

대학 규정집 기반 RAG 질의응답 시스템에 대한 탐색적 연구

(An Exploratory Study on a RAG-based Question-Answering System for University Regulations)

이상훈*, 이찬명**, 임희정***

(Sang-Hun Lee*, Chan-Myeong Lee**, Hoi-Jeong Lim***)

요약

본 연구는 대학교 규정집 검색 시스템이 기존의 키워드 검색의 한계를 극복하기 위해, 최신 규정 데이터를 반영한 검색 증강 생성(Retrieval-Augmented Generation, RAG) 기반 질의응답 시스템을 제안한다. 검색 단계에서는 벡터 기반 검색과 키워드 기반 검색을 결합한 하이브리드 검색과 재순위화 기법을 적용하였으며, 생성 단계에서는 규정 출처를 명시하도록 제한하는 근거 인용형 프롬프트를 통해 답변의 신뢰성을 확보하였다. RAGAS 기반의 정량 평가를 통해 총 30건의 대표 질의에 대해 평균 Answer Relevancy 0.87, Faithfulness 0.91을 기록하여, 제안 시스템이 규정에 근거한 신뢰성 있는 응답을 생성할 수 있음을 확인하였다.

■ 중심어 : 검색 증강 생성(RAG) ; 대학 규정집 ; 의미론적 청킹 ; 하이브리드 검색 ; RAGAS(Retrieval-Augmented Generation Assessment)

Abstract

This study proposes a Retrieval-Augmented Generation (RAG) based QA system to overcome the limitations of conventional keyword-based search in university regulation retrieval. The system integrates a hybrid search strategy with re-ranking to enhance retrieval precision and employs citation-based prompting to ensure response reliability. Quantitative evaluation using the RAGAS framework on 30 representative queries yielded an average Answer Relevancy of 0.87 and Faithfulness of 0.91. These results demonstrate that the proposed system generates reliable, regulation-grounded responses with enhanced accessibility.

■ keywords : Retrieval-Augmented Generation (RAG) ; University Regulations ; Semantic Chunking ; Hybrid Search ; RAGAS (Retrieval-Augmented Generation Assessment)

1. 서론

대학 규정집은 다음과 같은 도메인적 특성을 지닌다. 첫째, 규정은 학칙, 시행규칙, 지침 등 다수의 문서로 분산되어 있으며, 동일한 주제에 대한 규정이 여러 문서와 조항에 걸쳐 중복·분산되어 존재하는 경우가 있다. 둘째, '수강', '휴학', '복학', '졸업', '징계'와 같이 학생과 교직원에게 반

복적으로 질의되는 개념들이 조항마다 상이한 표현으로 기술되어 있어, 단순 키워드 일치 기반 검색에서는 의미적으로 관련된 규정을 충분히 회수하기 어렵다. 셋째, 규정 문서는 장-절-조의 계층 구조와 함께 개정 이력, 부칙, 참조 조항 등 문맥적 연결성이 강하게 요구되는 특성을 지닌다.

이러한 특성으로 인해 대학 규정집 도메인에서는 단순 문서 검색이나 LLM 단독 질의응답 방

* 준회원, 전남대학교 데이터사이언스대학원

** 준회원, 전남대학교 통계학과

*** 정회원, 전남대학교 데이터사이언스대학원, 공공데이터분석센터

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(RS-2023-00242528, 50%)과 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-지역지능화혁신인재양성사업의 지원을 받아 수행된 연구임(IITP-2026-RS-2022-00156287, 50%)

접수일자 : 2026년 01월 02일

게재확정일 : 2026년 01월 29일

수정일자 : 2026년 01월 23일

교신저자 : 임희정 e-mail : hjlim@jnu.ac.kr

식만으로는 사용자의 질의 의도를 정확히 반영한 응답을 제공하는 데 한계가 존재한다. 특히 규정 해석에 있어서는 응답의 자연스러움보다도 근거 조항의 명확성, 출처 추적 가능성, 그리고 규정 원문과의 정합성이 중요하게 요구된다.

국내 대학들은 방대한 규정을 운영하고 있으나, 이러한 규정들을 사용자 관점에서 효율적으로 탐색하고 활용하기에는 여전히 어려움이 존재한다[1]. 현재 다수의 대학에서 운용 중인 규정 검색 시스템은 사용자의 질의 의도나 의미적 맥락을 충분히 반영하지 못하고, 주로 텍스트의 형태적 일치 여부에 기반한 단순 키워드 매칭 방식에 의존하고 있다. 이로 인해 사용자가 원하는 정보를 정확히 찾아내거나 관련 규정을 종합적으로 이해하는 데 한계가 있다[2].

사용자는 필요한 정보를 얻기 위해 다수의 규정 문서를 개별적으로 열람하며 관련 조항을 스스로 탐색해야 하고, 이는 정보 탐색 과정에서 시간 소모와 인지적 부담을 증가시킨다. 그 결과, 기존 규정 검색 시스템은 사용자의 질의에 대해 구체적인 해답을 제공하기보다는 관련성이 불명확한 문서 목록을 제시하는 데 그치며, 학사·행정 정보 접근의 효율성과 활용도를 저해하는 요인으로 작용하고 있다.

이러한 정보 접근성의 한계를 개선하기 위해, 대규모 언어 모델(Large Language Model, LLM)의 언어 이해 능력과 외부 지식 검색을 결합한 RAG(Retrieval Augmented Generation) 기법의 도입이 요구된다. 일반적인 LLM은 방대한 지식을 학습했으나, 특정 도메인 데이터는 학습되어 있지 않아 답변이 불가능하거나, 사실이 아닌 정보를 마치 사실인 것처럼 생성하는 환각 현상을 일으킬 위험이 있다[3].

RAG(Retrieval-Augmented Generation) 기술은 대규모 언어 모델이 지니는 지식 한계를 보완하기 위한 방법으로, 사용자의 질문이 입력되면 외부 지식베이스로부터 관련 문서를 실시간으로 검색하고, 해당 문서를 문맥 정보로 활용하여 응

답을 생성하는 구조를 가진다. 이러한 방식은 최신성 및 근거 기반 응답이 요구되는 도메인 지식 환경에서 대규모 언어 모델을 단독으로 활용할 때 발생할 수 있는 부정확성 문제를 완화하는 데 효과적인 대안으로 주목받고 있다[4].

본 연구의 목적은 대학 규정집이 지니는 도메인 특성을 고려하여 RAG 기반 질의응답 시스템을 구현하고, 이를 통해 규정 정보에 대한 접근성과 응답 품질의 개선 가능성을 분석하는 것이다. 특히 문서의 분산성, 표현의 다양성, 장-절-조로 구성된 계층적 구조라는 대학 규정집의 특성을 반영하여 규정 문서를 조항 단위로 구조화하고, 계층 및 페이지 정보를 메타데이터로 결합한 지식베이스를 구축하여 의미 기반 검색과 키워드 기반 검색을 결합한 하이브리드 검색 전략과, 검색된 규정 문맥 내에서만 응답을 생성하도록 제한하는 근거 인용형 프롬프트를 적용하였다. 최종적으로 RAGAS 기반의 정량적 평가를 통해 기존 키워드 중심 검색 방식 대비 제안 시스템의 우수성을 실증적으로 검증하고자 한다.

II. 관련 연구

1. RAG의 개념 및 패러다임 변화

RAG(Retrieval-Augmented Generation)는 외부 지식베이스에서 관련 문서를 검색해 대규모 언어 모델의 입력 문맥으로 제공함으로써, 근거 기반 응답 생성을 가능하게 하는 자연어 처리 프레임워크이다. 이 기술은 2020년 Facebook AI Research에 의해 제안되었으며, LLM이 최신 정보를 반영하지 못하거나 사실과 다른 내용을 생성하는 환각 문제를 완화하기 위한 대안으로 주목받았다[5]. RAG는 LLM이 사전 학습하지 못한 분야에 대해서도 추가 학습 없이 외부 지식을 활용하여 고품질의 응답을 생성할 수 있도록 지원한다. 이에 따라 의료, 법률, 금융 등 다양한 전

문 도메인에서의 활용 가능성이 확대되고 있으며, 최근에는 RAG를 적용한 LLM의 한계를 개선하고 내부 지식과 외부 정보를 효과적으로 통합하기 위한 연구가 활발히 진행되고 있으며[6]. 검색 단계와 생성 단계의 두 과정으로 구성된다. 먼저 검색 단계에서는 사용자 질의를 벡터화하여 문서 임베딩과의 유사도를 계산하고, 상위 K개의 관련 문서를 회수한다. 이후 생성 단계에서는 검색된 문서를 프롬프트에 통합하여 언어모델에 입력함으로써, 외부 지식에 기반한 최종 응답을 생성한다[7].

2. 도메인 특화 법령과 규정 질의응답 시스템 선행 연구

법령과 규정 문서는 조문 중심의 계층 구조와 도메인 특화 용어 및 판례, 부칙, 참조조항 등 정확한 근거 인용이 요구된다. 이러한 특성으로 인해 단순 키워드 검색이나 외부 근거 없이 LLM이 답변을 생성하는 질의응답 방식에는 한계가 있다. 이에 최근 연구들은 RAG를 중심으로 문서를 구조화하고 검색과 생성 과정에서 답변의 근거를 명확히 제시하도록 하였다. Seo 등[8]은 형사법 영역에서 20개 핵심 법률(조문 12,183)과 판례 10,707건(1975 - 2023)을 구축하고, 조 - 항 - 호의 3계층 구조로 112,206개의 최소 단위로 분할하였다. 쿼리 확장(3~5개)과 Maximum Marginal Relevance (MMR) 검색을 적용한 결과, 227개 사례에서 5-shot 정확도 61.7%를 보고하였으나, 판례 선정 오류가 23.7%로 높아 환각 위험이 핵심 병목으로 확인되었다(법규 적용 오류 0.8%).

Shin 등[9]은 민법 문서를 장 - 절 - 조의 구조와 조문 간 참조 관계로 모델링하고, LangGraph 기반의 모듈형 RAG로 적용하여 질의 유형에 따라 처리 워크플로우를 동적으로 선택하는 방식을 제안하였다. 1,118개 조문과 150개 질의를 대상으로 실험을 수행하여 검색 정확도 88%, 응답 적절성 92%, MRR 0.84를 달성하였으며, 직접 조

문 질의에 대해서는 100%의 정확도를 보고하였다. 그러나 주요 한계로는 유사 표현에 대한 처리와 법령 개정 이력이 반영되지 않았다는 점이 제시되었다.

Seo 등[10]은 의미 기반 청킹과 E5 임베딩의 도메인 적응 파인튜닝이 검색 성능을 개선할 수 있음을 보였고, BM25와 파인튜닝된 E5 임베딩을 결합한 하이브리드 검색 방식이 검색과 생성 품질 평가에서 가장 균형 잡힌 성능을 제시하였다. 또한 RAGAS와 BERTScore 평가 지표를 통해 해당 방법의 효과를 정량적으로 검증하였다.

이들 선행연구를 종합하면, 도메인 특화 질의응답 시스템 설계의 핵심 원리로 네 가지 요소가 도출된다. 즉, 계층 구조를 반영한 단위 분할 및 메타데이터 설계, 의미 기반 검색과 키워드 기반 검색을 결합한 하이브리드 검색, 조문이나 판례를 정확히 인용할 수 있는 검증 메커니즘, 그리고 질의 유형에 따라 서로 다른 처리 단계를 적용하는 모듈화 전략이 그것이다. 따라서 규정 질의응답 시스템은 단순한 검색 성능 향상을 넘어 답변의 근거의 신뢰성까지 보장하도록 설계되어야 한다.

본 연구의 기여는 다음과 같다. 첫째, 국내 대학 규정집이라는 행정 문서 도메인을 대상으로 검색 증강 생성 기반 질의응답 시스템을 설계하고 구현하여, 기존 키워드 중심 검색 방식의 한계를 실증적으로 분석하였다.

둘째, HWPX 기반 규정 문서를 장-절-조 단위로 구조화하고 계층적인 메타데이터를 결합한 지식베이스 구축 방안을 제시하였다.

셋째, 하이브리드 검색과 근거 인용형 프롬프트를 적용해 규정 기반 응답의 신뢰성을 강화하고, 이를 정량 지표를 통해 검증하였다.

III. 연구 방법

1. 데이터 수집 및 전처리

지식베이스 구축을 위한 원천 데이터는 2025년

11월 24일 기준 전남대학교 규정집 홈페이지에 게시된 최신 규정 자료를 대상으로 하였다. 데이터는 홈페이지에서 제공하는 원문 파일을 한글 문서의 개방형 XML 표준 형식인 HWPX 형식으로 제공받아 수집하였다.

수집 대상은 전남대학교 학칙, 전남대학교 교학규정, 대학원 교학규정, 데이터사이언스대학원 교학규정의 총 4종 문서로, 대학의 조직 및 학사 운영, 학생 활동 등 핵심 행정 절차를 규율하는 상위 규정에 해당한다. 이들 문서는 질의응답 시스템에서 활용 빈도가 높을 것으로 판단되어 선정하였다. 수집 데이터의 상세 구성은 표 1에 제시하였다.

가. 문서의 구조화 및 정제

수집된 규정 문서는 텍스트, 표, 다양한 서식 요소가 혼재된 비정형 데이터로, RAG 시스템이 효과적으로 검색하고 LLM이 이해할 수 있는 형태로의 변환이 필요하다. 이를 위해 본 연구에서는 문서의 논리적 구조를 최대한 보존하면서 데이터 품질을 확보하는 3단계 전처리 파이프라인을 구축하였다.

표 1. 수집 데이터 목록 및 구성

Document Title	File format	Number of pages (based on main text)	Notes (Included content)
University Regulations	.hwp	20	Includes 10 Attachment
Academic Affairs Regulations	.hwp	10	Includes 3 Attachment
Graduate School Academic Affairs Regulations	.hwp	6	-
Academic Affairs Regulations for the Graduate School of Data Science	.hwp	4	-
Total (4 documents)	-	40	Total of 13 Attachment

(1) 지능형 문서 파싱

HWPX 문서에는 표, 다단 구성 등 복잡한 레이아웃 요소가 포함되어 있어 기존의 단순 텍스트 추출 방식으로는 문서 구조를 온전히 보존하는

데 한계가 있다. 이를 해결하기 위해 본 연구에서는 상용 문서 파싱 도구[11]를 활용하여 문서를 분석하고, 시각적 레이아웃 정보를 최대한 유지한 상태에서 Markdown 형식의 구조화된 텍스트로 변환하였다.

(2) 텍스트 정제 및 계층 구조 분석

파싱 과정에서 머리글과 바닥글 등 실제 규정 내용과 직접적인 관련이 없는 요소를 제거하였다. 이후 규정 문서를 조항(article) 단위로 분할하고, 장(chapter) - 절(section) - 조(article)로 구성된 계층 구조를 분석하여 각 조항이 속한 상위 구조 정보를 기록하였다. 이를 통해 검색 단계에서 조항 간의 계층적 연결 관계를 활용할 수 있으며, 답변 생성 시 규정의 논리적 흐름을 유지함으로써 보다 정확한 정보 제공이 가능하도록 하였다.

(3) 메타데이터 추출 및 구조화된 저장

각 텍스트 청크가 원문 문서의 어느 위치에 해당하는지 식별할 수 있도록 페이지 범위를 매핑하였다. 또한, 본문텍스트와 메타데이터를 분리하여 관리하기 위해 YAML Frontmatter 형식을 적용함으로써, 각 청크가 위치 및 계층 정보를 모두 포함하도록 구성하였다. 전처리 결과로 구축된 지식베이스의 구조적 특징은 표 2에, 실제 전처리가 완료된 샘플 데이터는 그림 1에 각각 제시하였다.

표 2. 문서 전처리 및 인덱싱 세부사항

Category	Processing Details	Purpose
Document Format	HWPX -> Markdown	Preserving layout information
Segmentation Unit	Article-level segmentation	Achieving semantic completeness
Refining Process	Removing noise from headers/footers...	Understanding the hierarchical structure (chapters/sections/articles)
Metadata	Page number, parent information, title	Using YAML frontmatter
Tracking Information	Original page number, chunk order	Facilitating source citation

```

---
page: 5
total_page: 19
chunk: "27/92"
chapter: "제3장"
chapter_title: "학사운영"
section: "제1절"
section_title: "학사일반"
article: "제23조"
article_title: "수업일수"
---
---text
제23조(수업일수) ① 수업일수는 매 학년도 30주(매 학기 15주) 이상으로 한다. 다만, 수업을 집중하여 실시할 경우 수업일수를 달리할 수 있다.
② 천재지변이나 그 밖의 부득이한 사유로 제1항의 수업일수를 충족할 수 없는 경우에는 매 학년도 2주의 범위 내에서 수업일수를 감축할 수 있다.
---
    
```

그림 1. 전처리가 완료된 샘플 데이터 예시

나. 메타데이터 설계 및 텍스트 청킹

구조화된 규정 문서를 벡터 데이터베이스에 저장, 관리하기 위해 AI 애플리케이션 구축용 오픈소스 플랫폼인 Dify를 활용하였다. Dify의 지식베이스(Knowledge Base)를 기반으로 벡터 인덱싱 파이프라인을 구성하였으며, 검색 효율성을 향상시키는 동시에 각 응답이 어떤 문서와 조항을 근거로 생성되었는지를 추적할 수 있도록 메타데이터 스키마와 청킹 전략을 설계하였다.

단순 텍스트 임베딩만으로는 대규모 언어 모델이 답변 생성 과정에서 장 - 절 - 조 단위의 구체적인 근거를 명확히 제시하기 어렵다. 이에 본 연구에서는 전처리 단계에서 각 조항의 계층 정보를 나타내는 메타데이터(chapter, section, article)를 정의하여 활용하였으며, 적용된 메타데이터의 상세 명세는 표 3에 제시하였다.

표 3. 시스템 구성 및 RAG 구현 세부 사항

key	Description
page	The original page number where the chunk is located
total_page	Total number of pages in the document
chunk	Chunk Index (Current Number / Total Number of Chunks)
chapter	Upper Layer Information 1 (Chapter Number)
chapter_title	Top-level heading 1
section	Upper Layer Information 2 (Section Number)
section_title	Top-level heading 2
article	Article Number
article_title	Article Title

규정집은 논리적 완결성이 중요한 문서이므로, 문장이 중간에 분절되거나 서로 다른 조항이 하

나의 청크에 혼합될 경우, 검색 및 응답 품질이 저하될 수 있다. 이를 방지하기 위해 본 연구에서는 물리적 길이 기준이 아닌 논리적 구조를 우선적으로 고려하는 의미론적 청킹 전략을 적용하였다.

구체적으로, 텍스트 전처리 단계에서 문서의 구조를 분석하여 모든 조항의 시작과 끝 경계를 식별하고, 각 경계에 특수 식별자를 삽입하였다. 이는 문서를 청킹할 때 단순한 글자 수 제한이나 문단 구분이 아닌, 조항 단위의 논리적 경계를 기준으로 분할이 이루어지도록 유도하기 위함이다.

이후 Dify의 청크 설정에서 삽입된 특수 식별자를 세그먼트 식별자로 지정하여 청킹을 수행하였다. 이를 통해 각 청크는 하나의 조항 본문과 해당 조항의 메타데이터가 결합된 독립적인 단위로 보존되며, 검색 시 잘린 문장이 아닌 완결된 조항 전체와 상세한 메타 정보가 LLM의 문맥(Context)으로 함께 제공되도록 구성하였다.

2. 벡터 저장소 구축

가. 임베딩 모델 선정 및 적용

국내 대학 규정 문서를 고차원 벡터로 변환하기 위해 한국어 행정 및 규정 텍스트의 의미 표현에 적합한 상용 임베딩 모델[12]을 활용하였다. 임베딩은 자연어 텍스트를 기계가 처리 가능한 벡터 표현으로 변환하여, 의미적으로 유사한 텍스트들이 벡터 공간상에서 서로 가깝게 배치되도록 하는 RAG 시스템의 핵심 기술이다. 해당 임베딩 모델은 입력 텍스트의 역할에 따라 질의(Query)와 문서(Passage)를 구분하여 처리하는 이원화된 아키텍처를 제공한다.

본 연구에서는 규정 조항 임베딩에 적합한 solar-embedding-1-large-passage 모델을 선정하여 적용하였다. 구축된 벡터 저장소 내에서 사용자의 질의와 가장 관련성이 높은 규정 조항을

탐색하기 위해 코사인 유사도(cosine similarity)를 사용하여 질의 벡터와 규정 조항 벡터 간의 의미적 유사도를 계산하였다. 코사인 유사도는 두 벡터 간 각도의 코사인 값을 기반으로 텍스트 간 의미적 유사도를 정량화하는 지표로, 벡터의 크기 차이에 영향을 받지 않는다는 장점을 가진다.

나. 벡터 DB 구성 및 벡터 인덱싱

고차원 벡터로 변환된 규정 데이터를 효율적으로 저장하고 검색하기 위해 벡터 데이터베이스를 구축하였다. 기존의 관계형 데이터베이스(RDBMS)가 주로 키워드 기반 검색에 의존하는 것과 달리, 벡터 데이터베이스는 대규모 문서 청크 집합에서 사용자 질의와 의미적으로 가장 유사한 콘텐츠를 고속으로 탐색하는 데 특화되어 있다.

이를 구현하기 위한 인프라로 Dify Knowledge Base를 활용하였다[13]. Dify는 RAG 파이프라인 전반을 시각적으로 관리할 수 있는 환경을 제공하며, 업로드된 문서를 구조화된 벡터 인덱스로 자동 변환한다. 이를 통해 정적인 사전 학습 데이터셋에 의존하는 대규모 언어 모델의 한계를 보완하고, 최신 규정 정보가 검색 및 응답 생성 과정에 반영될 수 있도록 하였다.

벡터 인덱싱 단계에서는 각 청크의 텍스트를 임베딩 모델을 통해 고차원 벡터로 변환한 후, 이를 벡터 데이터베이스에 저장하였다. 이후 사용자 질의가 입력되면 동일한 임베딩 모델을 사용하여 질의를 벡터화하고, 질의 벡터와 저장된 청크 벡터 간의 유사도를 계산하는 벡터 검색(Vector Retrieval)을 하였다.

벡터 검색은 자연어처리(NLP)와 머신러닝 분야에서, 입력된 데이터를 고차원 벡터 형태의 임베딩으로 변환하여 의미 기반의 검색을 수행하는 방식으로, 기존의 키워드 기반 검색의 한계를 보완한다[14].

이 과정에서 코사인 유사도에 기반한 벡터 유

사성 검색을 통해 질의와 의미적으로 가장 유사한 조항 청크들이 상위 후보로 반환되었으며, 해당 청크들은 대규모 언어 모델의 문맥(Context)으로 제공되어 최종 답변 생성을 위한 근거로 활용되었다. 이러한 벡터 DB 기반 인덱싱 및 검색 구조를 통해 단순 키워드 일치 여부를 넘어, 표현은 다르지만 의미적으로 유사한 규정 조항을 효율적으로 탐색할 수 있는 RAG 지식베이스를 구현하였다.

3. 시스템 설계 및 구현

가. 시스템 아키텍처

본 연구에서 제안하는 시스템은 대학 규정집 데이터의 구조적 특성을 반영하여 근거 기반 응답을 생성하기 위해, 문서 수집 및 적재(Ingestion) 파이프라인, 코어 RAG 엔진(Core RAG Engine), 사용자 인터페이스(User Interface)의 3계층 모듈로 구성하였다. 전체 시스템 아키텍처는 그림 2에 제시하였다.

문서 수집 및 적재 파이프라인은 규정 원문을 검색 및 생성에 적합한 지식베이스로 변환하는 전처리 계층이다. 입력으로는 규정 문서(예: HWPX 원문)를 사용하며, 문서 파싱 및 구조화 단계에서 장·절·조의 계층 구조를 식별하고 표 및 다단 구성 등 복합 레이아웃 요소를 포함한 내용을 추출·정제한 뒤 Markdown 형식으로 변환한다. 이후 정제된 텍스트를 조항 중심의 의미 단위로 청킹하고, 각 청크에 대해 임베딩을 생성하여 벡터 저장소에 저장 및 색인화함으로써 검색 가능한 규정 지식베이스를 구축한다.

코어 RAG 엔진은 LangChain 기반으로 질의 처리의 전 과정을 제어하는 핵심 모듈로, 검색 단계와 생성 단계로 구성된다. 먼저 하이브리드 검색 전략을 적용하여 의미 기반의 벡터 검색과 용어 기반의 BM25 키워드 검색을 결합함으로써 상위 K개의 후보 문서를 도출한다. 이후 Cross-Encoder 기반 재순위화(reranking)를 통

해 후보 문서들의 문맥 적합성을 정밀하게 평가하고, 답변 생성에 기여도가 높은 규정 조항을 최종적으로 선별한다. 다음으로 선택된 조항에 문서 식별자와 장·절·조 등 구조적 메타데이터를 태깅하여 프롬프트용 문맥을 체계적으로 구성한다. 마지막으로 시스템 프롬프트, 구조화된 문맥, 사용자 질의를 결합한 입력을 대규모 언어 모델에 제공하여 규정에 근거한 응답을 생성하며, 본 연구에서는 생성 모델로 상용 LLM(Solar-Pro 2)을 활용하였다.

사용자 인터페이스 계층은 Streamlit 기반의 웹 서비스로 구현되며, 사용자 인증(로그인)과 질의 입력 기능을 제공한다. 사용자가 자연어 질의를 입력하면 웹 서비스는 코어 RAG 엔진의 처리 결과를 전달받아 답변을 제시하고, 동시에 참조된 규정조항과 출처 메타데이터를 함께 제공한다. 이를 통해 사용자는 응답의 근거를 직접 확인할 수 있으며, 시스템 응답의 투명성과 신뢰성을 제고할 수 있다.

나. 검색 및 재순위화

사용자의 질의에 대해 적합한 규정 정보를 제공하기 위해, 1단계 하이브리드 검색과 2단계 Cross-Encoder 기반 재순위화로 구성된 검색 파이프라인을 구축하였다.

(1) 하이브리드 검색 알고리즘

1차 검색 단계는 벡터 기반 의미 검색과 키워드 기반 검색을 결합한 하이브리드 검색으로 구성하였다. 규정집 청크는 앞서 구축한 벡터 저장소에 임베딩 형태로 저장되며, 검색 시 Dify Knowledge Base를 외부 지식베이스로 활용하여 질의 - 조항 간의 의미 유사도를 계산하였다. 먼저 사용자의 자연어 질의는 임베딩 모델을 통해 고차원 벡터로 변환되고, 해당 벡터와 각 조항 임베딩 간의 코사인 유사도를 기반으로 의미 검색 점수를 산출하였다. 동시에 BM25 기반 키워드 검색을 수행하여 질의에 포함된 핵심 용어

와 조항 텍스트 간의 어휘적 일치 정도를 점수화하였다.

최종 검색 점수는 의미 검색 점수와 키워드 검색 점수를 가중 합(weighted sum) 방식으로 통합하여 계산하였다. 실험 결과, 의미 검색과 키워드 검색의 가중치를 각각 0.7과 0.3으로 설정했을 때 규정 용어의 정확한 매칭과 의미적 유사성을 균형 있게 반영하는 것으로 나타나, 이를 최종 설정값으로 채택하였다.

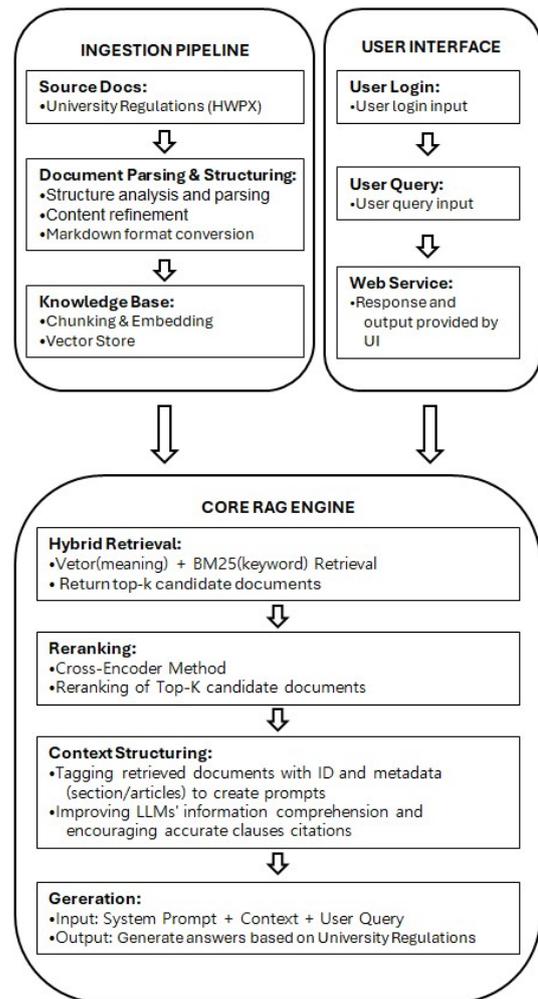


그림 2. RAG 시스템 아키텍처

최종 검색 점수는 의미 검색 점수와 키워드 검색 점수를 가중 합(weighted sum) 방식으로 통합하여 계산하였다. 실험 결과, 의미 검색과 키워드 검색의 가중치를 각각 0.7과 0.3으로 설정했을 때 규정 용어의 정확한 매칭과 의미적 유사성을 균형 있게 반영하는 것으로 나타나, 이를 최종

설정값으로 채택하였다.

Top-K 후보 개수는 K=3, 5, 6, 10에 대해 비교 실험을 수행한 결과, K=6일 때 규정 조항의 맥락을 충분히 포함하면서도 응답 시간을 과도하게 증가시키지 않는 최적의 균형점을 보였으며, 이에 따라 1차 검색 단계의 기본값으로 설정하였다. 또한 유사도 임계값(score threshold)을 0.1, 0.2, 0.3으로 비교한 결과, 0.2 미만의 문서는 실제 답변 생성에 기여도가 낮고 노이즈를 증가시키는 경향을 보여, 하한값을 0.2로 설정하여 낮은 관련도의 청크를 필터링하였다.

종합하면, 본 연구에서는 임베딩 기반 의미 검색 70%와 BM25 키워드 검색 30%의 가중치를 적용한 하이브리드 검색 전략을 사용하였으며, 상위 6개 문서(Top-K=6)를 1차 후보로 선정하고 유사도 점수가 0.2 미만인 문서를 제외하도록 설계하였다. 이를 통해 규정집 내에서 표현은 상이하더라도 의미적으로 유사한 조항을 안정적으로 1차 검색 후보로 추출할 수 있도록 하였다.

(2) Cross-Encoder 기반 재순위화

하이브리드 검색을 통해 추출된 상위 6개 후보 문서는 여전히 일부 노이즈를 포함할 가능성이 있으므로(Top-K=6), 2차 단계에서 Cross-Encoder 기반 재순위화를 수행하였다. 이를 위해 재순위화 모델로 bge-reranker-v2-m3를 적용하였다.

Cross-Encoder 방식은 질의와 문서를 각각 독립적으로 임베딩한 후 유사도를 계산하는 Bi-Encoder 방식과 달리, 질의와 문서를 하나의 입력 쌍으로 결합하여 두 텍스트 간의 관련도를 직접 추정하는 구조를 가진다. 본 연구에서는 각 후보 조항에 대해 사용자 질의와 조항 텍스트를 하나의 쌍으로 구성하여 Cross-Encoder에 입력하였으며, 1차 검색 단계에서 도출된 6개 문서를 대상으로 관련도 점수를 재계산하였다. 이후 재계산된 점수를 기준으로 상위 3개 문서만을 최종 컨텍스트 후보로 선택하여 검색 결과를 압축

하였다.

최종 컨텍스트로 선택할 문서 수는 다양한 설정에 따른 성능 비교 실험을 통해 결정하였다. 구체적으로 문서 수를 1개, 2개, 3개, 5개로 변화시키며 평균 응답 시간과 답변 정확도를 비교하였다. 실험 결과, 문서 수를 1개로 제한한 경우 응답 속도는 가장 우수하였으나 일부 질의에서 컨텍스트 부족으로 인한 정보 누락이 발생하였다. 문서 수를 2개로 설정한 경우 정확도는 일부 개선되었으나, 복수의 규정이 복합적으로 연관된 질의에서는 충분한 근거를 제공하지 못하는 사례가 관찰되었다. 반면 문서 수를 3개로 설정한 경우 평균 응답 시간과 답변 정확도 모두에서 가장 균형 잡힌 성능을 보였다. 문서 수를 5개로 증가시킨 경우 컨텍스트의 풍부도는 향상되었으나, 응답 시간이 증가하고 LLM 입력 길이가 과도하게 확장되는 문제가 나타났다.

이러한 실험 결과를 종합적으로 고려하여, 본 연구에서는 컨텍스트의 충분성과 응답 속도 간의 균형이 가장 우수한 문서 수인 3개를 최종 설정값으로 채택하였다. 재순위화 단계에서 최종적으로 선택된 3개의 문서에 대해서는 조항 번호, 문서 유형 등 관련 메타데이터를 포함하는 컨텍스트 포매팅을 수행하여 LLM 입력에 적합한 구조화된 텍스트 형태로 변환하였다. 이후 시스템 지시문, 재구성된 컨텍스트, 사용자 질의를 순차적으로 결합하여 프롬프트를 구성하고, 이를 선택된 대규모 언어 모델에 입력하여 최종 답변을 생성하였다.

이와 같은 하이브리드 검색과 Cross-Encoder 기반 재순위화를 결합한 2단계 검색 구조를 통해, 제안한 시스템은 키워드 일치 여부에만 의존하는 방식과 비교하여 규정 문서의 의미적 맥락을 보다 효과적으로 반영할 수 있었다. 또한 재순위화 과정을 통해 질의와의 관련성이 낮은 문서를 사전에 제거함으로써, 불필요한 컨텍스트 입력을 줄이고 답변 생성 단계의 효율성을 향상시켰다.

다. 프롬프트 엔지니어링 및 답변 생성 RAG 시스템의 성능은 대규모 언어 모델에 제공되는 문맥의 품질과 구조에 크게 의존한다. 기존 방식과 같이 검색된 텍스트를 단순히 나열하여 입력할 경우, 문서 간 경계가 명확히 구분되지 않아 LLM이 서로 다른 문서를 혼합하여 해석하거나, 답변의 출처를 정확히 식별하지 못하는 문제가 발생할 수 있다. 이러한 문제를 완화하기 위해 본 연구에서는 소스코드 수준에서의 전처리를 통해 문맥 구조화 단계를 설계·구현하였다.

(1) 메타데이터 기반 문맥 구조화

하나의 문서 청크는 YAML frontmatter 메타데이터 블록과 커스텀 태그로 구분된 본문 영역으로 구성되며, 이에 대한 예시는 그림 3에 제시하였다. 상단에는 검색 및 재순위화 점수, 원본 문서의 물리적 페이지 위치, 장·절·조에 해당하는 논리적 계층 정보가 포함된 YAML frontmatter가 배치되고, 하단에는 시작 및 종료 태그로 명확히 구분된 본문 영역이 위치한다.

본 연구에서는 이러한 구조를 활용한 메타데이터 기반 문맥 구조화 방식을 적용하여, 대규모 언어 모델이 본문을 처리하기 이전에 각 청크의 문맥적 위치와 참조 우선순위를 사전에 인식할 수 있도록 설계하였다. 특히 본문 영역은 경계 식별의 안정성을 확보하기 위해 커스텀 태그로 감싸고, 태그 헤더에는 핵심 메타데이터를 파이프(|) 기호로 구분하여 포함함으로써 문맥 해석 과정에서 출처 정보와 구조 정보를 동시에 참조할 수 있도록 하였다. 또한 종료 태그를 명시적으로 정의하여, 서로 다른 조항이 하나의 문맥으로 혼합되는 현상을 효과적으로 방지하였다.

```

...
doc_count: 51
rerank_model: bge-reranker-v2-m3-ko-q4_k_m
rerank_score: 0.2311
reference_score: 0.5939
word_count: 431
tokens: 647
source: 한국대학교 교육규정
source_link: https://rule.jnu.ac.kr/Rule/5ub.aspx?m=2&sub=2
category_name: 교육행정
parent_doc_name: 알 수 없음
attached_num: 알 수 없음
page: 2
total_page: 9
chapter: 제2장 입학관리
section: 제4절 입학관리
article: 제11조 입학관리
...
<< $1 | source=한국대학교 교육규정 | source_link=https://rule.jnu.ac.kr/Rule/5ub.aspx?m=2&sub=2 | category_name=교육행정 | parent_doc_name=알 수 없음 | attached_num=알 수 없음 | page=2 | total=9 | chapter=제2장 입학관리 | section=제4절 입학관리 | article=제11조 입학관리 | rerank_model=bge-reranker-v2-m3-ko-q4_k_m | rerank_score=0.2311 | reference_score=0.5939 >>
...
chapter: "제2장"
section: "입학관리"
section_title: "제4절"
section_subtitle: "대학 입학관리"
article: "제11조"
article_title: "대학 입학관리"
...
text
제11조(대학 입학관리) ① 대학은 입학하는 정해진 등록기간의 일정을 통하여 복학신청을 하고 등록을 마쳐 -야 한다. 그 기간 종료 후에는 복학을 허가하지 아니한다.
② 학기 개시 전 3주 이내의 전역이 가능한 교원은 입학 등록기간 중에 복학할 수 있다.
③ 재입학을 원하는 학생은 재입학원서를 정해진 기간 내에 소속대학(원)장에게 제출하여야 한다.
④ 학위를 원하는 학생은 학업일수를 수석 대학(원)장에게 제출하여야 한다.
...
<< <END $1 >>

```

그림 3. 문맥 구조화가 적용된 샘플 데이터 예시 (2) 근거인용 및 환각제어

시스템 프롬프트 단계에서는 답변의 근거 일관성과 신뢰성을 확보하기 위해 엄격한 그라운드잉(grounding) 전략을 적용하였다. 구체적으로, 프롬프트를 통해 대규모 언어 모델이 제공된 태그 내부의 텍스트만을 근거로 답변을 생성하도록 제한하였다. 또한 각 응답에 대해 논리적 계층 정보와 물리적 페이지 정보를 포함한 출처를 명시하도록 요구함으로써, 생성된 답변의 근거를 원본 규정 문서에서 직접 확인할 수 있도록 설계하였다.

아울러 규정 해석의 정확성을 제고하기 위해 추가적인 제약 조건을 설정하였다. 숫자, 기간, 자격 요건 등 정량적 정보가 포함된 규정에 대해서는 임의적인 요약이나 재해석을 제한하고 원문 표현을 유지하도록 하였으며, 검색 결과에 포함된 문서 간 내용이 상충하는 경우에는 재순위화 단계에서 산출된 관련도 점수를 반영하여 우선순위가 높은 문서를 기준으로 답변을 구성하도록 로직을 정의하였다. 또한 제공된 컨텍스트 내에서 질의에 대응하는 정보를 확인할 수 없는 경우에는 답변 생성을 시도하지 않고 해당 정보가 존재하지 않음을 명시적으로 응답하도록 설계함으로써, 부정확한 정보 생성 가능성을 최소화하였다.

(3) 웹 인터페이스 구현

구축된 RAG 시스템의 실사용 가능성을 확인하기 위해 Streamlit 기반의 웹 인터페이스를 구현하였다. 해당 인터페이스는 사용자가 자연어

로 질의할 수 있는 대화형 입력창과, 생성된 답변에 대한 근거 규정 문서를 함께 제시하는 구조로 설계되었다. 이를 통해 사용자는 답변 내용과 원문 규정 간의 대응 관계를 직접 확인할 수 있다. 구현된 웹 인터페이스의 예시는 그림 4에 제시하였다.



그림 4. 전남대학교 규정집 질의응답 화면

IV. 연구 결과 및 토론

1. 실험개요 및 환경

본 실험은 제안한 RAG 시스템의 응답 품질 경향과 실용적 타당성을 분석하기 위한 탐색적 평가로서, 총 30건의 테스트 데이터셋을 구성하고 생성된 답변만을 평가하는 reference-free evaluation 방식을 적용하였다. 평가는 검색 중강 생성(RAG) 시스템의 검색·생성 품질을 참조 정답 없이 평가하기 위해 제안된 RAGAS(Retrieval-Augmented Generation Assessment) 프레임워크를 통해 수행되었다 [15]. RAGAS는 LLM을 평가자로 활용하여 생성된 답변의 질의 적합성과 검색 문맥에 대한 근거 충실도를 각각 Answer Relevancy와 Faithfulness 지표로 정량적으로 측정하는 평가 방법이다.

본 연구에서는 최종 사용자가 체감하는 답변

품질을 중점적으로 분석하기 위해 위 두 지표를 핵심 평가 기준으로 선정하였다. 실험 시스템은 연산 효율성과 재현성을 고려하여 Dify 플랫폼을 기반으로 구성하였으며, LLM 및 임베딩 모델로는 Upstage API를 활용하였다.

가. 평가 데이터셋 구축

평가 데이터셋은 대학 규정집 4종 문서를 기반으로, 실제 대학 행정 환경에서 발생할 수 있는 질의 30건으로 구성하였다. 질의 생성에는 Google NotebookLM을 활용하였으며, 사람이 작성한 정답과의 직접 비교 대신 reference-free evaluation 방식을 적용하여 생성된 답변이 질의 의도를 충족하는지와 검색된 문맥에 충실한지를 평가하였다.

나. 평가 지표 선정

시스템 성능 평가는 LLM을 평가자로 활용하는 RAGAS 프레임워크를 이용하여 수행하였다. 평가자 모델로는 답변 생성에 사용된 모델과 동일한 Upstage Solar-Pro 2를 적용함으로써, 한국어 규정 문맥에 대한 이해 수준과 평가 기준의 일관성을 확보하였다. 본 연구에서는 최종 사용자가 체감하는 서비스 품질을 중점적으로 분석하기 위해 생성 단계 중심의 평가 지표인 Answer Relevancy와 Faithfulness를 선정하였다.

Answer Relevancy는 생성된 답변이 사용자의 질의 의도에 얼마나 부합하는지를 평가하는 지표이며, Faithfulness는 답변 내용이 검색된 컨텍스트에 근거하여 생성되었는지를 측정한다. 두 지표를 통해 검색과 생성 과정을 포함한 RAG 시스템의 End-to-End 응답 품질을 종합적으로 평가하였다.

다. 시스템 환경 및 파라미터 설정

실험 환경과 주요 모델 및 파라미터 설정은 표

4에 제시하였다. 대규모 언어 모델과 임베딩 모델은 Upstage API를 통해 제공받았으며, 규정집 지식베이스 구축과 벡터 인덱싱에는 Dify 플랫폼을 활용하였다. 검색 및 생성 파이프라인의 제어는 LangChain을 기반으로 수행하였고, 사용자 인터페이스는 Streamlit 기반의 웹 애플리케이션으로 구현하였다.

2. 정량적 평가 결과

총 30건의 질의에 대해 평가를 수행하였으며, 각 질의는 검색 단계에서 상위 3개의 문서 조각을 기반으로 답변을 생성하였다. 평가 결과는 지표별 통계 분석과 문서 유형별 비교를 통해 제시한다.

표 4. 시스템 구성 및 RAG 구현 세부 사항

Category	Component	Specification
Software & Platform	Language	Python
	Framework	LangChain(Pipeline Orchestration) Streamlit(UI)
	Knowledge Base Platform	Dify(Vector Indexing & Management)
	Document Parser	Upstage Document Parse (API)
Model	LLM	Upstage Solar-pro2 (API)
	Embedding	Upstage solar-embedding-1-large-passage (API)
	Reranker	bge-reranker-v2-m3-ko-q4_k_m
RAG Parameters	Chunking	Semantic Chunking(Article-based)
	Retrieval	Hybrid Search(Semantic 0.7 + Keyword 0.3)
Generation Settings	Top-K (Initial Retrieval)	6(Score Threshold: 0.2)
	Top-N (Final Selection)	3 (Final Selection after Reranking)
	Temperature	0.0(Deterministic)
	System Prompt	Context-based Grounding, No Summarization

가. 생성 성능 평가

표 5에 제시된 Method별 생성 성능을 비교한 결과, Keyword-only RAG는 평균 Answer Relevancy 0.701, 평균 Faithfulness 0.781, Semantic-only RAG는 각각 0.812와 0.842로 Keyword-only RAG 대비 전반적인 평균 성능이 향상되었다. 본 연구에서 제안한 Hybrid RAG는 평균 Answer Relevancy 0.872, 평균 Faithfulness 0.912로 두 베이스라인 대비 가장 높은 평균 성능을 기록하였다.

Keyword-only RAG는 질의어와 규정 문서 간

표현이 직접적으로 일치하는 경우에는 일정 수준의 Answer Relevancy를 보였으나, 표현 다양성과 문서 분산성이 큰 질의에서는 관련 문서 회수 실패로 인해 Faithfulness와 최소값 측면에서 불안정한 성능을 나타냈다. Semantic-only RAG는 의미 기반 검색을 통해 전반적인 Answer Relevancy와 Faithfulness의 평균 성능을 개선하였으나, 의미적으로 유사하지만 질의와 직접적인 근거 관계가 약한 문서가 함께 회수되는 경우가 발생하여 일부 문항에서 성능 변동성이 관찰되었다.

반면, 제안한 Hybrid RAG는 키워드 기반 검색과 의미 기반 검색의 장점을 결합함으로써 질의와 직접적으로 관련된 규정 문서를 우선적으로 회수하면서도 표현 다양성으로 인한 누락을 효과적으로 보완하였다. 그 결과, Answer Relevancy와 Faithfulness 모두에서 가장 높은 평균 성능과 상대적으로 안정적인 분포를 보였으며, 대학 규정집과 같이 표현 다양성과 문서 구조의 복잡성이 높은 도메인에서 효과적인 검색 전략임을 확인하였다.

제안된 Hybrid RAG의 Answer Relevancy는 평균 0.872(± 0.176)와 중앙값 0.920으로 나타나 대부분의 질의에서 질문 의도에 부합하는 답변이 생성되는 경향을 확인하였다. 전체 응답 중 90.0%가 0.8 이상, 60.0%가 0.9 이상으로 집계되어 전반적으로 높은 관련성이 관측되었다. 다만 최솟값이 0인 사례가 1건 존재하여, 일부 질의에서는 질문 요구를 충족하지 못하는 실패 사례가 확인되었다.

Faithfulness는 평균 0.912(± 0.115), 중앙값 0.973으로 전반적으로 생성된 답변이 검색된 문서에 충실하게 근거하고 있음을 보였다. 전체 문항 중 50.0%가 1.0을 기록하였고, 70.0%가 0.9 이상으로 나타났다. 다만 일부 문항에서는 점수가 0.600까지 낮아져, 답변이 문서 근거를 충분히 반영하지 못하는 경우도 관찰되었다. 두 지표를 동시에 고려할 때, Answer Relevancy와 Faithfulness가 모두 0.8 이상인 문항

은 전체의 76.7%로 확인되었다.

문서 유형별 성능 비교 결과는 표 6에 제시하였다. 대학교 학칙은 Answer Relevancy 평균 0.815, 최솟값 0으로 다른 문서 유형에 비해 성능 편차가 컸으며, Faithfulness 역시 평균 0.845로 낮게 나타났다. 이는 분량이 크고 규정 범위가 넓은 학칙 문서의 특성상 동일한 검색 설정에서 핵심 조항이 상위 문서 조각으로 회수되지 못한 경우가 있었기 때문으로 해석된다. 반면 교학 규정 및 대학원 교학 규정은 두 지표 모두에서 상대적으로 높은 성능을 보였으며, 데이터사이언스대학원 교학 규정은 Faithfulness 평균 0.969로 가장 높은 근거 충실도를 보였다.

표 5. RAGAS 정량 평가 결과 요약 (N=30)

Method	Metric	Mean	Standard deviation	Median	Maximum	Minimum
Keyword-only RAG	Answer Relevancy	0.701	0.214	0.74	0.91	0
	Faithfulness	0.781	0.182	0.83	0.95	0.4
Semantic-only RAG	Answer Relevancy	0.812	0.189	0.86	0.96	0.2
	Faithfulness	0.842	0.154	0.89	0.98	0.5
Hybrid RAG (Proposed)	Answer Relevancy	0.872	0.176	0.92	0.982	0
	Faithfulness	0.912	0.115	0.973	1	0.6

표 6. 문서 유형별 평가 결과 (Hybrid RAG 기준)

Document Title	N	Answer Relevancy Mean	Answer Relevancy Minimum	Faithfulness Mean	Faithfulness Minimum
University Regulations	10	0.815	0	0.845	0.600
Academic Affairs Regulations	10	0.896	0.784	0.938	0.833
Graduate School Academic Affairs Regulations	5	0.925	0.868	0.942	0.800
Academic Affairs Regulations for the Graduate School of Data Science	5	0.887	0.792	0.969	0.900

나. 종합 성능 분석 및 고찰

정량적 평가 결과를 종합하면, 제안한 Hybrid RAG 시스템은 전반적으로 높은 응답 품질과 안

정적인 성능을 보였으나, 일부 질의에서는 성능 저하가 관찰되었다. 특히 여러 규정 문서를 종합적으로 참조해야 하는 질의의 경우, 문서별 적용 범위와 기준이 상이하어 핵심 조항이 상위 문서 조각으로 일관되게 회수되지 못하면서 Answer Relevancy가 하락하는 경향이 나타났다. 이러한 현상은 키워드 기반 검색만을 사용하는 Keyword-only RAG에서 가장 두드러지게 나타났으며, 의미 기반 검색을 단독으로 적용한 Semantic-only RAG에서도 유사한 문제가 부분적으로 관찰되었다.

한편, Answer Relevancy가 0으로 평가된 일부 사례에서도 Faithfulness가 1.0으로 측정된 경우가 확인되었다. 이는 모델이 검색된 문서 범위를 벗어난 정보를 생성하지 않고, 제공된 근거 내에서만 보수적으로 응답했기 때문으로 해석할 수 있다. 이러한 결과는 환각 억제 측면에서는 긍정적인 특성이지만, 사용자 질의에 대한 실질적인 해답을 제공하지 못했다는 점에서 검색 단계의 커버리지 부족이라는 한계를 동시에 드러낸다. 즉, 생성 모델의 한계라기보다는 검색 단계에서의 문서 회수 실패가 응답 품질 저하의 주요 원인으로 작용한 사례로 볼 수 있다.

이러한 분석을 바탕으로 향후 개선 방향으로서는 검색 단계와 생성 단계 간의 정렬(alignment)을 강화하는 방안을 제시한다. 우선, 규정의 장·절·조 및 적용 대상 정보를 포함한 메타데이터 필터링을 통해 질의와 직접적으로 관련된 문서군으로 검색 범위를 제한함으로써, 불필요한 다문서 혼합에 따른 문맥 희석 현상을 줄일 필요가 있다. 또한 질의 재구성, 확장 및 분해 기법을 도입하여 사용자의 불완전하거나 모호한 질문을 규정 문맥에 적합한 형태로 보정함으로써 핵심 조항 회수 가능성을 높일 수 있다. 이러한 개선은 검색 안정성과 답변 일관성을 동시에 향상시켜, 대학 행정 환경에서 규정 질의응답 시스템의 실질적 활용 가능성을 더욱 높일 것으로 기대된다.

V. 결 론

본 연구는 대학 행정 규정의 방대한 분량과 기존 규정 검색 시스템의 한계로 인해 발생하는 정보 접근성 문제를 해결하기 위해, 대학 규정집 도메인에 특화된 검색 증강 생성(RAG) 기반 질의응답 시스템을 설계·구현하였다. HWPX 형식의 규정 문서를 Markdown 형식으로 변환하고 전처리하여 장·절·조 단위의 구조화된 지식베이스를 구축하였으며, 의미 기반 검색과 키워드 기반 검색을 결합한 하이브리드 검색과 Cross-Encoder 기반 재순위화를 통해 질의와의 관련성이 높은 규정 조항을 효과적으로 선별하였다. 나아가 메타데이터 기반 문맥 구조화와 근거 인용형 프롬프트 설계를 적용함으로써, 생성된 답변에 규정 출처를 명시하고 응답의 신뢰성과 검증 가능성을 확보하였다.

제안한 시스템의 성능은 RAGAS 기반의 정량적 평가를 통해 검증하였으며, 30건의 대표 질의에 대해 Answer Relevancy와 Faithfulness 지표 모두에서 전반적으로 양호한 결과를 보였다. 이를 통해 본 연구는 대학 규정집 도메인에 RAG 기법을 적용함으로써 기존 키워드 중심 검색 방식의 한계를 보완하고, 사용자에게 정확하고 맥락화된 규정 정보를 제공할 수 있는 실질적인 접근 방안을 제시하였다는 점에서 탐색적 의의를 지닌다. 이는 특수 도메인 지식베이스 구축의 실질적 접근 방안을 마련함과 동시에, 향후 지능형 학사 행정 시스템으로의 발전 가능성을 확인한 탐색적 시도라 할 수 있다.

REFERENCES

[1] H.S. Lee. "RAG-based Intelligent College Information Question and Answer System" *Journal of Internet of Things and Convergence* Vol. 11, no. 3, pp. 161-166, 2025

[2] Y.S. Kim and J.M. Lee, "A Study on Usability Evaluation and Improvement of

University Academic Information Systems," *The Journal of Korean Association of Educational Information and Media*, Vol. 29, no. 2, pp.231-259, 2023.

[3] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu et al., "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," *arXiv preprint*, arXiv:2309.01219, 2023.

[4] Y. Gao, Y. Xiong, X. Gao, K. Jia et al., "Retrieval-augmented generation for large language models: A survey," *arXiv preprint*, arXiv:2312.10997, 2023.

[5] P. Lewis, E. Perez, A. Piktus, F. Petroni et al., "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*. Vol. 33, pp. 9459-9474, 2020.

[6] C.B. Yang, Y.S. Kim, "Implementation of Retrieval Augmented Generation (RAG) Model Using LLM: A RapidMiner-Based Approach," *Smart Media Journal*, Vol. 14, no. 2, pp.34-42, 2025

[7] G.H. Kim, D.G. Kim, "Comparison of Fine-tuning vs. RAG in sLLM - Focusing on Machine Reading Comprehension and Sentiment Analysis," *Smart Media Journal*, Vol. 14, no. 6, pp.50-59, June. 2025

[8] K. Seo and T. Utsuro, "RAG based Question Answering of Korean Laws and Precedents," In Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER), Vienna, Austria, *Association for Computational Linguistics*, pp. 91-100, 2025.

[9] E.S. Shin, H. So, Y.J. Yoon, C.W. Jung et al., "Modular RAG-based Legal Question Answering System Reflecting the Hierarchical Structure of Civil Law," *Journal of Korean Institute of Information Technology*, Vol. 23, no. 11, pp. 27-38, 2025.

[10] J.W. Seo and J.H. Min. "Optimization of a Hybrid RAG System for Korean Legal QA," *Journal of The Korea Society of Computer and Information*, Vol. 30, no. 8, pp. 53-63, 2025.

[11] Upstage, Document Parse API(2025). <https://www.upstage.ai> (accessed Dec.,10,2025).

[12] Upstage, Embedding API(2025). <https://www.upstage.ai> (accessed Dec.,10,2025).

[13] Dify, Dify Knowledge Base(2023). <https://dify.ai> (accessed Dec.,10,2025).

[14] J.H Lee, "Research on Generative AI Chatbot Service based on Academic Data," *Smart Media Journal*, Vol. 13, no. 12, pp.122-129, 2024

[15] R. Es, J. Gao, A. Chen, and A. Liu, "RAGAS: Automated Evaluation of Retrieval-Augmented

Generation,” *arXiv preprint*, arXiv:2309.15217, 2023.

저 자 소 개



이상훈(준회원)

2026년 전남대학교 데이터사이언스 대학원 석사과정.

<주관심분야 : Large Language Model, Retrieval-Augmented Generation, Model Context Protocol, Multi-Agent System>



이찬명(준회원)

2021년-현재 전남대학교 통계학과 학부과정.

<주관심분야 : Large Language Model, Retrieval-Augmented Generation, Machine Learning, Deep Learning, Data Science>



임희정(정회원)

1988년 이화여자대학교 수학교육학과 학사 졸업.

1994년 Columbia 대학교 보건통계학과 석사 졸업.

2001년 Columbia 대학교 보건통계학과 박사 졸업.

1998년-2000년 Neurological Institute of New York 보건통계학자

2001년-2004년 서울대학교 의과대학 선임연구원

2005년-2022년 전남대학교 치의학전문대학원 교수

2022년-현재 전남대학교 데이터사이언스 대학원 교수

<주관심분야 : Large Language Model, Retrieval-Augmented Generation, Model Context Protocol, 자연어처리, 빅데이터분석>