JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 8, pp. 53-63, August 2025
https://doi.org/10.9708/jksci.2025.30.08.053

Optimization of a Hybrid RAG System for Korean Legal QA

Jun-Won Seo*, Junghye Min**

*Student, Dept. of Computer Science, Inha Technical College, Incheon, Korea
**Professor, Dept. of Computer Science, Inha Technical College, Incheon, Korea

[Abstract]

Legal question-answering systems demand high reliability and accuracy, and large language models (LLMs) have recently been actively explored to meet these requirements. However, pretrained LLMs often struggle to reflect the most recent case law or specific legal provisions, which can lead to so-called “hallucination” — the generation of factually incorrect information. To address this issue, Retrieval-Augmented Generation (RAG), which generates responses based on external documents, has received growing attention. This study aims to develop a RAG system tailored to the Korean legal domain by optimizing key components including document chunking, embedding models, and retrieval strategies. Experimental results show that combining BM25 with a fine-tuned embedding model trained on Korean legal data, applied to semantically chunked documents, yields the best performance. The proposed hybrid retrieval approach outperformed baseline methods in both retrieval accuracy and factual consistency of the generated answers.

▸ **Key words**: Retrieval-Augmented Generation (RAG), Legal QA, Hybrid Retrieval, Semantic Chunking, Embedding Fine-tuning

[요 약]

법률 질의응답 시스템에는 높은 수준의 신뢰성과 정확성이 요구되며, 이를 위한 방법으로 최근 대규모 언어 모델 (LLM)을 활용한 연구가 활발히 진행되고 있다. 그러나 사전학습 기반의 LLM은 최신 판례나 세부 법령의 반영이 어려워, 사실과 다른 내용을 생성하는 이른바 ‘환각 (hallucination)’ 현상이 발생할 수 있다. 이를 보완하기 위해, 외부 문서를 기반으로 응답을 생성하는 검색 증강 생성 (Retrieval-Augmented Generation, RAG) 기법이 주목받고 있다. 본 연구에서는 한국어 법률 도메인에 특화된 RAG 시스템을 구축하고자, 문서 분할, 임베딩 모델, 검색 기법을 조합하여 최적의 구조를 설계하고 성능을 분석하였다. 실험 결과 의미 기반으로 청킹된 문서를 대상으로, 한국어 법률 데이터로 파인튜닝한 E5 임베딩 모델과 BM25를 결합한 하이브리드 검색 전략을 적용했을 때 가장 우수한 성능을 보였으며, 검색 정확도와 응답의 사실성 모두에서 기존 방법을 상회하는 결과를 얻었다.

▸ **주제어**: 검색 증강 생성, 법률 질의응답, 하이브리드 검색, 의미 기반 청킹, 임베딩 파인튜닝

• First Author: Jun-Won Seo, Corresponding Author: Junghye Min
*Jun-Won Seo (joonone.seo@gmail.com), Dept. of Computer Science, Inha Technical College
**Junghye Min (jhmin@inhatc.ac.kr), Dept. of Computer Science, Inha Technical College
• Received: 2025. 07. 15, Revised: 2025. 08. 07, Accepted: 2025. 08. 11.

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

I. Introduction

최근 법률 정보에 대한 대중의 관심이 증가함에 따라, 법령 및 판례를 이해하고 활용할 수 있는 자동 질의응답 (QA) 시스템의 필요성이 커지고 있다. 특히 대규모 언어 모델 (LLM)의 발전에 힘입어 법률 분야에서도 자연어처리 (Natural Language Processing, NLP) 기반의 법률 AI 응용 연구가 활발히 이루어지고 있다. 이러한 흐름 속에서, 외부 문서로부터 신뢰할 수 있는 정보를 검색하고 이을 생성 모델의 응답 근거로 활용하는 Retrieval-Augmented Generation (RAG) 방식이 주목받고 있다 [1]. RAG 방식은 외부 지식을 생성 과정에 활용함으로써, 대규모 언어 모델이 자주 겪는 환각 (hallucination) 문제를 줄이고 보다 사실에 기반한 응답을 생성하는 데 효과적인 접근법이다.

그러나 법률 도메인에 RAG 방식을 적용할 때는 여러 기술적 과제들이 존재한다. 법률 정보 탐색은 단순한 문서 검색을 넘어, 여러 조문과 판례를 종합적으로 고려해야 하며, 유사한 문구 간의 미세한 의미 차이조차 법적 해석에 영향을 미칠 수 있다는 특징이 있다. 따라서 단순히 관련 문서를 찾는 것만으로는 부족하며, 판결의 핵심 근거가 되는 정확한 문장을 효과적으로 찾아 답변 생성에 활용할 수 있는 문맥 확보 과정이 매우 중요하다 [2]. 이러한 법률 도메인의 특성상 기존의 영어 중심 RAG 아키텍처나 범용 사전학습 임베딩 모델을 그대로 적용할 경우, 검색 성능 저하와 생성 일관성 문제 등의 한계가 발생할 수 있다 [3].

따라서 한국어 법률 도메인의 특성을 반영한 문서 분할 (chunking) 기법, 임베딩 모델의 도메인 특화 파인튜닝, 검색 구조의 개선이 함께 고려되어야 한다. 본 연구에서는 한국어 법률 QA 분야에 특화된 RAG 시스템 구축을 목표로, 시스템의 핵심 구성 요소인 문서 분할, 임베딩, 검색 전략의 성능을 체계적으로 평가하고 최적의 아키텍처를 도출하고자 한다.

본 논문의 구성은 다음과 같다. 제2장에서는 관련 연구를 검토하고, RAG 시스템의 핵심 기술들을 살펴본다. 제3장에서는 실험을 위한 데이터셋 구축 과정과 시스템 설계 방안을 설명한다. 제4장에서는 각 구성 요소별 성능을 실험 결과에 기반하여 비교·분석하며, 제5장에서는 연구의 결론과 향후 연구 방향을 제시한다.

II. Related works

1. Retrieval-Augmented Generation (RAG)

검색 증강 생성 (RAG)은 대규모 언어 모델 (LLM)의 지식 한계를 외부의 최신 정보로 보완하는 생성 모델링 방식이다. Lewis 등이 제안한 RAG [1]는 답변을 만드는 생성기 (Generator)와 외부 문서에서 필요한 정보를 찾아내는 검색기 (Retriever)를 함께 사용한다. 검색기는 주어진 입력에 대해 관련성이 높은 문서 또는 구절을 우선적으로 식별하고, 생성기는 검색된 내용을 바탕으로 최종 응답을 생성한다. RAG 방식은 이러한 구조를 통해 별도의 추가 학습 없이도 외부 지식에 접근할 수 있다는 점에서 주목받고 있다. 특히, 답변의 근거를 외부 문서에 두기 때문에 LLM의 대표적인 한계인 '환각 (hallucination)' 현상을 완화할 수 있다는 점이 RAG의 주요한 강점으로 뽑힌다.

그러나 RAG 시스템의 검색 및 생성 단계에서도 새로운 유형의 환각이 발생할 수 있다. 예를 들어, 검색기가 부적절한 문서를 검색하거나, 생성기가 검색된 정보를 잘못 해석하여 비사실적인 응답을 생성하는 경우 [4]가 이에 해당한다. 따라서 RAG의 각 구성 요소를 정교하게 최적화하는 것은 환각 현상을 최소화하고 시스템의 신뢰성을 향상시키는 데 필수적이다.

이러한 배경 속에서 RAG는 사실 기반 추론이 요구되는 과제, 특히 법률과 같은 전문 분야에서 활발히 연구되고 있다. 본 연구 또한 RAG의 기본 프레임워크를 바탕으로, 한국어 법률이라는 특수 도메인에서 발생하는 고유한 문제들을 해결하고 시스템 성능을 최적화하는 것을 목표로 한다.

2. Retrieval Technologies

RAG 시스템의 전체 성능은 사용자의 질의에 대해 관련성 높은 문서를 얼마나 정확하게 검색할 수 있는지에 크게 영향을 받는다 [2]. 이에 따라 검색 성능 최적화는 RAG 연구 분야에서 중요한 연구 과제로 부각되고 있다.

기존 정보 검색은 TF-IDF [5]나 BM25 [6]와 같은 어휘 기반 (Lexical) 검색, 즉 희소 (Sparse) 검색 기법에 의존해왔다. 반면, 밀집 (Dense) 검색은 문서와 쿼리를 고차원 임베딩 공간에 매핑하여 의미적 유사성을 측정하는 방식이다. 밀집 검색은 단어 단위 매칭에 의존하지 않고, 문맥적 의미를 포착할 수 있다는 장점이 있다.

희소 검색 기법 중 하나인 BM25는 단어 빈도와 역문서 빈도, 문서 길이 등의 요소를 종합적으로 고려하여 문서의 관련성을 평가하는 확률적 랭킹 함수로, 키워드 매칭이 중요한 환경에서 널리 활용되어 왔다. 하지만 이 방식은 동의어

나 문맥의 의미를 충분히 반영하지 못한다는 한계가 있다.

이러한 문제를 보완하기 위해, 최근에는 밀집 검색 기법이 주목받고 있으며, 그 예로 Dense Passage Retrieval (DPR)이 있다. Karpukhin 등은 질문-정답 쌍 데이터를 활용해 BERT 모델을 미세조정 (fine-tuning)하는 방식을 통해 밀집 임베딩 기반 검색 성능을 크게 향상시켰으며, Top-5 검색 정확도 기준으로 BM25를 능가하는 성과를 보였다 [7].

이후 BEIR 벤치마크를 통해 검증된 E5와 같은 범용 임베딩 모델이 등장하며, 밀집 검색의 성능은 더욱 향상되었다. E5는 대규모 텍스트 쌍 데이터 (CCPairs)를 이용한 대조 학습 (contrastive learning) 기반 훈련으로, 별도 레이블 없이도 제로샷 (zero-shot) 설정에서 기존 BM25 기준선을 초과하는 성능을 달성하였다 [8].

한국어 환경에서는 의미 기반 검색의 대표적인 방법 중 하나로 KoSBERT [9]가 널리 사용된다. KoSBERT는 Sentence-BERT (SBERT) 구조 [10]를 기반으로 대규모 한국어 데이터셋에서 학습된 문장 임베딩 모델로, 코사인 유사도 기반의 문장 간 의미 비교에 최적화되어 있다.

그러나 긴 문서 구조나 복잡한 맥락을 다루는 경우에는 여전히 밀집 검색만으로는 정밀한 검색에 한계가 존재한다. 이를 보완하고자, 희소 검색과 밀집 검색의 장점을 결합한 하이브리드 검색 기법이 활발히 연구되고 있다. Luan 등 [11]은 고정된 벡터 크기를 가지는 밀집 검색의 한계를 지적하고, 희소-밀집 하이브리드 모델이 대규모 검색 환경에서 기존 모델들을 능가하는 성능을 보임을 입증하였다.

한편, 법률과 같은 전문 도메인에서는 범용 언어 모델의 성능이 제한적일 수 있다. Chalkidis 등 (2020)은 법률 분야 정보 검색에서 범용 BERT와 Legal-BERT의 성능을 비교하여, 도메인 특화 모델이 판례 검색 과제에서 우수한 성능을 보임을 확인하였다 [12]. 이러한 결과는 본 연구에서 한국어 법률 데이터셋 기반 E5 모델 파인튜닝 실험을 설계하게 된 주요 동기이다.

3. Document Chunking Techniques

RAG 시스템의 성능은 효과적인 검색기 설계뿐만 아니라, 검색 대상 데이터의 처리 방식에도 크게 영향을 받는다. 문서를 적절한 단위의 청크로 분할하는 것은 RAG의 기본적인 전처리 단계이다. 그러나 단순한 청킹 방식은 긴 문서의 전체 맥락을 손실시킬 위험이 있다 [13]. 이러한 문제를 완화하기 위해 다양한 청킹 기법이 제안되었다. 예를 들어, 지식 그래프 추출 과제에서는 지역적 문맥 보존을 위해 중첩된 슬라이딩 윈도우 (overlapping windows)를 활용하는 SLIDE 전략이 제안되었으며 [14], 질의 특성에 맞춰 동적으로 청크 단위를 선택하는 MoG (Mix-of-Granularity) 전략도 연구되었다 [15]. 또한, 문장 간 의미적 유사도를 활용해 문서 분할 여부를 결정하는 의미기반 청킹 기법이 연구되고 있으며, 최근에는 이를 특정 과제에 적용하여 그 효과를 검증하는 국내 연구 [16]가 제안되기도 하였다. 현재까지 최적의 청킹 전략에 대한 정답은 존재하지 않으며, 이는 본 연구에서 한국어 법률 문서의 특성을 고려하여 다양한 청킹 방식의 성능을 비교하는 이유이기도 하다.

4. Legal RAG Evaluation Benchmark

법률 질의응답에서는 과거 유사 판례를 참조하는 추론 과정이 요구되며, 이로 인해 검색 정확도와 응답의 신뢰성이 더욱 중요하다. 특히 법률 분야에서는 정확한 정보 근거에 기반한 응답 생성이 필수적이며, 이를 위해서는 검색 단계에서 신뢰도 높은 관련 문서를 찾아내는 과정이 선행되어야 한다. 이러한 필요에 따라, 검색 단계의 성능을 정밀하게 평가할 수 있는 벤치마크인 LegalBench-RAG가 제안되었다 [2].

그러나 LegalBench-RAG는 영어 법률 체계와 언어에 기반하여 설계되었기 때문에, 법률 용어와 문장 구조가 상이한 한국어 환경에 RAG 시스템의 성능을 평가하는 벤치마크로 직접 적용하는 데에는 한계가 있다. 이러한 문제를 반영하여, 최근에는 변호사와의 긴밀한 협력을 통해 실무 시나리오를 기반으로 개발된 KBL (Korean Legal Benchmark)과 같이, 한국어 법률 시스템의 특수성을 고려한 평가 체계를 구축하려는 연구 [17]가 등장하고 있다. 이러한 배경을 바탕으로, 본 연구는 국내 공개 법률 판례 데이터셋을 활용하여 한국어 RAG 시스템의 성능을 실증적으로 평가하고, 이를 통해 국내 법률 환경에 최적화된 시스템 구성 방안을 모색하고자 한다.

III. System Architecture

1. System Pipeline

본 장에서는 한국어 법률 질의응답 시스템을 구축하고, 각 구성 요소에 대한 실험 환경을 설계하기 위한 연구 절차와 시스템 아키텍처를 상세히 기술한다. 본 연구의 시스템은 검색 (Retrieval)과 생성 (Generation) 단계로 구성된 RAG 아키텍처를 기반으로 한다. 전체 시스템의 주요 동작 흐름은 다음과 같다. 사용자가 질문을 입력하면, 사전에 의미 단위로 나누고 벡터화한 법률 문서들로 구성된

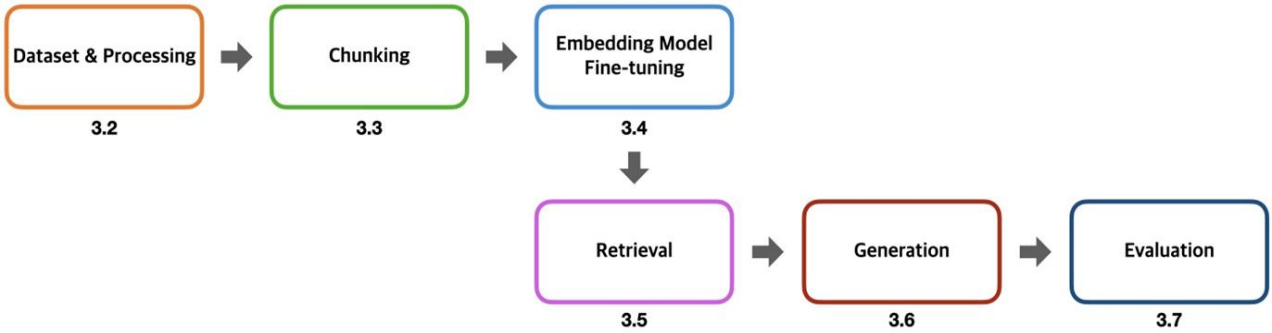


Fig. 1. System Pipeline

데이터베이스에서 관련 정보를 검색한다. 이때, 한국어 법률 도메인에 특화된 임베딩 모델을 활용한 밀집 검색과, BM25 기반의 희소 검색을 결합한 하이브리드 전략을 통해 가장 관련성 높은 문서들을 선택한다. 이렇게 추출된 근거 문서들은 대규모 언어 모델 (LLM)에 전달되어, 최종 답변 생성에 활용된다.

Fig. 1.은 본 연구의 3장의 구성과 각 단계에서 고려된 핵심 실험 요소들을 시각적으로 요약한 것이다. 연구 절차는 다음과 같은 순서로 진행된다. 먼저 3.2절에서는 실험에 활용된 데이터셋의 구성과 전처리 과정을 설명한다. 3.3절에서는 법률 문서의 길이와 구조적 특성을 고려하여 적용한 두 가지 청킹 접근법 (Fixed-Size, Semantic)의 특성을 비교한다. 이어서 3.4절에서는 법률 분야 검색 성능 향상을 위해 임베딩 모델의 도메인 적합성을 높이는 파인튜닝 과정을 소개한다. 3.5절에서는 구축된 인덱스를 바탕으로 사용자 질문에 대해 관련 문서를 추출하는 과정과 밀집 검색, 희소 검색, Hybrid 검색 방식에 대해 설명한다. 3.6절에서는 검색된 문서 조각을 기반으로 생성 모델에 입력할 문맥을 구성하고, 이를 통해 최종 답변을 생성하는 과정을 설명하며, 마지막으로 3.7절에서는 각 실험 단계에서 수집된 결과의 성능을 정량적으로 평가하기 위해 적용된 주요 지표들을 정의한다. 최종적인 성능 분석과 결과 비교에 대한 상세 논의는 이후 4장 결과 분석에서 다룬다.

2. Dataset and Preprocessing

본 연구에서는 AI 허브 (AI Hub)에서 제공하는 '법률/규정 텍스트 분석 데이터(고도화) - 상황에 따른 판례 데이터셋'을 실험의 기반으로 활용했다 [18]. 전체 데이터셋 중 일관된 도메인 내 실험 환경을 조성하기 위해, 민사 판례 원문 13,582건을 실험 데이터베이스로 선정하였다. 각 데이터의 '판례내용' 필드는 원문 (Source Document)으로 사용되었다. 또한, 데이터셋에 포함된 질의응답 (QA) 쌍은 임베딩 모델 파인튜닝과 최종 성능 평가에 사용되었다.

전처리 과정에서는 원문 텍스트 내 불필요한 메타 정보 및 특수문자를 정규표현식 (Regular Expression)을 이용해 제거하였다. 이후 정제된 텍스트는 kss (Korean Sentence Splitter) 라이브러리를 통해 문장 단위로 분리하여, 후속 청킹 (chunking) 단계에서 사용할 기본 처리 단위로 활용하였다.

최종적으로 전체 QA 데이터를 학습 (Training), 검증 (Validation), 평가 (Test) 세트로 분할하였다. 구체적으로, 1,545개의 QA 쌍을 학습용으로, 317개를 검증용으로, 346개를 테스트용으로 활용하였다. 모델 파인튜닝 과정에서는 학습 및 검증 세트를 사용하였으며, 본 논문에서 보고하는 모든 성능 지표는 최종 평가 (Test) 세트에서 산출된 결과를 기준으로 한다.

3. Chunking Strategy

법률 문서는 조문 간의 논리적 연결이 강하고 문장도 길게 구성되는 경우가 많다. 이러한 특성으로 인해, 정보를 의미 있는 단위로 분할하는 청킹 전략은 RAG 시스템의 검색 성능에 큰 영향을 미친다. 이에 본 연구에서는 두 가지 서로 다른 청킹 방식을 적용하고, 각각의 성능을 비교하였다.

첫째, 고정 청킹 (Fixed-Size Chunking)은 문서의 내용을 일정 단위로 자르는 단순한 방식으로, 이번 연구에서는 3문장을 하나의 청크로 묶고 한 문장씩 이동하는 방식 (chunk size=3, stride=1)을 적용하였다. 이 방식은 구현이 간단하다는 장점이 있지만, 문장 간 의미 흐름이나 경계를 고려하지 못하는 한계가 있다.

둘째, 의미 기반 청킹 (Semantic Chunking)은 문장 간 의미적 연관성을 바탕으로 분할 지점을 결정하는 방식이다. 구체적으로는, E5(Base) 모델을 이용해 각 문장을 임베딩한 후, 인접 문장 간 코사인 유사도를 계산하였다. 이후 문서 내 유사도 분포에서 하위 p% 구간(p=50, 60, 70)을 임계값 (threshold)으로 설정하고, 유사도가 해당 임계값보다 낮아지는 지점을 의미적 분할 경계로 활용하였다.

이후, 청크가 지나치게 짧거나 길어지는 것을 방지하기 위해 길이를 일정 범위 내로 조정하였다.

4. Embedding Model and Fine-tuning

본 절에서는 한국어 법률 질의응답 환경에 최적화된 임베딩 생성을 위해 수행한 E5 모델의 도메인 적응 파인튜닝 (domain-adaptive fine-tuning)과정을 설명한다. 본 연구에서는 E5(Base)모델을 한국어 법률 QA 데이터로 추가 학습하여 법률 도메인에 특화된 임베딩 모델을 구축하고, 검색 성능 향상을 목표로 하였다.

우선, E5 모델이 한국어 법률 문서에서도 강력한 성능을 발휘하는지 검증하기 위해, 3.2절에서 소개한 테스트 세트를 활용해 사전 실험을 수행하였다. 이 실험에서는 모두 사전학습 (pre-trained) 상태인 KoSBERT와 E5(Base) 모델의 검색 성능을 비교하였다. 그 결과, Table 1.과 같이 E5(Base) 모델이 전반적으로 더 높은 검색 관련성을 나타냈다.

Table 1. Performance Comparison between KoSBERT and E5(Base)

Retrieval Model	ref_Recall@3	ref_MRR@3	case_MRR@3
E5(Base)	**0.6534**	**0.6314**	**0.4537**
KoSBERT	0.5400	0.5043	0.2809

이러한 사전 성능 비교 결과를 바탕으로, 본 연구에서는 E5(Base) 모델을 임베딩 생성의 기본 모델로 선정하고, 도메인 적응 파인튜닝을 진행하였다. Fig. 2.에는 전체 학습 과정의 개요를 나타내었다.

학습 데이터로는 AI Hub의 한국어 법률 QA 데이터셋 (3.2절 참조)을 활용하였다. 이 데이터셋에서 질문 (question) 필드를 쿼리로, 해설 (commentary) 필드를 정답 문서 (positive passage)로 사용하여 학습 쌍을 구성하였다.

학습 방식으로는 In-Batch Negatives 전략을 적용하였다. 이는 별도의 오답 수집 과정 없이, 각 미니배치 내 다른 정답 문서들을 부정 샘플로 간주하는 방식이다. 즉, 하나의 배치에 포함된 모든 질문-정답 문서 쌍 중 특정 쿼리의 정답을 제외한 나머지 문서들이 해당 쿼리의 부정 샘플로 활용된다.

이 방법은 법률 문서의 의미적 유사성을 세밀하게 구분할 수 있는 임베딩을 학습하는 데 적합하고, 모델은 관련 있는 질문-정답 쌍은 더 가깝게, 무관한 쌍은 더 멀게 임베딩 공간에 배치되도록 학습되며 [8], 이 과정에서 내부 가중치는 반복적으로 조정된다. 손실 함수로는 Multiple Negatives Ranking (MNR) Loss를 사용하였다. 모든 training batch에 대해 최적화가 수행되었으며, 세부 실험 설정 및 하이퍼파라미터 튜닝 과정은 4.2절에 기술되어 있다. 최종적으로, 파인튜닝된 E5 모델을 활용해 실험에 사용된 모든 문서 청크의 벡터 표현을 생성하고, 이를 벡터 데이터베이스에 저장하였다.

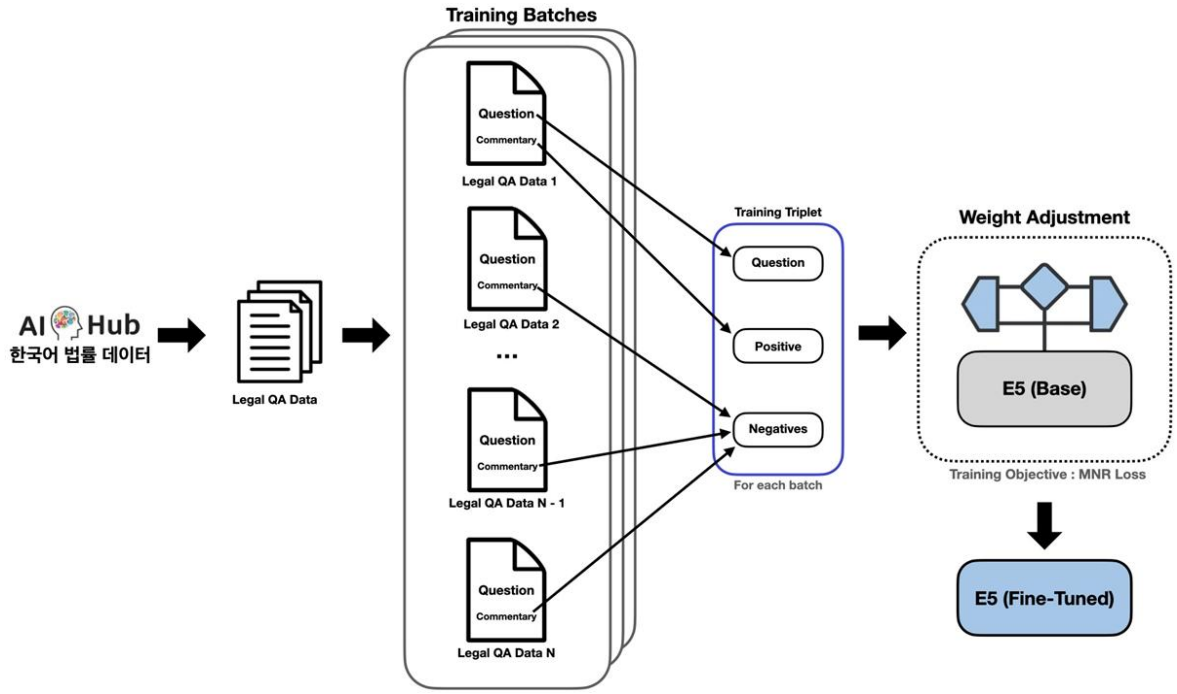


Fig. 2. Overview of domain-adaptive fine-tuning of E5 for legal retrieval using in-batch negatives

5. Retrieval Methods

본 연구에서는 검색 성능 비교를 위해 희소 검색, 밀집 검색, 그리고 이 둘을 결합한 하이브리드 검색 방식을 실험했다.

먼저, 희소 검색 기법인 BM25는 내부 랭킹 함수를 통해 키워드 일치도를 기반으로 각 청크의 관련도 점수를 계산한다. 이 기법은 법률 문서의 정형화된 용어나 사건번호 등 명시적 키워드 검색에 강점이 있어, 본 연구의 비교 실험에 포함했다.

다음으로, 밀집 검색 방식은 질문 (Query)과 문서 (Passage)를 각각 독립된 인코더로 임베딩한 후, 코사인 유사도 (Cosine Similarity)를 계산하여 의미적 관련성을 평가한다. 본 연구에서는 한국어 법률 도메인에 특화해 파인튜닝한 E5 모델 E5(Fine-tuned)을 밀집 검색기로 활용하였다.

마지막으로, 하이브리드 검색 (Hybrid Search) 방식은 BM25의 키워드 기반 정밀성과 E5 밀집 검색의 문맥적 이해 능력을 결합하여 두 기법의 강점을 모두 활용한다. 법률 QA 과제에서는 특정 용어나 사건번호의 정확한 일치가 중요할 뿐만 아니라, 질의 의도와 법적 개념을 깊이 있게 이해하는 능력도 필요하기 때문이다. 이 두 방식은 서로 다른 원리로 유사도를 평가하므로 스케일 (Scale)이 다르다. 이를 해결하기 위해, 본 연구에서는 각 방식의 평가값을 모두 [0, 1] 범위로 조정한 후 선형 결합 (Linear combination)을 수행했다. 구체적으로, BM25 점수는 검색된 후보군 내에서 최소-최대 정규화 (Min-Max Normalization)를 적용했고, [-1, 1] 범위의 E5의 코사인 유사도는 $\frac{Sim_{E5}+1}{2}$식을 통해 정규화했다. 가중치 α를 적용한 최종 하이브리드 점수는 아래 수식으로 계산된다.

$$Score_{hybrid} = \alpha \cdot norm(Score_{BM25}) + (1-\alpha) \cdot norm(Score_{E5})$$

여기서 $norm(Score_{BM25})$는 정규화된 BM25 점수를, $norm(Score_{E5})$는 정규화된 코사인 유사도를 의미한다.

앞서 설명한 세 가지 검색 방식은 각각의 기준에 따라 모든 청크에 대한 관련도 점수를 산출한다. 이후 점수가 높은 순서대로 청크를 정렬하고, 목록의 최상위에 위치한 K개의 청크를 선택하여 답변 생성을 위한 최종 문맥으로 활용하였다.

6. Generation Method

최종 답변 생성을 위한 근거 문맥 (context)은 3.5절에서 기술한 검색 모델이 선별한 상위 K개의 청크 (chunk)를 기반으로 구성하였다. 본 연구에서는 LLM이 처리할 수 있는 문맥 길이와 연산 효율성을 고려하여 K를 3으로 설정하였다. 그러나 검색 모델이 반환한 청크만으로는 답변 생성에 필요한 충분한 맥락을 제공하기 어렵다는 한계가 존재한다.

이러한 문제를 해결하기 위해 검색된 청크의 앞뒤 인접 청크를 함께 포함하는 슬라이딩 윈도우 방식을 적용해 문맥을 확장하였다. 구체적으로는, 검색된 청크를 기준으로 앞뒤에 위치한 청크를 함께 포함하여 총 3개의 연속된 청크로 문맥 블록을 구성하였다. 이 방식은 법률 문서에서 핵심 내용과 함께 인접한 정의나 예외 조항까지 자연스럽게 포함해, LLM이 답변의 타당한 근거를 확보하고 신뢰도 높은 응답을 만들 수 있도록 돕는다. 이렇게 구성한 확장 문맥 블록과 사용자 질의를 하나의 프롬프트로 통합해, Google의 Gemma-3-27b-it 모델에 입력하여 최종 답변을 생성했다.

7. Evaluation Metrics

7.1 Retrieval Performance Metrics

본 연구의 정량적 평가는 3.2절에서 기술한 데이터셋의 질의응답 (QA) 쌍을 활용하여 수행하였다 [18]. 각 질의 (Query)에는 정답 기준으로 사용될 '참조조문'과 '사건번호'가 라벨링되어 있으며, 리트리버가 검색한 청크의 정답 여부는 다음 기준에 따라 판별하였다.

참조조문 (Reference Rules) : 질의에 포함된 reference_rules 필드 (정답 조문 집합)와 검색된 청크의 원문 판례 내 참조조문 집합 간에 하나 이상의 공통 조문이 존재할 경우 정답으로 간주하였다.

사건번호 (Case ID) : 질의의 reference_court_case 필드에 명시된 사건번호와, 검색된 청크가 속한 판례의 사건번호가 일치할 경우 정답으로 판단하였다.

이러한 정답 판별 기준을 바탕으로, 본 연구에서는 검색 성능 평가를 위해 표준 정보 검색 지표들을 활용하였다. Recall@k는 검색된 상위 k개 결과 내에 실제 정답이 포함되었는지를, MRR@k는 얼마나 높은 순위에 첫 정답이 위치하는지를 측정한다. 모든 평가에서 K=3으로 고정했으며, 최종 성능은 각 지표를 질문 단위로 계산 후 산술 평균 (Macro-average)하는 방식으로 산출했다. 이때 평가의 정확성을 높이기 위해, 검색 결과와 정답 집합에서 중복 항목은 제거한 후 고유한 값만을 기준으로 지표를 계산하였다. 사용된 각 평가 지표의 세부 정의는 Table 2.에 정리하였다.

Table 2. Definitions of Retrieval Performance Metrics

Criteria	Metric	Description
Reference Rules	ref_Recall@k	The probability that a correct rule is included in the top-k results
Reference Rules	ref_MRR@k	The average of the reciprocal ranks of the first correct rule found
Case ID	case_id_MRR@k	The average of the reciprocal ranks of the first correct case ID found

7.2 Generation Performance Metrics

앞선 3.5절에서 비교한 각 검색 모델이 선택한 문서 조각 (chunk)이 최종 답변 품질에 미치는 영향을 다각도로 분석하기 위해 다양한 자동화 평가 지표를 활용하였다.

생성된 답변의 품질 평가는 RAG 평가 프레임워크인 RAGAs [19]와 표준 유사도 기반 지표인 BERTScore [20]를 기반으로 이루어졌다. 구체적으로, RAGAs를 통해서는 답변의 충실성 (Faithfulness), 관련성 (Answer Relevancy), 정확성 (Answer Correctness)을 측정했으며, BERTScore를 추가적인 지표로 사용했다. RAGAs 기반의 평가는 GPT-3.5-Turbo를 평가자 (Evaluator)로 사용하여 각 항목을 [0, 1] 척도로 점수화했으며, 모든 지표는 최종적으로 전체 테스트 데이터셋에 대해 평균 내어 계산했다. 사용된 각 평가 지표의 상세한 정의는 Table 3.와 같다.

IV. Experiments and Results

1. Performance by Chunking Strategy

RAG 시스템의 검색 성능에 직접적인 영향을 미치는 청킹 전략의 효과를 비교하기 위해, 3.3절에서 설계한 고정 청킹 및 의미 기반 청킹 방식을 동일한 임베딩 모델 E5(Base)을 기준으로 성능을 실험하였다. 각 청킹 전략에 따른 검색 성능은 주요 평가 지표별로 측정되었으며, 그 결과는 Table 4.에 제시하였다.

실험 결과, 의미 기반 청킹 (Semantic Chunking) 전략이 고정 청킹 (Fixed-Size)에 비해 전반적으로 더 우수한 검색 성능을 보였다. 특히, 코사인 유사도 기반의 문장 간 의미 유사도를 활용해 청크 경계를 동적으로 설정하는 방식이, 단순히 고정된 문장 수로 분할하는 방법보다 더 효과적인 것으로 나타났다.

Semantic Chunking의 다양한 임계값 설정 (p=50%, 60%, 70%)에 따른 실험 결과 중, p=50% 설정은 전체 청크 수를 약 26% 줄이면서도 ref_Recall@3 (0.6562), ref_MRR@3 (0.6330) 등 주요 지표에서 가장 높은 성능을 기록했다. 이는 청크 수를 줄여 효율성을 확보하면서도 검색 정확도를 동시에 달성할 수 있음을 보여준다.

다만, 본 연구의 최종 목표가 법률 질의응답인 점을 감안할 때, 단순한 관련성뿐 아니라 특정 사실 (Fact)을 정확히 식별하는 능력이 중요하다. Table 4.에 따르면, Semantic (p=60%) 설정은 ref_Recall@3 지표에서 p=50% 모델에 근소하게 뒤졌지만, 핵심 개체인 '사건번호'에 대한 정확도를 나타내는 case_MRR@3에서는 가장 뛰어난 성능을 보였다. 이는 p=60% 설정이 전반적인 검색 성능과 함께 법률 QA에 필수적인 고유 정보 식별 능력 사이에서 가장 균형 잡힌 결과를 제공함을 보여준다. 따라서 이후 진행된 RAG 시스템의 종단 (end-to-end) 평가에서는 Semantic (p=60%)을 최적의 청킹 전략으로 채택하여 실험을 수행하였다.

Table 3. Definitions of Generated Answer Quality Evaluation Metrics

Metric	Description
Faithfulness	Evaluates wether the generated answer is factually grounded in the provided context
Answer Relevancy	Evaluates how relevant the generated answer is to the original user query
Answer Correctness	Evaluates how semantically similar the generated answer is to the ground truth answer
BERTScore	Calculates semantic similarity between the generated answer and the ground truth answer using BERT embeddings.

Table 4. Comparison of Retrieval Performance by Chunking Strategy

Category	Model (p=bottom % of similarity)	No. of Chunks	ref_Recall@3	ref_MRR@3	case_MRR@3
Fixed-Size Chunking	Fixed-Size (Baseline)	74,665	0.6182	0.5831	0.4401
Semantic Chunking	Semantic (p=50%)	55,286	0.6562	0.6330	0.4408
	Semantic (p=60%)	**62,973**	**0.6534**	**0.6314**	**0.4537**
	Semantic (p=70%)	70,751	0.6423	0.6183	0.4463

2. Embedding Model Fine-tuning

법률 도메인에 특화된 최적의 임베딩 모델을 구축하기 위해, 3.4절에서 제안한 방법으로 E5 모델의 도메인 적응 파인튜닝을 수행하였다. 최적의 파인튜닝 조건을 찾기 위해 배치 사이즈 (Batch Size)에 따른 성능 변화를 추가로 분석하였으며, 배치 사이즈 16과 32를 비교한 결과, 배치 사이즈 32에서 보다 안정적이고 우수한 검색 성능을 나타냈다. 이에 본 연구에서는 안정성과 성능을 종합적으로 고려하여 배치 사이즈 32를 최적 학습 조건으로 선정하였다.

파인튜닝된 E5 모델 E5(Fine-tuned)는 파인튜닝 이전의 기본 모델 E5(Base)에 비해 모든 평가 지표에서 성능 향상을 보였다. 특히, 전반적인 검색 품질을 나타내는 ref_Recall@3와 ref_MRR@3 지표는 각각 약 3.5%p 개선되었으며, 법률 도메인의 핵심 정보인 '사건번호'를 정확히 찾아내는 능력을 평가하는 case_MRR@3 지표는 약 5%p의 성능 향상을 기록하였다. 이는 도메인 적응 파인튜닝이 범용 모델을 한국어 법률 분야 특성에 맞게 효과적으로 특화했음을 보여주며, 이후 진행되는 모든 비교 실험에서는 이러한 최적 조건으로 파인튜닝된 E5(Fine-tuned) 모델을 '최적화된 밀집 검색 모델'로 활용하였다.

Table 5. Performance Comparison of Embedding Model Before and After Fine-tuning

Retrieval Model	ref_Recall@3	ref_MRR@3	case_MRR@3
E5 (Base)	0.6534	0.6314	0.4537
E5 (Fine-tuned)	**0.6884**	**0.6661**	**0.5040**

3. Retrieval Model Performance Comparison

이전 단계에서 최적화한 청킹 전략 (Semantic, p=60%)과 파인튜닝된 E5 모델을 바탕으로, 본 연구에서는 BM25와 E5 모델의 가중치를 조합하는 하이브리드 검색 방식을 실험했다. 하이브리드 모델의 가중치 a를 0.1부터 0.9까지 변화시키며 주요 성능 지표를 평가한 결과 (Table 6.), a =0.8일 때 가장 우수한 성능을 보여 해당 값을 최종 가중치로 선정하였다.

Table 7.은 이 설정을 적용한 하이브리드 모델과 기존 검색 모델들의 주요 성능 지표를 비교한 결과이다. 비교 결과, BM25는 파인튜닝된 E5(Fine-tuned) 모델보다 모든 평가 지표에서 우수한 성능을 보임을 확인할 수 있었다. 주목할 점은, 최종적으로 제안한 Hybrid 방식이 모든 주요 평가 지표에서 가장 뛰어난 성능을 달성했다는 것이다. 특히 ref_Recall@3 및 ref_MRR@3 지표에서는 BM25 단독 모델보다 소폭 개선된 결과를 보여주었으며, 이는 희소 검색과 밀집 검색의 결합이 실질적인 성능 향상으로 이어졌음을 보여주며, 두 방식의 상호 보완적 특성이 검색 품질을 끌어올렸음을 의미한다.

Table 6. Comparison of Retrieval Performance by α

a	ref_Recall@3	ref_MRR@3	case_MRR@3
0.1	0.7093	0.6889	0.5598
0.5	0.7096	0.7068	0.5772
0.8	**0.7118**	**0.7085**	**0.5872**

Table 7. Final Retrieval Model Performance Comparison

Retrieval Model	ref_Recall@3	ref_MRR@3	case_MRR@3
E5 (Fine-tuned)	0.6884	0.6661	0.5040
BM25	0.7047	0.7047	0.5771
Hybrid (BM25+E5) (a = 0.8)	**0.7118**	**0.7085**	**0.5872**

4. Quantitative Evaluation of Generation Performance

앞선 검색기 (Retriever) 성능 평가에 이어, 각 검색기가 최종 생성되는 답변 품질에 미치는 영향을 RAG 평가 프레임워크인 RAGAs를 활용해 분석하였다. 세 가지 방식에 대한 핵심 평가 결과는 Fig. 3.에 제시되어 있다.

분석 결과, 단일 검색기 모델 간에는 명확한 장단점이 나타났다. 파인튜닝된 E5 모델 E5(Fine-tuned)는 Faithfulness (0.6584) 지표에서 최고점을 기록하며, 의미적으로 정제된 문맥을 바탕으로 환각 없이 답변하는 능력이 가장 우수함을 보여주었다. 반면 BM25는 정답과의 의미적 일치도를 평가하는 Answer Correctness (0.6319)에서 가장 높은 성능을 보였다.

하이브리드 방식은 Answer Relevancy에서 0.7044를 기록하며, 평가에 사용된 모든 방식 중 가장 높은 성능을 보였다. 이는 사용자의 질문 의도에 가장 부합하는 답변을 생성했음을 의미한다. 또한, 정답과의 의미적 유사도를 나타내는 BERTScore (0.6938)에서도 최고점을 보여, 내용의 정확성과 표현의 품질을 모두 확보했다. 검색 단계에서 BM25의 정확성과 E5(Fine-tuned)의 의미적 관련성을 균형 있게 결합한 결과, 생성 단계에서 최종 답변의 관련성과 품질이 향상되었음을 확인할 수 있다.

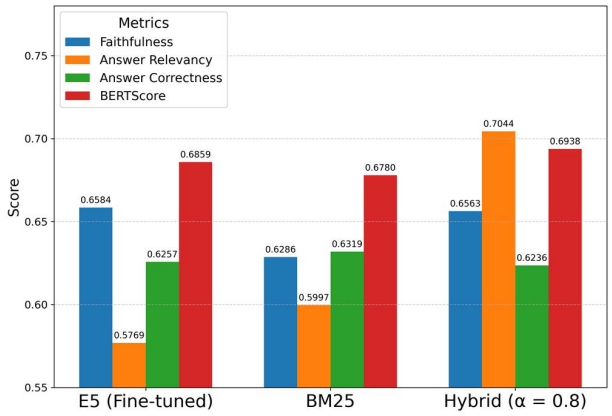


Fig. 3. Comparison of Final Answer Generation Performance by Retrieval Method

5. Qualitative Analysis of Generation Performance

앞선 정량적 평가 결과, 제안한 하이브리드 검색 방식이 단일 검색기 모델 보다 우수한 성능을 보이는 것으로 확인되었다. 본 절에서는 이러한 수치적 차이가 실제 질의응답 과정에서 어떤 영향을 주는지 살펴보고, 특히 검색된 근거(청크)의 질이 최종 답변의 정확성에 미치는 영향을 자세히 분석하였다. 분석 대상은 동일한 생성기를 사용하였음에도, 검색 방식의 차이로 인해 상반된 법률적 결론이 도출된 QA_07684 사례이며, 해당 결과는 Fig. 4에 제시하였다.

Fig. 4.에서 확인할 수 있듯이, E5(Fine-tuned) 검색 방식을 적용한 경우에는 질문의 핵심 쟁점인 '양도담보'와 무관한 '매매계약 해제' 판례가 근거로 활용되어, "원칙적으로 거절할 수 있습니다" 라는 명백한 오답이 도출되었다. 이는 적절한 근거 문서의 미확보에 기인한 것으로 판단된다.

반면, 제안하는 하이브리드 검색 방식을 적용한 경우에는 질문의 특정 상황인 '양도담보권 설정 후 대항력을 갖춘 임차인'과 완전히 일치하는 근거를 확보함으로써, "원칙적으로 거절할 수 없습니다" 라는 올바른 답변을 도출할 수 있었다.

본 사례는 검색 방식의 차이가 최종 답변의 법률적 정확성에 결정적인 영향을 미치며, 제안하는 하이브리드 검색 방식이 정확한 근거 확보를 통해 RAG 시스템의 신뢰성과 안정성에 기여할 수 있음을 시사한다.

V. Conclusion and Discussion

1. Proposed Architecture

본 연구는 4장에서 수행한 일련의 실험을 토대로, 한국어 법률 질의응답에 특화된 최적화된 RAG 시스템 아키텍처를 제안한다. 전체 구성은 다음과 같은 주요 설계 요소로 요약할 수 있다.

먼저, 문서 분할 (chunking) 전략에서 고정 길이 방식과 의미 기반 방식을 비교한 결과, 의미 기반 청킹 전략 중 의미 유사도 하위 60%를 선택하는 방식이 가장 안정적인 성능을 보여 최종 청킹 방식으로 채택하였다.

임베딩 모델 측면에서는 사전 학습된 성능이 우수한 E5(Base) 모델을 기반으로, 법률 QA 데이터셋을 활용한 도메인 적응형 파인튜닝을 수행하였다. 그 결과, 전반적인

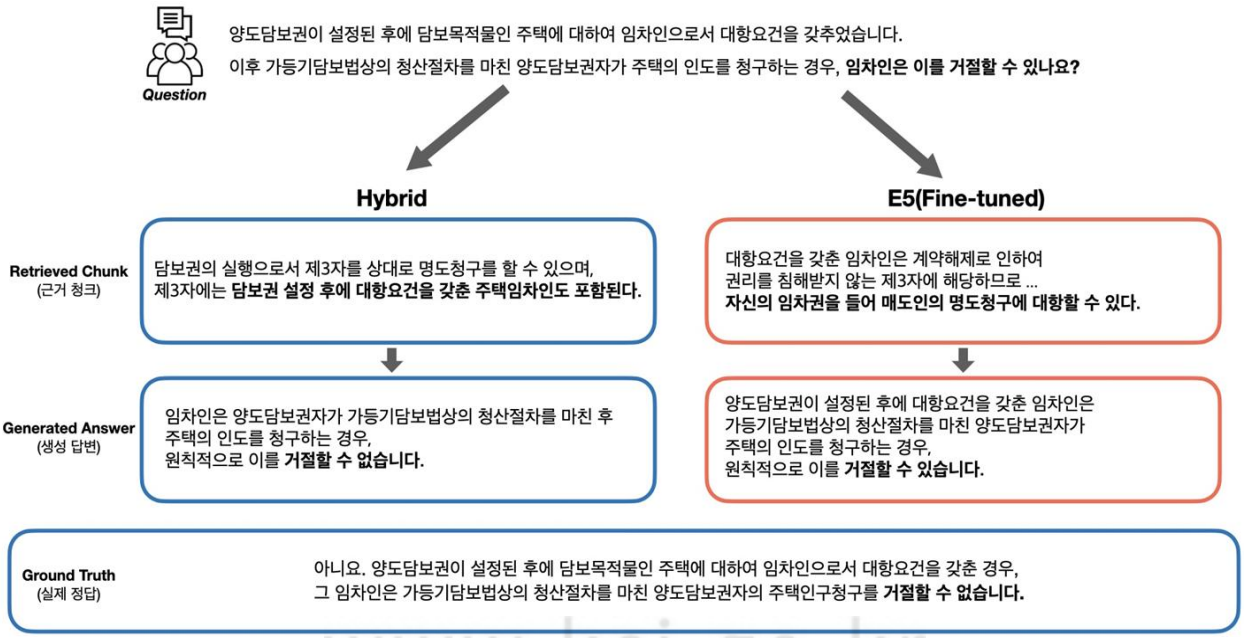


Fig. 4. Comparison of Retrieved Evidence and Generated Answers by Model

검색 성능이 향상되어, 해당 모델이 법률 도메인에 효과적으로 특화되었음을 확인할 수 있었다.

검색기 구성에서는 BM25, E5(Fine-tuned), 그리고 이 두 모델을 결합한 Hybrid 방식을 비교 분석하였다. 실험 결과, BM25와 E5를 8:2의 가중치 비율로 조합한 하이브리드 방식이 모든 주요 검색 지표에서 가장 뛰어난 성능을 보여 이를 최종 검색 방식으로 채택하였다.

생성 단계에서는 Google의 Gemma 3 모델을 활용하였으며, 4.4절의 RAGAs 기반 생성 평가 결과, 하이브리드 검색 모델이 제공한 문맥이 높은 관련성과 정확성을 바탕으로 우수한 답변을 생성함을 확인하였다.

이러한 실험적 검증을 바탕으로, 본 연구는

- 의미 기반 청킹
- 한국어 법률 QA 데이터로 파인튜닝한 E5 임베딩 모델
- 파인튜닝한 E5와 BM25를 결합한 하이브리드 검색

을 결합한 RAG 시스템을 최종 아키텍처로 제안하며, 전체 구조는 Fig. 5.에 제시하였다.

2. Limitations and Future Research

본 연구는 법률 QA에 특화된 RAG 시스템을 제안하고 그 효과를 검증했지만, 여전히 보완이 필요한 한계들이 존재한다. 이에 따라 후속 연구에서 이를 보완할 수 있는 방향을 제시한다.

- 평가 방식의 한계 : 본 연구는 RAGAs와 같은 자동 평가 지표를 통해 생성 답변의 품질을 측정하였으나, 법률적 맥락의 세밀한 뉘앙스나 실무적 타당성까지 완벽히 반영하지는 못한다. 예를 들어, 본문에서 다룬 사례와 같이 정답과 오답이 명확히 구분되는 경우에는 평가가 가능했으나, 복잡한 법리 쟁점에 대해서는 자동 평가만으로 신뢰도를 충분히 확보하기 어렵다. 따라서 연구 결과의 실용성을 높이기 위해서는 향후 변호사 등 법률 전문가의 참여를 통해 법리적 정확성, 근거 적절성, 실무 유용성 등을 종합적으로 평가하는 정성적 평가가 반드시 병행되어야 한다.
- 데이터 범위의 한계 : 본 연구는 '민사 판례'라는 특정 분야에 한정하여 실험을 수행하였다. 민사 판례는 법률 QA 연구에서 중요한 영역임에도 불구하고, 형사나 행정 등 다른 법률 분야에서는 문서 구조와 용어가 다르기 때문에 본 연구에서 제안한 최적화 기법이 동일한 효과를 낼지 검증되지 않았다. 따라서 제안 시스템의 범용성을 확보하기 위해서는 다양한 법률 도메인의 데이터를 활용한 추가 실험과 검증이 필요하다.

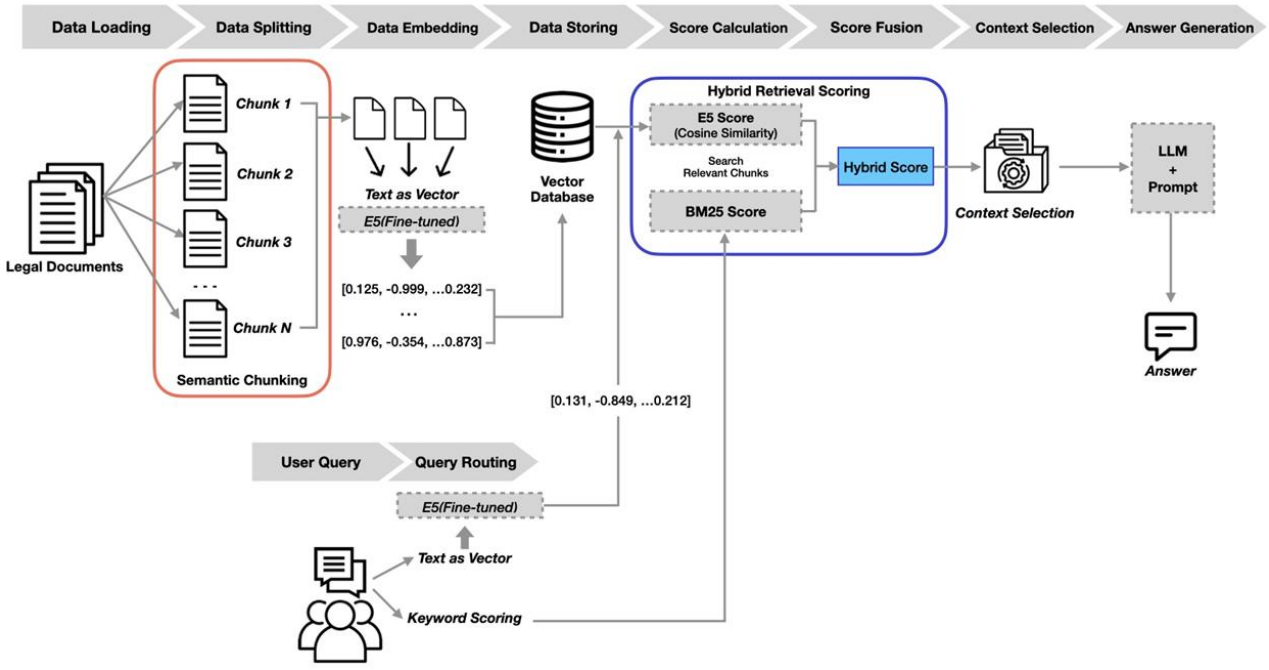


Fig. 5. Overall Architecture of the Proposed System

REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. Adv. Neural Inf. Process. Syst., pp. 9459-9474, 2020.

[2] N. Pipitone and G. H. Alami, "LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain," arXiv, *https://arxiv.org/abs/2408.10343*, 2024.

[3] Y. Kim, Y. Choi, E. Choi, J. Choi, H. J. Park, and W. Hwang, "Developing a Pragmatic Benchmark for Assessing Korean Legal Language Understanding in Large Language Models," in Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 5573-5595, Miami, Florida, USA, Nov. 2024. DOI: 10.18653/v1/2024.findings-emnlp.319

[4] W. Zhang and J. Zhang, "Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review," Mathematics, Vol. 13, No. 5, p. 856, Mar. 2025. DOI: 10.3390/math13050856

[5] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Inf. Process. Manage., Vol. 24, No. 5, pp. 513-523, Jan. 1988.

[6] S. Robertson and S. Walker, "Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval," in Proc. SIGIR, pp. 232-241, London, U.K., Aug. 1994.

[7] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-T. Yih, "Dense Passage Retrieval for Open-Domain Question Answering," in Proc. EMNLP, pp. 6769-6781, Online, Nov. 2020. DOI: 10.18653/v1/2020.emnlp-main.550

[8] L. Wang et al., "Text Embeddings by Weakly-Supervised Contrastive Pre-training," arXiv preprint arXiv:2212.03533, 2022. DOI: 10.48550/arXiv.2212.03533.

[9] J. H. Gan, "jhgan/ko-sbert-multitask," Hugging Face, *https://huggingface.co/jhgan/ko-sbert-multitask*, Dec. 24, 2021.

[10] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. EMNLP-IJCNLP, pp. 3982-3992, Hong Kong, China, Nov. 2019. DOI: 10.18653/v1/D19-1410

[11] Y. Luan et al., "Sparse, Dense, and Attentional Representations for Text Retrieval," Trans. Assoc. Comput. Linguist. (TACL), Vol. 9, pp. 329-345, 2021. DOI: 10.1162/tacl_a_00369

[12] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets straight out of Law School," in Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2898-2904, Online, Nov. 2020. DOI: 10.18653/v1/2020.findings-emnlp.261

[13] W. Cao, J. Wang, Y. Zheng, L. Bao, Q. Zheng, T. Berg-Kirkpatrick, R. Paturi, and L. Bergen, "Efficient Full-Context Retrieval for Long Documents," OpenReview, *https://openreview.net/forum?id=NJUzUq2OIi*, 2025.

[14] D. Singh, M. N. Martinez, B. J. Dorr, and S. S. Galunder, "SLIDE: Sliding Localized Information for Document Extraction," arXiv, *https://arxiv.org/abs/2503.17952*, Mar. 2025. DOI: 10.48550/arXiv.2503.17952

[15] ZHONG, Zijie, et al. Mix-of-granularity: Optimize the chunking granularity for retrieval-augmented generation. arXiv preprint arXiv:2406.00456, 2024.

[16] L. Kim, G. Lee, S. Choi, J. Lee, K.-Y. Kwahk, and N. Kim, "Semantic Document Segmentation Using Language Models," Proceedings of KIIT Conference, pp. 102-104, Jeju, Korea, May 2024.

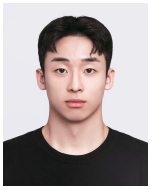
[17] Y. Kim, Y. Choi, E. Choi, J. Choi, H. J. Park, and W. Hwang, "Developing a Pragmatic Benchmark for Assessing Korean Legal Language Understanding in Large Language Models," arXiv preprint arXiv:2410.08731, 2024.

[18] AI Hub, "Legal Regulation Text Analysis Data (Advanced) - Case Law Data by Situation," AI Hub, *https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=71723*, Dec. 4, 2024.

[19] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," in Proc. Eur. Chapter Assoc. Comput. Linguist. (EACL), pp. 150–158, St. Julians, Malta, Mar. 2024.

[20] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," arXiv preprint arXiv:1904.09675, 2020.

Authors



Jun-Won Seo received an A.S. degree in Computer Science from Inha Technical College in 2025 and is currently pursuing a B.S. degree in Computer Science at Inha Technical College, Incheon, Korea. He is interested in natural language processing, retrieval-augmented generation (RAG).

Junghye Min received the B.S. degree in mathematics from Ewha Women's University, Seoul, Korea, in 1995, and the M.E. and Ph.D. degrees in computer science and engineering from Pennsylvania State University, U.S.A. in 2003 and 2005, respectively. From 2005 to 2021 she was a Principal Engineer with Samsung Research, Samsung Electronics, Seoul, Korea. Dr. Min joined the faculty of the Department of Computer Science at Inha Technical College, Incheon, Korea in 2022 and is currently an assistant Professor. Her research interests include image enhancement, image style transfer, and natural language processing.