. Gupta, S. Karn, and

www.kci.go.kr

V. Raghavan, “Named Entity Recognition in Indian Court Judgments,” Proceedings of the Natural Legal Language Processing Workshop 2022, pp. 184-193, 2022. DOI: 10.18653/v1/2022.nllp-1.15.

[18] K. Sharma, P. Kumar, and Y. Li, “OG-RAG: Ontology-Grounded Retrieval-Augmented Generation for Large Language Models,” Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 32962–32981, Association for Computational Linguistics, 2025. DOI: 10.18653/v1/2025.emnlp-main.1674.

[19] A. Hogan, E. Blomqvist, M. Cochez, C. d’Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann, “Knowledge Graphs,” ACM Computing Surveys, Vol. 54, No. 4, Article 71, pp. 1-37, 2021. DOI: 10.1145/3447772.

[20] A. Singh, A. Ehtesham, S. Kumar, and T. T. Khoei, “Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG,” arXiv preprint arXiv:2501.09136, 2025.

[21] T. Nguyen, P. Chin, and Y.-W. Tai, “MA-RAG: Multi-Agent Retrieval-Augmented Generation via Collaborative Chain-of-Thought Reasoning,” arXiv preprint arXiv:2505.20096, 2025.

[22] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, “Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions,” Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 10014-10037, 2023. DOI: 10.18653/v1/2023.acl-long.557.

[23] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, “Corrective Retrieval Augmented Generation,” arXiv preprint arXiv:2401.15884, 2024.

Authors



Tae-Kyu Hwang received his B.S. and M.S. degrees in Computer Engineering from Chung-Ang University in 1998 and 2000, respectively. He is currently pursuing a Ph.D. at the Graduate School of Business IT, Kookmin University. His research interests include natural language processing, legal artificial intelligence, retrieval-augmented generation (RAG), knowledge graphs, multi-agent systems, and intelligent document understanding based on large language models (LLMs).



Namgyu Kim received the B.S. in Computer Engineering from Seoul National University in 1998, M.S. and Ph.D. degrees in Management Engineering from KAIST, Korea, in 2000 and 2007, respectively. Dr. Kim joined the faculty of the School of Management Information Systems at Kookmin University, Seoul, Korea, in 2007. He served as the Dean of the Graduate School of Business IT at Kookmin University and is currently a professor at the Business IT. He is interested in text mining, LLMs, and multi-agent systems.

www.kci.go.kr