

기록관리기준표·규정 기반 RAG 질의응답 시스템에서 질의 유형별 BM25와 임베딩 검색 성능 비교 분석

Comparative Analysis of BM25 and Embedding Retrieval Performance by Question Type in a Records Schedule and Regulations-Based RAG(Retrieval-Augmented Generation) Question Answering System

최 유 나 (Yuna Choi)* 박 종 도 (Jong-Do Park)**
이 해 인 (Haerin Lee)*** 조 재 인 (Jane Cho)****

목 차

- | | |
|-------------------|------------|
| 1. 서 론 | 4. 실험 결과 |
| 2. 이론적 배경 및 관련 연구 | 5. 논의 및 결론 |
| 3. 연구의 설계 | |

초 록

RAG(Retrieval-Augmented Generation) 구조는 지식기반을 검색한 후 이를 참조해 생성형 AI가 응답을 생성하는 기술로, 신뢰성 있는 답변을 제공할 수 있다. 그러나 RAG 기반 질의응답 시스템은 지식기반의 검색 성능에 따라 답변의 품질이 달라질 수 있다. 특히 구조화된 속성이나 도메인 고유 용어로 구성되어 있는 전문 영역의 규범을 지식기반으로 하는 경우에는 의미기반의 임베딩 방식이 무조건 우수한 성능을 보장할 것이라고 보기 어렵다. 따라서 본 연구는 대학기록관의 기록관리기준표와 규정을 기반으로 한 지식기반을 구축하고, 동일한 생성형 AI 모델을 사용하되, 검색 방식이 BM25와 임베딩으로 상이한 다른 두 개의 시스템을 구축하였다. 그리고 검색 방식이 이용자의 질의 유형과 지식기반의 구조에 따라 어떻게 다르게 나타날 수 있는지 비교·평가해 보았다. 실험 결과, 직답형과 오류전제형 질의에서는 BM25에서 우세한 경향이 나타났으며, 용어변형형과 추론형 질의에서는 임베딩 방식에서 우세한 경향이 나타났다.

ABSTRACT

The Retrieval-Augmented Generation structure is a technology that generates responses by referencing a knowledge base, thereby supporting reliable responses. However, the quality of answers from a RAG structure can vary depending on the search performance of the knowledge base. In particular, when the knowledge base consists of norms in various specific domains composed of structured attributes or domain specific vocabulary, it is difficult to assume that embedding-based methods will unconditionally guarantee superior performance. Thereby, this study created a knowledge base based on the rules and regulations of a university's archive. Then, two different systems using the same LLM were built, with two different retrieval methods, BM25 and embedding. And then, the study compared and evaluated how search methods can vary depending on the user's query type and the structure of the knowledge base. Experimental results showed that BM25 demonstrated superior performance direct-answer and error-premising queries, where exact keyword matching is crucial, while the embedding method showed superior performance in inference and term-variation queries.

키워드: 생성형 AI, 검색증강생성, 기록관리기준표, 임베딩, BM25

Generative AI, RAG (Retrieval-Augmented Generation), Records Retention Schedule, Embedding, BM25

* 인천대학교 문헌정보학과 대학원 기록관리전공 석사과정
(cya3327@naver.com / ISNI 0000 0005 3053 4537) (제1저자)

** 인천대학교 문헌정보학과 부교수(jdp23@inu.ac.kr / ISNI 0000 0004 7358 748X) (공동저자)

*** 인천대학교 문헌정보학과 대학원 기록관리전공 석사과정
(seain2729@inu.ac.kr / ISNI 0000 0005 3077 9374) (공동저자)

**** 인천대학교 문헌정보학과 교수(chojane@inu.ac.kr / ISNI 0000 0004 6016 5878) (교신저자)
논문접수일자: 2026년 7월 16일 최초심사일자: 2026년 8월 2일 게재확정일자: 2026년 8월 5일
한국문헌정보학회지, 60(3): 125-143, 2026. <http://dx.doi.org/10.4275/KSLIS.2026.60.3.125>

* Copyright © 2026 Korean Society for Library and Information Science

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (<https://creativecommons.org/licenses/by-nc-nd/4.0/>) which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.

1. 서론

대학기록관의 운영 목적은 행정, 연구, 교육 과정 등에서 생산된 기록을 체계적으로 관리 및 보존하여 업무의 투명성과 책임성을 확보하며, 대학이 생산한 각종 지식 자산을 수집, 보존, 활용하도록 하는 것이다. 대학의 기록관리 규정은 대학이 생산한 기록을 어떻게 관리, 보존할 것인지에 대한 운영 원칙과 절차를 제공하며, 기록관리기준표는 단위 업무의 보존기간과 관리 기준을 제시한다. 이 두 가지 규범은 대학 기록관리의 일관성을 유지하고 근거 기반 업무 수행을 가능하도록 하는 필수적인 도구라고 할 수 있다. 한편 대학기록관은 대개 기록관리전문요원 1인 중심으로 운영되면서(서주은, 이성신, 2021), 대체 복무 요원과 계약직 인력의 순환 근무, 그리고 처리과 담당자의 잦은 변경이라는 구조적, 조직적 한계를 가지고 있다. 이러한 현상은 대학 현장에서 업무 담당자의 기록관리 규범에 대한 이해도가 축적·유지되는 것을 어렵게 만들고 개인적 경험이나 해석에 의존할 가능성을 높일 수 있다. 또한 단순하고 반복적인 조직 구성원의 질의가 기록관리전문요원에게 집중되어 업무적 부담을 가중시킬 수 있다.

따라서 RAG(Retrieval-Augmented Generation) 구조를 적용한 생성형 AI 기반의 기록관리기준표·규정 질의응답 시스템이 구축된다면, 개인의 기억과 경험이 아닌 규범에 근거한 빠른 판단을 지원할 수 있으며, 대체복무자, 처리과 관련자 등 기록관리와 관련된 각종 규범에 대한 질문에 대하여 신속하고 즉각적인 응답 제공이 가능하다. RAG은 거대언어모델이 답변을 생

성할 때 학습된 지식에 의존하지 않고 지식기반에서 관련 문서를 검색해서 그 내용을 근거로 응답을 생성하는 구조이다. 그러나 RAG 기반 질의응답 시스템은 검색 단계를 어떻게 설계하는가에 따라 답변의 품질에 차이가 나타날 수 있다. RAG 구조는 지식기반의 검색 단계에서 얼마나 적절한 정보를 선별할 수 있는지에 크게 의존하기 때문이다. 지식기반을 검색하는 방식은 크게 키워드 매칭 방식과 의미기반 임베딩 검색 방식이 존재하는데, BM25와 같은 키워드 매칭 방식은 지식기반에 질문에 포함된 용어가 명시적으로 존재할 경우 높은 정확도를 보이나 유사 개념 또는 동의어가 사용된 경우에는 탐색하지 못한다. 반면 임베딩 기반 검색은 질문과 규범을 벡터 공간에 의미적으로 표현하기 때문에 맥락상 비슷한 개념을 찾을 수 있으나, 규정에 직접적으로 일치하지 않는 유사 조항이나 항목을 검색 결과로 제시할 가능성도 가진다.

이러한 맥락에서 본 연구는 동일한 지식기반과 생성형 AI 모델을 사용하되 검색 방식을 BM25와 임베딩 기반으로 구분한 두 개의 RAG 기반 질의응답 시스템을 개발하고 검색 방식이 이용자의 질의 유형과 지식기반의 구조에 따라 어떻게 다르게 나타나는지 비교·평가해 보고자 한다. 구체적으로 이용자의 질의를 직답형, 용어변형형, 추론형, 오류전제형으로 구분해 검색 방식별 성능 차이를 분석할 뿐 아니라, 구조화된 형태를 가지는 기록관리기준표와 서술형 구조를 가지는 규정을 대상으로도 검색 방식에 따라 성능에 차이를 보이는지 비교·평가해 보고자 한다. 본 연구는 이와 같은 목적을 달성하기 위하여 다음과 같은 연구문제를 설정한다. 첫

째, 동일한 기록관리규범에 대한 지식기반을 사용하는 RAG 기반 질의응답 시스템에서 BM25와 임베딩 검색 방식은 질문 유형(직답형, 용어 변형형, 추론형, 오류전제형)에 따라 성능 차이를 보이는가? 둘째, BM25와 임베딩 검색 방식은 기록관리기준표와 규정이라는 서로 다른 구조의 규범 문서를 대상으로 하는 질의에 있어 성능 차이를 보이는가?

2. 이론적 배경 및 관련 연구

2.1 대학 기록관리 규범

행정적 책임성 및 투명성 확보와 대학의 역사적 가치 보존을 목적으로 하는 대학 기록관리는 「공공기록물 관리에 관한 법률(이하 공공기록물법)」에 근거하여 수행되는 법적 의무이다(공공기록물 관리에 관한 법률, 법률 제21066호). 「공공기록물법 시행령」 제3조 제5호는 「유아교육법」, 「초·중등교육법」 및 「고등교육법」, 그 밖에 다른 법률에 따라 설립된 각급 학교를 공공기관의 범위에 포함하고 있어, 국립·사립을 불문하고 모든 대학이 해당 법령의 적용을 받는다(국가기록원, 2026b, 8). 이에 따라 대학의 교직원들은 공공기관의 임직원으로서 법 제4조에 명시된 기록물 보호 및 관리의 의무를 지니며, 연간 기록물 생산량이 1천 권 이상이거나 보존 대상 기록물이 5천 권 이상인 대학은 법 제13조 및 같은 법 시행령 제10조에 따라 기록관 설치 대상이 된다(국가기록원, 2026b, 10-12). 또한 기록관에는 법 제41조 및 같은 법 시행령 제78조에 따라 전체 정원의 4분의 1 이상을 기

록물관리 전문요원으로 배치하여야 한다(국가기록원, 2026b, 237). 이러한 제도적 장치에 의해 대학은 자체적인 기록관리 규범을 수립하여 행정의 투명성을 확보하고 기록관리 업무의 일관성과 체계성을 유지하고 있다.

다만 대학 기록관리는 일반 행정기관과 차별화되는 고유한 특수성도 함께 지닌다. 대학은 교육, 연구, 행정이라는 복합적인 기능을 수행하기 때문에 기록 생산의 주체와 유형이 다양하고 대학의 정체성 확립을 위한 역사 자료 아카이브 기능도 동시에 수행하고 있기 때문이다. 더불어 국가적 보존가치가 높아 영구기록물관리기관이 이관을 요청하는 경우를 제외하고는 대학의 기록물을 대학이 자체적으로 보존·관리하기 때문에 영구기록물에 대한 높은 자율성과 관리 책임도 가진다(국가기록원, 2026b, 19).

대학기록관 운영에 필요한 핵심 규범은 기록관리규정, 기록관 운영규정, 기록관리기준표이다. 「2026년 대학 기록물관리 지침」에 따르면 법 제13조 및 시행령 제10조를 근거로 대학기록관은 기록관리규정과 기록관 운영규정을 제정하여야 하는데, 기록관리규정은 임직원의 기록관리 의무와 역할, 관리 대상 기록물의 범위, 기록물의 등록·분류 등 생산 절차, 기록물 정리와 생산현황 보고, 기록관으로의 이관, 열람, 평가·폐기에 이르는 관리 절차를 정하며, 기록관 운영규정은 기록관의 기능과 조직·인력 구성, 기록물 인수·재분류·평가·보존 등 기록관이 수행하는 제반 업무를 규정한다(국가기록원, 2026b, 28). 이는 상위 법령과 기관 내부 규정이 결합된 구조로서 법적 틀은 「공공기록물법」이 정하며, 대학의 특수성을 반영한 운영 규정은 대학이 직접 제정한다. 더불어 업무의 절

차적 규범과 행정 책임의 범위를 제시하고 있어 기록관리전문요원의 근거 기반 업무 수행이나 업무 담당자의 기록물 생산, 이관, 공개 여부와 같은 관리 행위에 있어 매우 중요한 기준이 되고 있다.

한편, 기록관리기준표란 정부기능분류체계(BRM)에 따라 단위과제별로 작성하여 각각의 단위과제에 대한 관리기준을 제시한 핵심 도구이며, 「공공기록물법 시행령」 제25조 제1항에 따라 업무설명, 보존기간 및 보존기간 책정사유, 비치기록물 해당 여부, 보존장소, 보존방법, 공개 여부 및 접근 권한 등을 포함한다(공공기록물 관리에 관한 법률 시행령, 대통령령 제35506호). 단위과제는 해당 표의 기본 구성단위이며, 정부기능분류체계(BRM)의 최하위 단위이자 행정업무 수행의 기준으로 업무 간 유사성과 독자성을 고려하여 소기능을 세분화한 업무 영역 단위로 정의된다. 기록관리기준표의 분류체계는 정부기능분류체계(BRM) 적용을 원칙으로 하며, 이는 정책분야-정책영역-대기능-중기능-소기능-단위과제로 이어지는 6단계 계층 구조를 형성한다(국가기록원, 2026a, 58-59). 다만 같은 조 제2항은 기록관리기준표를 정부기능분류체계의 단위과제별로 작성하는 것을 원칙으로 하되, 정부기능분류체계를 적용할 수 없는 경우에는 관할 영구기록물관리기관의 장과 협의하여 별도의 기능분류방식을 사용할 수 있다. 대학과 같이 교육·연구·행정 기능이 복합적으로 수행되는 기관의 경우 BRM 체계를 그대로 적용하기 어려운 경우가 많으며, 실제 운영에서는 관할 영구기록물관리기관과 협의를 통해 별도의 기능 분류 방식을 사용하는 경우가 일반적이다. 조직의 기능과 기록물

이 생산된 업무 맥락을 구조적으로 반영하는 기능 중심의 계층적 분류체계는 대학 기록관리의 체계성 및 일관성 확보와 기록정보의 효과적 탐색과 통제를 도모할 수 있다.

이렇게 대학 기록관리기준표와 규정은 기록관리 업무에 근간이 되는 핵심 규범이지만, 기준표는 단위과제, 처리부서, 보존 기간과 같은 구조화된 속성으로 구성되어 있으며, 규정은 기록관리에 관한 절차와 기준을 서술형으로 제시하고 있다. 지식기반으로 활용될 수 있는 이러한 규범의 구조적 특성은 질의의 유형과 함께 질의응답 시스템의 검색 성능에 영향을 미칠 수 있기 때문에, 검색 전략 수립에 있어 중요한 고려 요소가 될 수 있다.

2.2 RAG 개념 및 지식기반 검색 방식

대규모 언어모델(LLM)은 사전 학습된 지식을 바탕으로 답을 생성하기 때문에 최신 정보나 특정 분야의 지식이 충분히 반영되지 않을 수 있다. 또한 생성 단계에서 근거에 없는 내용이 나타나는 환각 현상이 일어날 수도 있다(Martineau, 2023). 이러한 언어 모델의 한계를 보완하기 위하여 외부 지식기반을 참고해 답을 생성하는 RAG 구조를 활용할 수 있다. RAG 구조는 검색 단계와 생성 단계로 나눌 수 있다. 검색 단계에서는 사용자 질문을 벡터로 만들고 그 벡터를 이용해 지식기반에서 비슷한 문서를 찾는다. 다음 단계에서는 검색한 문서를 근거로 답변을 합성해 이를 활용해 답변을 제공하게 된다. 이 구조는 생성형 AI 모델을 활용하여 실제 외부 데이터에 대한 질의 응답을 제공하므로 환각을 줄이고 신뢰성 있는 답변을

할 수 있다(Gao et al., 2023).

그러나 검색 단계에서 적절한 문서를 정확히 찾지 못하면, LLM이 외부 지식 대신 사전 학습한 내부 지식을 사용해 답변을 생성하게 된다. 그 과정에서 실제 근거가 없는 내용을 생성할 수 있어, 시스템의 신뢰도를 떨어뜨릴 수 있다(Martineau, 2023). 따라서 RAG 기반 질의응답 시스템의 성능은 생성 모델의 능력뿐만 아니라, 검색 단계에서 얼마나 정확하고 적절한 문서를 골라내는지가 성능에 영향을 줄 수 있다. 일반적으로 RAG 구조에서 지식기반의 검색에는 BM25 기반의 키워드 검색과 임베딩 기반의 의미 유사도 검색이 활용되고 있다.

BM25는 전통적인 정보검색 기법이다. 특정 단어가 문서에 얼마나 자주 나타나는지와 문서의 길이를 기준으로 점수를 계산한다. 구체적으로 이 알고리즘은 질의에 들어 있는 단어가 문서에 얼마나 자주 나타나는지(TF: term frequency), 그 단어가 얼마나 드문지(IDF: inverse document frequency), 그리고 문서 길이를 보정한 값을 모두 고려해 관련성 점수를 매긴다(Manning et al., 2008). 이 방식은 용어가 정확해야 하는 전문 영역의 규범과 행정 문서 검색에서 강점을 가지며, 정해진 형식과 고유 용어가 반복되는 검색 환경에서 이용자가 의도한 정보를 정확히 찾아낼 수 있다.

한편, 임베딩은 문장이나 문서의 의미를 고차원 숫자인 벡터로 바꾸는 기술이다. 이는 문장의 의미를 숫자로 바꾸고, 숫자를 방향과 거리로 나타낸다. 그 다음 숫자를 공간에 좌표값으로 매핑함으로써 문장 사이의 의미가 얼마나 비슷한지를 계산하기 위한 핵심기술이라고 할 수 있다(강한훈 외, 2024). 이 방식은 단어의

형태가 다르더라도 문맥적 의미가 유사하다면 벡터 공간 내에서 인접하게 배치되도록 하여 의미기반 검색을 가능하게 한다. 그렇기 때문에 자연어 질의의 다양한 표현 방식에 유연하게 대응할 수 있으며, 동일한 용어가 사용되지 않더라도 관련성이 높은 문서를 효과적으로 검색할 수 있다.

정리하자면, BM25는 질의와 문서에 포함된 용어의 일치 정도를 중심으로 관련성을 계산하므로 동일한 용어가 명시적으로 포함되어 있을 경우 높은 점수를 부여하지만, 임베딩 기반 검색은 의미적 유사성을 계산하는 방식이므로 동일한 용어가 사용되지 않더라도 의미적으로 유사한 문서를 검색할 수 있다. 그러나 기록관리 규범과 같이 도메인 고유의 통제 용어가 중심이 되는 지식기반에서는 정확한 조항과 단위과제를 식별하는 것이 중요하므로 의미적 유사성만으로는 우수한 검색 결과를 보장하기 어렵다.

2.3 선행연구

최근 법률, 금융, 의료 분야 등에서 검색과 생성 기술을 결합한 질의응답 시스템을 만들기 위해 활발히 연구가 진행되고 있다. 이동현 외(2025)는 이 방법으로 금융용어 질의응답 시스템을 개발하였다. 연구팀은 ChatGPT 모델을 기반으로 시스템을 설계하였고 한국은행이 제공한 경제·금융용어 700개와 TESAT 핵심 문항 70개를 데이터로 사용하였다. 유경도와 임희정(2025)도 비슷한 방법으로 건강 관리 질의응답 프로그램을 개발하였다. AIHub 한국어 의료 질문답변 데이터를 바탕으로 지식기반을 만들고 이를 근거로 질문에 대한 답을 생성하는 시스

템을 설계하였다. 한편, 신은성 외(2025)는 방법을 중심으로 신뢰할 수 있는 답변을 생성하는 RAG 시스템을 개발하였다. 법률문서는 장-절-조 계층 구조와 조문 간 참조 관계 같은 구조적 특성을 가지고 있는데, LangGraph를 통해 질의 분석, 검색, 생성 과정을 구현하였다. 이 연구에서 눈에 띄는 점은 지식기반 검색에 키워드 매칭과 임베딩 방식을 동시에 사용했다는 것이다. 조문 제목을 기준으로 한 질의에서는 위 두 가지 방식을 함께 사용하였으며, 추론을 포함한 복합적인 질의에 대한 처리를 위해서는 온톨로지를 활용하였다.

한편, 본 연구와 같이 대학 안에서 사용할 수 있는 챗봇 구현 사례도 아래와 같이 살펴볼 수 있다. 오재광과 김수균(2025)은 대학의 학사 규정, 수강 편람, 입시 요강 같은 형식이 정해지지 않은 행정 문서를 빠르고 정확하게 활용하기 위해 검색과 생성이 결합된 질문 답변 시스템을 구현하였다. 이들이 진행한 연구에서는 BM25와 임베딩을 결합하여 정보를 찾아내도록 설계하였으며, 구현된 시스템은 대학 구성원에게 필요한 정보를 빠르고 정확하게 제공함으로써 행정 업무의 편리함을 높일 수 있다고 보고하였다. 이현섭(2025)도 대학 환경에서 검색과 생성이 결합된 기술을 이용하여 학사 정보와 규정에 관한 질문 답변 시스템을 설계하였다. 제안된 RAG 기반 챗봇은 규정 내용의 정확성을 유지하면서도 자연스러운 대화형 응답을 제공하여 사용자 만족도와 정보 접근성을 향상시킬 수 있다고 하였다. 그 밖에 김선욱(2025)도 대학기록관 이용안내와 FAQ를 중심으로 자연어 질의 응답이 가능한 RAG 기반 챗봇을 구축하고 그 성능을 평가한 바 있다. 이 연구는 기록관리 현

장을 대상으로 한다는 점에서 본 연구와 유사하지만, 이용자 서비스 지원을 목적으로 하고 있다는 점에서 차이가 있다.

앞에서 살펴본 대로 최근 신뢰할 수 있는 질의응답 시스템을 만들기 위해 RAG 구조를 사용한 사례가 여러 분야에서 보고되고 있다. 대부분의 사례는 임베딩 방식을 전제로 설계하고, RAG 구조의 구현과 활용 가능성을 검증하고 있지만, 법률과 규범을 지식기반으로 하는 일부 사례에서는 BM25와 임베딩을 같이 쓰는 방법을 선택하기도 한다(신은성 외, 2025; 오재광, 김수균, 2025). 이들 사례에서는 단순 임베딩만으로는 구조화된 속성과 고유 용어를 사용하는 지식기반에서 답변의 품질을 보장하기 어렵고 확실히 처리만으로는 최상의 성능을 내기 어렵다고 주장하고 있다. 또한 복잡한 문서 구조나 긴 문맥이 포함된 질의에 대해서는 여전히 환각 현상이 관찰되고 있음을 보고하면서(오재광, 김수균, 2025), 향후 특정 도메인 고유 용어의 유사 표현 처리, 복잡한 참조 관계 처리 등의 과제가 아직 남아 있음을 언급하고 있다.

3. 연구의 설계

본 장에서는 지식기반 구축, BM25와 임베딩 기반의 RAG 기반 질의응답 시스템 구조 설계 및 개발, 질문 유형의 정의, 실험 설계 및 평가 방법에 대하여 설명하도록 한다.

3.1 지식기반 구축

먼저 지식기반 구축을 위한 대상 문서는 I 대

학 기록관리기준표와 기록관운영규정이다. 이 두 개의 규범을 대상으로 CSV 기반의 지식레코드를 설계하되, 기록관리기준표는 담당부서, 단위과제명, 보존기간, 기록물철에 대한 정보를 문서 단위로 구성해 단일 셀에 텍스트로 패키징하였다. 즉, 단위과제를 하나의 독립된 검색 단위로 설정하고 관련 속성 정보를 하나의 청크로 묶어, 기준표의 구조가 검색 단위에 반영되도록 하였다. 규정에는 기록관 조직 및 분장 업무, 기록물의 관리와 열람 등 기록관리 규정과 기록관 운영 규정에 관한 주요 내용이 포함되어 있는데, 개별 조문을 하나의 문서 단위로 설정하고 조제목과 본문 내용을 하나의 청크로 구성하였다. 조문 하나를 전체로 묶은 이유는 하나의 독립적인 의미 단위로 정의, 조건, 절차, 책임 주체 등을 포함해 의미적 완결성이 훼손되지 않도록 하기 위함이다. 기록관리기준표의 평균 문서 길이는 약 89자이며, 규정은 244자로 지식기반의 문서가 비교적 짧고 자체적으로 완결된 의미 단위를 형성하므로 별도의 청킹은 수행하지 않았다. 한편 기록관리기준표의 경우 단위과제 고유 목록을 기반으로 총 426개의 독립된 문서 단위를 구축하였으며, 규정은 조 단위 조항 28개를 문서 단위로 구축하였다.

3.2 질의응답 시스템의 구조

본 연구의 질의응답 시스템은 RAG 구조를 기반으로 하며, 검색 단계만 상이하고 생성 단계는 동일한 2개의 시스템으로 구축하였다. 검색 단계에서는 BM25 방식의 키워드 매칭과 임베딩 기반의 의미 유사도 검색 방식을 각각 사

용하되, 지식기반의 양이나 문서의 구성이 결과에 미치는 영향을 통제하고 순수하게 검색 방식에 따라 나타날 수 있는 성능 차이를 비교하였다. 시스템 개발 과정에서 Anaconda와 Python 3.14를 사용하였고, LLM 엔진은 LM Studio를 사용하였다. 그리고 생성 모델은 Mistral 계열 14B 규모 언어모델을 사용하였다. Mistral은 오픈소스 기반의 대형 언어 모델로 이를 로컬 환경에 다운로드하고 직접 실행하였다. 시스템은 macOS가 설치된 Apple Mac Studio(M3 Ultra, unified memory 512GB)에서 개발되었다. 시스템의 주요 파이프라인은 다음과 같다. 이용자가 BM25와 임베딩 시스템에 각각 자연어 질문을 입력하면, 동일한 지식기반을 사용해 가장 높은 순위의 문서 한 개를 선택하고 결과를 반환한다. 이후 생성 단계에서 검색된 문서를 LLM(Mistral)에 입력하여 검색된 문서를 근거로 답변 생성하는 방식이다. BM25 기반 시스템과 임베딩 기반 시스템의 파이프라인을 자세히 설명하면 다음과 같다.

3.2.1 BM25 기반 시스템

이 시스템은 지식기반에 구축된 문서를 사전에 토큰 단위로 계산하고 토큰화 결과를 메모리에 저장해 두었다가 검색 시 재활용함으로써 반복 과정을 줄이고 검색 속도를 향상시키도록 하였다. 전체 파이프라인은 <그림 1>과 같이 지식기반 로딩 - 문서토큰화 및 캐시저장 - 문서통계사전계산 - 사용자질의 - 질의토큰화 - BM25유사도계산 - TOP1문서선택 - 프롬프트구성 - LLM생성 - 답변반환의 순서로 작동한다. 지식기반은 CSV 문서 집합으로 구성되며, Kiwi 형태소 분석기(Kiwipiepy)를 이용하

여 문서를 사전에 토큰화하였다. 그리고 빠른 유사도 계산을 위해 토큰의 빈도(TF), 문서 빈도(DF), 문서 길이를 계산해 두고, 이용자가 질의를 입력하면 질의를 토큰화한 후 지식기반에 구축된 모든 문서와 비교해 BM25 점수를 계산하여 상위 1개의 문서를 선택하도록 하였다. 한편 생성 단계에서는 로컬 LLM(Mistral)을 호출하여 답변을 생성하도록 하였다. 프롬프트에는 근거 없는 정보를 추가하지 않도록 지시하여 결과가 근거를 기반으로 생성되도록 설계하였으며, 근거와 함께 최종 답변을 출력하도록 구성하였다.

3.2.2 임베딩 기반 시스템

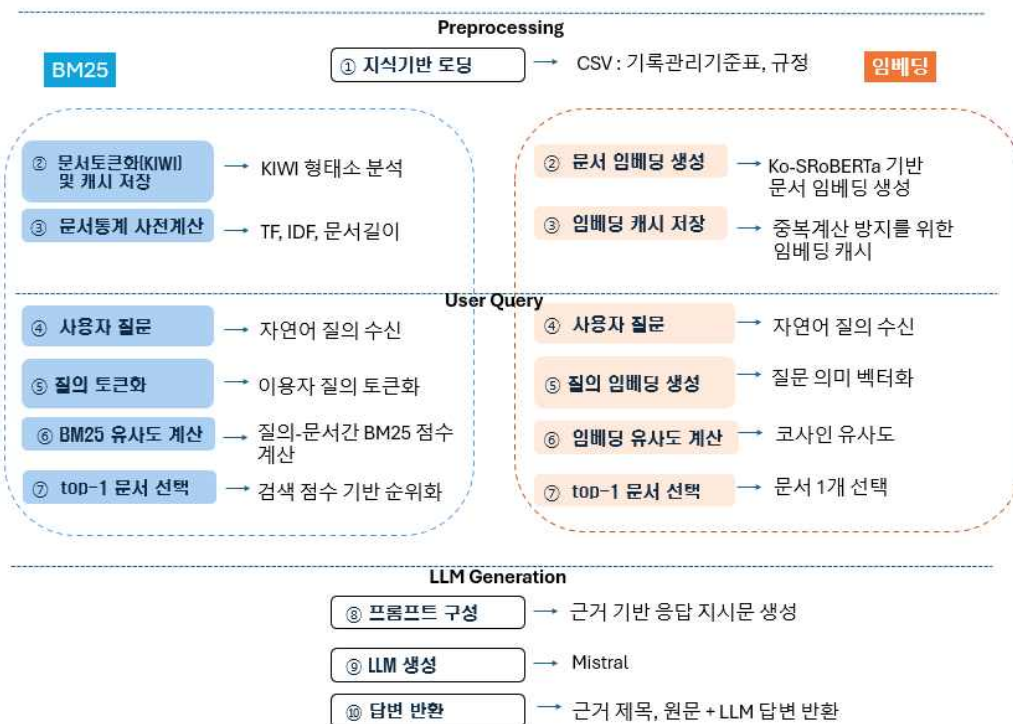
임베딩 기반의 의미 유사도 질의응답 시스템은 한국어 문장 유사도 태스크에 특화된 Ko-SRoBERTa(jhgan/ko-sroberta-multitask)를 사용하였다(Gan, n.d.). Ko-SRoBERTa는 한국어 문장 유사도 및 자연어 추론 데이터셋을 기반으로 학습된 문장 임베딩 모델로, 문장 간 의미 유사도를 반영하는 벡터 표현을 생성한다. 이 시스템은 BM25 시스템과 마찬가지로 검색 효율성을 확보하기 위하여 문서 임베딩을 사전에 계산하여 캐쉬 형태로 저장하는 구조를 마련하였다. 즉 CSV 파일의 캐시 키를 생성하여 지식기반에 변경이 없는 경우 기존 임베딩을 사용하도록 함으로써 중복 연산을 방지하였다. 또한 지식기반 문서의 임베딩 벡터를 정규화하여 벡터 크기의 영향을 제거하고 벡터 간 방향 유사도를 기반으로 코사인 유사도를 계산하였다. 이 시스템의 전체적인 파이프라인은 <그림 1>과 같이 지식기반로딩 - 문서임베딩생성 - 임베딩캐시저장 - 사용자질문 - 질의임베딩생성

- 임베딩유사도계산 - TOP1문서선택 - 프롬프트구성 - LLM생성 - 답변반환 순으로 이루어진다. 즉 이용자가 질문을 자연어로 입력하면 임베딩 모델을 사용해 질의 임베딩을 생성하고 사전에 계산된 문서의 임베딩과 코사인 유사도를 계산하여 가장 높은 유사도를 보이는 문서를 검색 결과로 반환시키는 방식이다. 그다음 앞의 모델과 동일하게 로컬 LLM(Mistral)을 호출하여 답변을 생성하는데, 프롬프트에는 근거 없는 정보를 추가하지 않도록 지시하여 결과가 근거를 기반으로 생성되도록 설계하였으며, 근거와 함께 답변을 출력하도록 구성하였다.

요약하면 두 시스템은 동일한 지식기반 CSV를 사용하여 검색 결과로 선택된 문서를 LLM으로 보내 동일한 프롬프트로 응답을 생성해 답변을 반환한다. 그러나 검색이 이루어지는 단계에서만 다음과 같이 차이를 가진다. 첫 번째, 검색을 위한 사전 처리 방식에 있어서 BM25 시스템은 구축된 지식기반의 문서 단위를 명사 중심으로 토큰화하여 빈도와 길이 등의 통계를 미리 계산해 놓는 반면, 임베딩 시스템은 문서 내용을 임베딩 모델을 이용해 벡터로 변환한 후 이를 사전에 저장해 두는 차이를 가진다. 두 번째, 사용자의 자연어 질의에 있어 BM25 시스템은 명사를 기반으로 토큰화해 질의와 문서 간의 유사도를 BM25 점수로 비교한다면, 임베딩 시스템은 질의를 임베딩 벡터로 변환해 사전에 생성된 문서 임베딩과 비교해 가장 유사한 문서를 선택한다는 점에서 차이가 있다. 두 시스템의 파이프라인과 검색 방식의 차이는 아래 <표 1>, <그림 1>과 같다.

〈표 1〉 BM25 및 임베딩 기반 기록관리 RAG 질의응답 시스템의 실험 환경 및 구성 비교

구분	BM25 기반 RAG	임베딩 기반 RAG
운영체제	macOS	
CPU	Apple M3 Ultra	
GPU	M3 Ultra Integrated GPU	
Unified Memory	512 GB	
개발 환경	Anaconda	
프로그래밍 언어	Python 3.14	
LLM 실행 엔진	LM Studio	
생성 모델	Mistral-3-14B-Reasoning-2512 Q4_K_M	
지식기반	CSV 기반 기록관리기준표·규정	
검색 방식	키워드 매칭 기반	의미 유사도 기반
주요 검색 모델	BM25	jhgan/ko-sroberta-multitask(한국어임베딩 모델)
입력 표현	형태소 기반 토큰	문장 임베딩 벡터
유사도 계산	BM25 점수	코사인 유사도
사전 처리	토큰화 및 빈도 통계 계산	문서 임베딩 사전 생성
질의 처리	질의 토큰화 후 점수 계산	질의 임베딩 생성 및 유사도 계산



〈그림 1〉 BM25 및 임베딩 검색 기반 기록관리 RAG 질의응답 시스템의 파이프라인

3.3 질문 유형 및 구성 방법

본 연구의 질문 유형은 기록관리기준표와 기록관운영규정이라는 지식구조를 기반으로 이용자의 정보 탐색 형태를 고려하여 구성되었다. 질문의 유형을 정의하기 위한 이론으로는 Taylor (2015)의 질문 형성 위계와 Belkin(1980)의 ASK (Anomalous State of Knowledge) 이론을 제시할 수 있다. Taylor는 정보 요구가 의식된 요구 (Conscious need), 형식화된 질문(Formalized need), 시스템에 맞게 조정된 질문(Compromised need)으로 점진적으로 정제되는 과정이라고 주장하면서, 사용자의 질의는 지식수준과 표현 단계에 따라 달라질 수 있다고 설명하였다. 본 연구에서는 질의의 유형을 추론형, 용어변형형, 직답형으로 구분함으로써 이용자의 질의가 의식화된 요구에서부터 정제된 질문으로 각기 다른 지식수준을 나타내는 현상을 반영하였다. 한편 Belkin(1980)은 정보 요구가 완전한 지식 상태가 아니라 지식 구조의 결손이나 불완전성에서 발생한다고 보았다. 이에 이용자의 불완전한 지식 상태가 질문의 전제로 반영되는 경우를 평가하기 위하여 오류전제형 질문을 추가하였다. 구체적으로 각 유형에 대한 정의를 설명하면 다음과 같다.

첫째, 직답형 질문은 사용자가 기록관리기준표에 명시된 정확한 명칭(예: 단위과제명, 기록물철명 등)을 인지하고 이를 그대로 사용하여 질의하는 경우를 의미한다. 이 유형은 Taylor의 질문 형성 위계상 상위에 위치하는 유형으로 시스템에 맞게 정제된 질문을 하는 단계를 의미한다. 이 질문은 명시적 사실 정보의 직접 회수를 목적으로 하며, “연구과제 관리 및 연구

지원의 보존기간은?”을 예로 들 수 있다.

둘째, 용어변형형 질문은 사용자가 지식수준이 어느 정도 정제되어 있으나 정확한 용어로 정제하지 못한 채 일상어, 비공식 표현, 축약어 등을 사용하여 질의하는 경우이다. 따라서 기준표의 통제어 또는 공식 용어로의 변환이 필수적인 질의 형태이다. 예시로 “ROTC 문서는 어느 부서 거야?”를 들 수 있다.

셋째, 추론형 질문은 Taylor의 질문 형성 위계상 하위에 위치하는 유형으로 정보 요구가 의식되고 사용자가 개념적 수준의 이해는 하고 있으나 이를 기준표의 공식 항목과 직접적으로 매핑하지 못하는 단계에서 나타나는 질문의 유형으로 정의한다. 그렇기 때문에 질의 처리 과정에서 논리적 추론이나 구조적 연결이 필요한데, 예를 들면, “도서관에서 책을 새로 구입할 때 발생하는 증빙 서류는 어디에 분류해야 해?”와 같은 질문을 예로 들 수 있다.

넷째, 오류전제형 질문은 Belkin이 지적한 바와 같이 지식 구조에 결손·모순이 존재하는 상태에서 이루어지는 질문이다. 이는 기록관리 제도나 절차에 대한 잘못된 인식을 전제로 질의하는 경우로, 단순한 정보 제공이 아니라 전제 자체를 교정하는 응답이 요구되는 경우를 의미한다. 예시로 “고등 교육통계 조사 보존기한이 30년 맞지?” 등을 들 수 있다.

한편, 각 질문은 특정 단위과제 또는 규정 조항을 기준으로 하나의 과제나 조항만으로 답변이 가능하도록 설계하였다. 이후 연구자가 기록관리기준표와 규정을 대조하여 모든 문항이 하나의 정답에 귀결하는지 검토하였다. 위의 네 가지 유형별로 각각 15개씩 총 60건의 질의를 구성하였으며, 기준표를 대상으로는 10개씩

〈표 2〉 유형별 질의와 정답 예시

유형	질의	정답
직답형	공동기기 관리 보존기한은 몇 년이야?	5년
용어변형형	학생 향의 관련 건 처리 보존 기간은?	10년
추론형	노동조합과의 협의, 회의, 합의 사항을 정리한 문서들은 인사 기록이 아니라 어느 업무 영역의 기록으로 관리해야 할까?	노동조합관리
오류전제형	대학 및 입학 홍보는 학술정보팀에서 관리하는거 맞지?	아니오, 입학관리과입니다.

총 40건, 규정을 대상으로 는 5개씩 총 20건으로 구성하였다. 〈표 2〉에서 유형별 질의와 정답 예시를 제시하였다. 앞에서 언급한 바와 같이, 본 연구에서 설계한 질의는 실제 이용자의 질의 로그를 기반으로 구축한 것은 아니다. 따라서 실험에 활용된 질의가 모든 기록관리 현장의 정보 요구를 대표하는 데에는 한계가 있을 수 있음을 밝힌다.

3.4 실험 및 평가 방법

실험은 동일한 질문을 대상으로 BM25 기반 검색과 임베딩 기반 검색 시스템에 질의를 던지고 그 결과를 비교 평가하는 방식으로 진행하였다. 검색 결과는 근거 제시의 적합성과 답변의 정확성을 기준으로 1과 0의 이진 분류로 판단하였다. 정답인 경우는 정확한 근거 문서 선택과 그 근거에 기반한 답변을 한 경우로 정의하였으며, 오답인 경우는 근거 문서가 적합하지 않거나 답변이 오류인 경우, 그리고 근거에 없는 내용을 생성한 경우로 판정하였다.

평가는 3인의 연구자가 독립적으로 수행하였으며, 사전에 확정된 정답을 기준으로 각 검색 결과가 정답과 일치하는지 판정하였다. 평가 과정에서 시스템의 종류에 대해서는 블라인드 처리하지 않았으나, 본 평가는 사전에 정의

된 정답과의 일치 여부를 기준으로 정답과 오답을 판정하기 때문에 평가자의 주관적 판단이 개입될 가능성은 크지 않다. 아울러 문항별 판정 값은 평가 결과를 기반으로 한 다수결 방식으로 결정하였으며, 각 연구자의 평가 결과에 대하여 Fleiss' kappa를 산출하여 평가자 간 일치도를 검증하였다. Fleiss' kappa는 R의 irr 패키지를 활용하여 산출하였으며, 분석 결과 BM25 시스템에 대한 평가는 $\kappa = 0.904(p < .001)$, 임베딩 시스템에 대한 평가는 $\kappa = 0.863(p < .001)$ 로 나타나 평가자 간 일치도는 매우 높은 수준임을 확인하였다. 마지막으로 SPSS 30을 이용하여 McNemar 검정을 통해 BM25 기반 검색과 임베딩 기반 검색의 전반적인 성능 차이를 분석하고 이후, 검색 정확률 비교를 통해 질문 유형과 지식기반 구조에 따른 차이를 분석하였다. 정확률은 정답을 제시한 질의 수를 해당 범주의 문항 수로 나누어 산출되었다.

4. 실험 결과

4.1 BM25와 임베딩 전체 성능 비교

먼저 60개의 전체 질의를 대상으로 한 평가 결과를 살펴보도록 한다. 〈표 3〉을 보면 두 시

〈표 3〉 BM25와 임베딩 모델의 전체 질의 응답 분포 및 McNemar 검정 결과

BM25 \ 임베딩	임베딩 오답	임베딩 정답	합계
BM25 오답	17	5 (b)	22
BM25 정답	8 (c)	30	38
합계	25	35	60

McNemar test $b=5$ $c=8$ $p=.581$

시스템에서 모두 정답인 경우가 30건, 모두 오답인 경우가 17건으로 나타났으며, BM25에서만 정답으로 나온 경우는 8건(c), 임베딩 시스템에서만 정답으로 나온 경우(b)가 5건으로 나타났다. BM25 시스템에서 정답으로 나온 경우가 더 많았지만, McNemar 검정 결과($p = 0.581$) 전체적인 정확도 면에서는 두 모델 간 유의한 성능 차이가 없는 것으로 검증되었다. 그러나 각 모델이 정답을 도출하는 양상은 질문의 유형과 특성에 달라질 수 있으므로 두 시스템의 정확률 비교를 통해 4.2장에서 심층적으로 분석해 보도록 하겠다.

4.2 질문 유형별 검색 방식의 정확률 비교

4.2.1 직답형

먼저 검색 정확률 차이를 나타낸 〈표 4〉를 기반으로 직답형 질문의 응답 성능 차이를 살펴해보도록 한다. BM25에서 정답이 나타난 사례는 13건(86.7%), 임베딩에서 정답이 나타난 사례는 11건(73.3%)으로 두 시스템 모두 우수하지만, BM25에서 정답 사례가 더 많이 나타났다. 구체적인 사례를 보면, “연구소 및 사업단 관리 보존기간은 몇 년이야?”라는 질문에 임베딩 시스템은 오답을 제시했는데 그 이유는 ‘연구사업’, ‘연구과제 관리’와 같은 유사 개념 단위과제를 후보로 함께 고려하여 단위과제 간

구분에 실패한 것으로 추정된다. 반면 BM25는 이 질문에 성공했는데, 이는 질의에 포함된 핵심 명사어와 문서 내 용어 간의 직접적인 일치 여부를 기반으로 관련성을 산출하는 BM25의 특성이 효과적으로 작용한 결과로 볼 수 있을 것이다. 이와 유사한 사례로 “감사수행 관련 기록물은 어느 부서에서 관리해?”라는 질문에 있어, 특정 행정 업무를 지칭하는 용어인 ‘감사수행’을 BM25 시스템에서는 정확히 찾아낸 반면, 임베딩 기반 시스템에서는 점검, 평가, 조사 등 감사와 관련된 유사 의미를 포괄적으로 반영하면서 정답과는 다른 문서를 상위에 반환하여 정확도가 저하된 것으로 추정된다.

4.2.2 용어변형형

용어변형형 질문에 대한 응답 결과를 살펴보면, BM25에서 정답이 나타난 사례는 4건(26.7%), 임베딩에서 정답이 나타난 사례는 7건(46.7%)으로 임베딩 시스템이 우세한 경향이 있는 것으로 나타났다. 이를 보여주는 사례로 “학생 징계 및 포상 관리 보존기간은?”과 같이 기준표의 어휘와 이용자가 던진 표현 간 불일치가 존재하는 질문의 경우, 임베딩 모델만이 정답을 도출한 것으로 나타났다. 실제 기준표 상의 단위 과제명은 ‘학생 상벌 관리’로 임베딩 시스템은 질의의 표현을 지식기반의 고유 용어로 변환하는데 성공한 것으로 보인다. 그러나 “학술지

구독 관련 업무 보존기간은?”, “학생 향의 관련 건 처리 보존기간은?”, “ROTC 기록은 어느 부서 관할이야?”와 같은 질문에 있어서는 임베딩 시스템 역시 검색에 실패하였다. 이는 임베딩 모델이 ‘학술지 구독’, ‘학생 향의’, ‘ROTC’를 지식기반 내 고유 용어인 ‘연속간행물 구독’, ‘민원사무처리’, ‘학군단’과 같은 용어로 변환하지 못했기 때문이다. ‘학술지 구독’과 ‘연속간행물 관리’는 일반적인 자연어 맥락에서 유의어 관계가 아니므로, 임베딩 모델이 두 용어 간의 의미적 연관성을 충분히 학습하지 못했을 가능성이 있다. 또한 ‘학생 향의’는 감정 행위 중심의 일상어로 ‘민원사무처리’와의 의미적 유사성이 처리되지 못했을 가능성이 있으며, ‘ROTC’ 또한 ‘학군단’이라는 고유 조직명으로 수렴하지 못했을 것으로 추정된다. 물론 사전학습 규모가 더 큰 고성능 임베딩 모델을 적용할 경우, 연관된 표현을 추가로 포착할 가능성을 가질 수는 있다. 그러나 ‘학술지 구독-연속간행물 구독’이나 ‘학생 향의-민원사무처리’와 같은 매핑 구조는 업무 기능에 기반한 비직관적 대응 관계로 범용 데이터 중심으로 학습된 대형 모델을 활용하더라도, 특수 도메인의 폐쇄적 지식 구조를 완벽히 이해하는 데에는 한계가 존재할 수 있을 것이다.

4.2.3 추론형

추론형 질문에 대한 응답 결과를 살펴보면, BM25에서 정답이 나타난 사례는 8건(53.3%), 임베딩에서 정답이 나타난 사례는 10건(66.7%)으로 이 유형의 질문에 있어서는 임베딩 검색이 우세한 것으로 보이나 전반적으로 직답형 질문보다는 검색 성능이 떨어지는 것으로 평가

된다. 구체적인 응답 사례를 살펴보면, “노동조합과의 협의, 회의, 합의 사항을 정리한 문서들은 인사 기록이 아니라 어느 업무 영역의 기록으로 관리해야 할까?”라는 질문에 임베딩 시스템만이 정답을 반환하였다. BM25는 인사라는 단어가 문장에 명시적으로 등장해 이를 우선적으로 매칭했을 가능성이 있으며, 동시에 문장 속 부정의 의미를 이해하지 못해 잘못된 단위과제를 후보로 제시했을 가능성이 있다. 반면 임베딩 시스템은 이를 노동조합 관련 공식 활동 기록이라는 개념으로 인식하고 인사관리 업무 영역과 구분함으로써 성공했을 것으로 추측된다.

반면, 아래와 같은 질문은 임베딩 시스템에서도 정답을 내놓지 못했다. “IRB와 관련해서 연구윤리 관련 연구계획서를 검토하는 업무를 담당하는 부서는 어디인가?”라는 질문에서 오답을 제시했는데, 이는 임베딩 시스템이 ‘IRB’가 조직 명칭인 ‘기관생명윤리위원회’와 대응된다는 점을 식별하지 못해, 더 광범위한 수준의 조직을 추천하는 오류를 범한 것으로 추정해 볼 수 있을 것이다.

4.2.4 오류전제형

오류전제형 질문에서는 BM25에서 정답이 나타난 사례는 13건(86.7%), 임베딩에서 정답이 나타난 사례는 7건(46.7%)으로 BM25가 우세한 경향이 있는 것으로 나타났다. 가령 “고등교육통계 조사 보존기한이 30년 맞지?”라는 질문에서 BM25는 핵심 명사 키워드를 기준표에서 정확하게 검색하였고, 그 결과 실제 보존기간이 포함된 단위과제가 LLM에 전달됨으로써 질문에 포함된 오류 전제를 교정할 수 있었던 것으로 보인다. 그러나 임베딩 시스템은 질문에

〈표 4〉 질의 유형에 따른 BM25와 임베딩 시스템의 검색 정확률

질의유형	BM25 정답	임베딩 정답
직답형 (15문항)	13 (86.7%)	11 (73.3%)
용어변형형 (15문항)	4 (26.7%)	7 (46.7%)
추론형 (15문항)	8 (53.3%)	10 (66.7%)
오류전제형 (15문항)	13 (86.7%)	7 (46.7%)

등장하는 ‘고등교육통계조사’를 정확한 단위과제로 식별하지 못하고 의미적으로 유사한 다른 단위 과제를 검색하여 LLM에 전달함으로써, 실제 보존기간에 대한 근거 정보를 제공하지 못한 것으로 보인다. 그 결과 질문에 포함된 오류전제를 교정하지 못한 것으로 추정된다.

4.3 기준표와 규정 간 검색 방식의 정확률 비교

여기에서는 행정적 속성값으로 구성된 기록관리기준표와 서술형으로 구성된 규정을 대상으로 한 질의에 있어 검색 방식에 따라 어떠한 성능 차이를 보이는지 평가한 결과를 〈표 5〉와 같이 살펴본다.

먼저 기록관리기준표를 대상으로 한 질의를 살펴보면 BM25는 24(60%)건에서 정답을 제시했으며, 임베딩은 21(52.5%)건에서만 정답이 나와 BM25가 약간 더 우세한 것으로 나타났다. 그 이유는 기록관리기준표가 단위과제, 담당부서, 보존기간, 기록물철과 같이 독립적 의미 단위로 구조화되어 있어 정확한 키워드 대응이 유리한 BM25 방식에서 효과를 발휘했기

때문이다. 반면 임베딩 방식은 유사한 행이 기준표에 흩어져 있을 경우, 의미적으로 유사한 다른 행을 선택할 가능성이 높아지므로 상대적으로 정확성이 떨어질 수 있다. 반면 규정의 특정 조항을 정답으로 고정한 질의에서는 BM25와 임베딩에서 정답이 나온 경우가 모두 동일하게 14건(70%)으로 나타나 두 방식 간 뚜렷한 우위는 확인되지 않았다. 가령 “업무가 끝난 기록을 기록관으로 넘기는 절차는 언제까지 해야 하나요?”와 같은 자연어 기반 질의의 용어 매칭에서 ‘이관’, ‘기한’이라는 용어로 변환하지 못한 BM25 방식은 검색에 실패하였으나, 의미적 동의어를 인지하여 문장 전체의 맥락을 보고 의도를 파악한 임베딩 방식은 성공하였다. 반면 “비공개 기록물도 직원이면 누구나 열람할 수 있죠?”라는 질문에 있어 BM25는 핵심 용어가 포함된 관련 규정 조항을 정확히 검색해 열람 제한 조건이 LLM의 답변 생성 과정에 활용될 수 있었다. 반면, 임베딩 기반 시스템은 실패했는데, 의미적으로 유사한 문서를 검색하는 과정에서 비공개 기록물의 열람 제한 조건이 명시된 핵심 조항을 검색하지 못한 것으로 보인다.

〈표 5〉 지식기반 구조에 따른 BM25와 임베딩 시스템의 검색 정확률

구조	BM25 정답	임베딩 정답
기준표 (40문항)	24 (60.0%)	21 (52.5%)
규정 (20문항)	14 (70.0%)	14 (70.0%)

4.4 BM25와 임베딩의 상대적 강·약점 종합

직답형 질문에서는 BM25(86.7%)와 임베딩(73.3%) 검색이 모두 비교적 안정적으로 작동하였으나, BM25에서 약간 더 우세한 성능이 나타났다. 오류전제형 질의에서도 명시적 불일치를 감지하는 데 있어 BM25(86.7%)가 임베딩(46.7%)에 비해 상대적으로 안정적인 성능을 나타냈다. 반면에 추론형 질의에서는 임베딩(66.7%)이 BM25(53.3%)에 비해 우세한 성능을 보였다. 그리고 용어변형형 질의에서도 임베딩(46.7%)이 BM25(26.7%) 보다 우세한 성능을 보였으나, 절반 이상의 질의에서 답변에 실패해 안정적 성능을 확보했다고 평가하기 어렵다. 한편, 서술형 조문으로 구성된 규정을 대상으로 하는 질의에서는 양 시스템이 각각 70%의 동일한 성능을 나타냈으나, 구조화된 속성으로 구성된 기준표를 대상으로 하는 질의에서는 BM25가 60%로 52.5%인 임베딩 시스템에 비해 높은 성능을 나타냈다.

양 시스템 검색 성능에 대한 비교·평가 결과를 정리하자면, 임베딩 검색이 용어변형형과 추론형 질문에 있어 BM25에 비해 강점이 있지만, 오류를 전제로 한 질문과 직답형 질문에 있어서는 상대적으로 낮은 성능을 보였다. 반대로 BM25는 직답형과 오류전제형 질문에서 우세한 성능을 보였으나 용어변형형과 추론형 질의에 있어서는 한계를 보였다. 한편, 문장이 길고 서술형 구조를 가진 규정 조문에서는 두 시스템이 비슷한 성능을 보였으나 기록관리기준표와 같이 단위과제명, 철명, 부서명 등 정확한 명칭 일치가 핵심인 정형 지식 구조에서는 BM25 검색이 상대적으로 유리하게 나타났다.

이러한 차이는 두 모델이 상호 보완적인 특성이 있음을 시사한다. 민법 조문을 지식기반으로 한 신은성 외(2025)의 연구에서도 질의 유형에 따라 요구되는 정보의 형태와 검색 전략이 달라지므로 확일적 처리만으로는 최적의 성능을 달성하기 어렵다고 지적하였다. 따라서 기록관리 규범 질의응답 시스템에서도 우세한 한 가지의 검색 방식을 선택하기보다, 질문의 특성에 따라 질의를 분기시키거나, 두 방식을 결합해 가중치를 조정하는 하이브리드 검색 모델이 적합할 수 있을 것이다. 한편, 용어변형형과 추론형에서 임베딩 검색이 BM25에 비해 성공한 사례가 많이 나타났지만, 일관된 성능을 보이지 못했다는 점도 주목해 볼 필요가 있다. 이는 의미 기반 검색이 BM25 검색에 비해 이용자의 다양한 표현을 수용하는 데 강점이 있지만, 특수 도메인의 전문 용어나 복잡한 문서 구조를 이해하지 못하고 환각을 발생시킬 가능성이 있음을 시사한다(오재광, 김수균, 2025). 따라서 임베딩 방식을 적용하더라도 조직기능을 반영한 온톨로지나 기록관리 어휘 사전 등을 결합함으로써 추가적인 검색 성능 향상을 도모할 필요가 있을 것이다.

5. 논의 및 결론

최근 인공지능 기술이 여러 분야에서 지식을 찾고 질문에 답하기 위해 사용되고 있으며, 여기에 응용되고 있는 RAG 구조는 신뢰성 있는 응답을 지원할 수 있는 핵심 기술로 주목되고 있다. RAG 구조는 지식기반 검색 단계의 성능에 따라 답변 품질이 크게 달라질 수 있기 때문

에, 지식기반의 특성과 질문 유형에 맞는 검색 방식이 선택되어야 한다. 일반적인 응용 분야에서 활용되고 있는 FAQ, 업무 매뉴얼 등과 같이 자연어 중심의 서술형 구조를 가진 지식기반이 활용된다면, 다양한 표현을 허용하는 임베딩 기반 방식이 검색 성능을 효과적으로 향상시킬 수 있다. 그러나 구조화된 형식과 속성, 고유 용어를 주로 사용하는 전문 분야의 기준과 규범을 지식기반으로 하는 경우에는 이용자가 던지는 자연어 질의와 지식기반 사이에 표현 차이가 상시 발생할 수 있다. 또한 의미가 비슷해도 이용자가 요구하는 정확한 정보 단위가 아니라면 오답이 될 가능성이 있다. 따라서 임베딩 방식이 일관되게 좋은 성능을 보일 수 있을 것이라고 단언하기는 어렵다.

기록관리기준표와 규정을 대상으로 이용자의 다양한 질의 유형에 따라 BM25와 임베딩 기반 검색의 성능이 어떻게 달라질 수 있는지 비교·평가한 본 연구에서는 다음과 같은 주요 사실을 발견하였다. 먼저, McNemar 검정 결과 두 시스템 간에 유의한 성능 차이가 나타나지 않았다. 그러나 검색 정확률을 비교한 결과, 질문 유형에 따라 양 시스템의 성능이 달라짐을 확인하였다. 직답형 및 전제오류가 있는 질의에서는 키워드 기반 검색(BM25)이 더 안정적인 결과를 나타냈다. 반대로 용어변형형과 추론형 질의에서는 임베딩 기반 검색이 상대적으로 우세한 성능을 보였다. 또한 서술형 구조를 가지는 규정 조문에서 답을 찾는 질의에서는 특정 방식이 우세하게 나타나지 않았으나 구조화된 속성 중심으로 구성된 기록관리기준표를 통해 답을 찾는 질의에 있어서는 BM25가 우세하게

나타났다. 이는 기록관리 규범 기반의 RAG 질의응답 시스템에서는 지식기반의 구조와 질의의 특성을 함께 고려하여 검색 전략을 설계할 필요성이 있음을 시사한다. 검색 방식 자체의 우열보다는 질문의 성격과 지식기반의 표현 방식에 따라 검색 방식을 동적으로 연결하거나 결합하는 방향으로 고도화할 필요성이 있다는 것이다. 본 연구 결과를 바탕으로 향후에는 이용자의 질문이 입력되면 질문의 유형을 자동으로 구분하여 적합한 검색 방식으로 분기시키거나 BM25와 임베딩 검색을 결합한 하이브리드 방식을 적용하여 가중치를 조절하는 전략을 고려할 필요가 있을 것이다. 또한 특수 도메인의 전문 용어를 학습하지 못한 임베딩의 한계를 보완하기 위하여 어휘 사전이나 온톨로지를 결합해 성능을 제고할 필요성도 있을 것이다.

한편, 본 연구는 검색 단계의 성능 비교에 초점을 둔 연구이기 때문에 생성 모델이나 프롬프트 구성 방식 등이 응답 성능에 미치는 영향은 다루지 못하였음을 한계로 밝힌다. 또한 단일 임베딩 모델을 적용한 실험 결과이므로 다양한 환경에서 일반화시키기 위해서는 다양한 임베딩 모델을 사용한 추가적인 검증도 필요하겠다. 이에 후속연구에서는 질의 유형에 따라 BM25와 임베딩 기반 검색으로 분기시키거나 가중치를 조절하는 하이브리드 검색 알고리즘의 개발뿐 아니라, 용어 사전을 활용한 질의 확장, 다양한 청킹 전략, 그리고 서로 다른 임베딩과 생성 모델을 적용한 비교 연구를 수행함으로써 기록관리 분야에 적합한 RAG 기반 질의응답 시스템의 고도화를 도모할 필요가 있을 것이다.

참 고 문 헌

- 강한훈, 정종균, 장진희 (2024). 임베딩 모델을 통한 문장의 의미 및 구조적 유사성 비교 평가. *한국IT정책경영학회 논문지*, 16(4), 3701-3709.
- 공공기록물 관리에 관한 법률 시행령. 대통령령 제35506호.
- 공공기록물 관리에 관한 법률. 법률 제21066호.
- 국가기록원 (2026a). 2026년 기록물관리 지침. (발간등록번호 11-1741050-100025-10).
- 국가기록원 (2026b). 2026년 대학 기록물관리 지침. (발간등록번호 11-1741050-100028-10).
- 김선욱 (2025). 대학기록관 챗봇 서비스 개발을 위한 RAG 적용: K 대학교를 중심으로. 석사학위논문, 경북대학교.
- 서주은, 이성신 (2021). 대학 행정기록물 관리현황과 개선방안: K 대학 사례를 중심으로. *한국비블리아학회지*, 32(1), 163-185. <http://dx.doi.org/10.14699/kbiblia.2021.32.1.171>
- 신은성, 소훈, 윤영준, 정찬우, 김동연, 정현준 (2025). 계층적 민법 구조를 반영한 모듈형 RAG 기반 한국어 법률 질의응답 시스템. *한국정보IT학회논문지*, 23(11), 27-38. <http://dx.doi.org/10.14801/jkiit.2025.23.11.27>
- 오재광, 김수균 (2025). 대학 학사 행정 보조를 위한 RAG 기반 질의응답 시스템 설계. *한국컴퓨터정보학회논문지*, 30(7), 79-86. <https://doi.org/10.9708/jksci.2025.30.07.079>
- 유경도, 임희정 (2025). AI 헬스케어 질의응답 데이터 기반 RAG 챗봇 시스템. *한국경영과학회 2025년 춘계 공동학술대회 논문집*, 3702-3706.
- 이동현, 최혁준, 강주영, 이한솔 (2025). RAG를 활용한 금융이해력 향상을 위한 질의응답 시스템 구축에 관한 연구. *경영정보학연구*, 27(3), 313-333. <http://dx.doi.org/10.14329/isr.2025.27.3.313>
- 이현섭 (2025). RAG 기반 지능형 대학정보 질의응답 시스템 설계. *사물인터넷융복합논문지*, 11(3), 161-166. <https://doi.org/10.20465/KIOTS.2025.11.3.161>
- Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5, 133-143.
- Gan, Junghyun (n.d.). *jhgan/ko-sroberta-multitask*. Hugging Face. Available: <https://huggingface.co/jhgan/ko-sroberta-multitask>
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: a survey. *arXiv*. <https://doi.org/10.48550/arXiv.2312.10997>
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

- Martineau, K. (2023, August 22). What is retrieval-augmented generation? IBM Research Blog.
Available: <https://research.ibm.com/blog/retrieval-augmented-generation-RAG>
- Taylor, R. S. (2015). Question-negotiation and information seeking in libraries. *College & Research Libraries*, 76(3), 251-267. <http://dx.doi.org/10.5860/crl.76.3.251>

• 국문 참고자료의 영어 표기

(English translation / romanization of references originally written in Korean)

- Enforcement Decree of the Public Records Management Act. Presidential Decree No. 35506.
- Kang, Hanhoon, Jung, Jong-kyun, & Jang, Jin-hee (2024). Evaluation of semantic and structural similarity of sentences using embedding models. *Journal of The Korea Society of Information Technology Policy & Management*, 16(4), 3701-3709.
- Kim, Sun-Wook (2025). Application of RAG for Developing a Chatbot Service in University Archives: focused on K University. Master's thesis, Kyungpook National University.
- Lee, Dong-hyun, Choi, Hyuk-jun, Kang, Ju-young, & Lee, Han-sol (2025). A study on the construction of a RAG-based question answering system for improving financial literacy. *Information Systems Review*, 27(3), 313-333. <http://dx.doi.org/10.14329/isr.2025.27.3.313>
- Lee, Hyun-seop (2025). Design of an intelligent university information question answering system based on RAG. *Journal of Internet of Things and Convergence*, 11(3), 161-166.
<https://doi.org/10.20465/KIOTS.2025.11.3.161>
- National Archives of Korea (2026a). Guidelines for Records Management. (Publication Registration No. 11-1741050-100025-10).
- National Archives of Korea (2026b). Guidelines for University Records Management. (Publication Registration No. 11-1741050-100028-10).
- Oh, Jae-gwang & Kim, Su-gyun (2025). Design of a RAG-based question answering system for university academic administration assistance. *Journal of The Korea Society of Computer and Information*, 30(7), 79-86. <https://doi.org/10.9708/jksoci.2025.30.07.079>
- Public Records Management Act. Act No. 21066.
- Seo, Ju-eun & Lee, Seong-sin (2021). Current status and improvement plans for university administrative records management: focusing on the case of K University. *Journal of the Korean Biblia Society for Library and Information Science*, 32(1), 163-185.
<http://dx.doi.org/10.14699/kbiblia.2021.32.1.171>
- Shin, Eun-seong, So, Hoon, Yoon, Young-jun, Jeong, Chan-woo, Kim, Dong-yeon, & Jeong,

Hyun-jun (2025). A modular RAG-based Korean legal question answering system reflecting the hierarchical structure of civil law. *Journal of Korean Institute of Information Technology*, 23(11), 27-38. <http://dx.doi.org/10.14801/jkiit.2025.23.11.27>

Yoo, Gyungdo & Lim, Hoi-Jeong (2025). AI healthcare question-answering data-based RAG chatbot system, *Proceedings of the 2025 Spring Joint Conference of the Korean Operations Research and Management Science Society*, 3702-3706.

