‘사서에게 물어보세요’ 서비스와 생성형 AI를 연계한 하이브리드 디지털 참고서비스 고도화 연구*

Advancing Digital Reference Services through a Hybrid RAG-Based Model: Integrating the “Ask a Librarian” Knowledge Base with Generative AI

임 정 훈 (Jeonghoon Lim)**
김 선 태 (Suntae Kim)***

초 록

본 연구는 국립중앙도서관 ‘사서에게 물어보세요’ 서비스에 축적된 지식정보 DB를 활용하여 키워드 기반 검색과 벡터 기반 검색을 결합한 하이브리드 검색 시스템의 성능을 실험적으로 분석하는 것을 목적으로 한다. 연구데이터는 지식정보 DB의 질의-응답 데이터 5,898건으로 구성되며, 한국십진분류법(KDC) 10개 대분류 체계를 포함한다. Python 기반 실험 환경에서 키워드 검색은 도치색인 기반 전문검색 엔진(Whoosh)을, 벡터 검색은 문장 임베딩 기반 벡터 데이터베이스(ChromaDB)를 적용하여 검색 시스템을 구현하였다. 실험데이터는 10개 대분류별로 10개씩 총 100개를 구성하고, 제안 시스템과 ‘사서에게 물어보세요’ 서비스 검색을 실제 호출하여 상위 10건 결과 및 응답시간을 수집하였다. 비교 결과, 제안 시스템은 평균 응답시간 0.21초, 검색 성공률 100%로 안정적인 검색 성능을 보인 반면, ‘사서에게 물어보세요’ 서비스는 평균 응답시간 13.12초, 검색 성공률 81%로 나타났다. 본 연구는 하이브리드 검색과 RAG 결합 방식이 검색 성공률과 결과 산출의 안정성 측면에서 기존 접근 방식에 비해 효과적임을 실험적으로 확인하였으며, 향후 연구 방향으로 한국어 특화 임베딩 모델의 적용과 적합도 기반 평가 체계의 확장을 제안하였다.

ABSTRACT

This study aims to experimentally analyze the performance of a hybrid search system that combines keyword-based and vector-based retrieval, utilizing the Knowledge Information Database accumulated through the National Library of Korea’s “Ask a Librarian” collaborative digital reference service. The dataset consists of 5,898 question-answer records from the Knowledge Information Database, categorized according to the ten main classes of the Korean Decimal Classification (KDC). The search system was implemented in a Python-based experimental environment, employing an inverted-index-based full-text search engine (Whoosh) for keyword retrieval and a sentence-embedding-based vector database (ChromaDB) for vector retrieval. A total of 100 test queries were constructed, with 10 queries for each of the 10 main classes, and both the proposed system and the “Ask a Librarian” service were invoked under identical conditions to collect the top 10 results and response times. The results showed that the proposed system achieved a mean response time of 0.21 seconds and a 100% search success rate, demonstrating stable retrieval performance, whereas the “Ask a Librarian” service recorded a mean response time of 13.12 seconds and an 81% search success rate. This study experimentally confirmed that the hybrid search and Retrieval-Augmented Generation (RAG) approach is more effective than the existing method in terms of search success rate and retrieval stability, and suggests future research directions including the application of Korean-language-specific embedding models and the expansion of relevance-based evaluation frameworks.

키워드: 디지털 참고서비스, 사서에게 물어보세요, 하이브리드 검색, 벡터 검색, 검색증강생성
Digital Reference Service, Ask a Librarian, Hybrid Search, Vector Search, RAG

---

* 이 논문은 2024년도 전북대학교 연구기반 조성비 지원에 의하여 연구되었음.
** 계명대학교 문헌정보학과 조교수(mictoxic@kmu.ac.kr / ISNI 0000 0004 8339 2694) (제1저자)
*** 전북대학교 문헌정보학과 교수, 연구데이터융복합연구소장
(kim.suntae@jbnu.ac.kr / ISNI 0000 0004 6492 6355) (교신저자)
논문접수일자 : 2026년 2월 19일 논문심사일자 : 2026년 2월 26일 게재확정일자 : 2026년 3월 18일
한국비블리아학회지, 37(1): 287-310, 2026. http://dx.doi.org/10.14699/kbiblia.2026.37.1.287

1. 서 론

디지털 정보 환경의 발전에 따라 이용자의 정보 요구는 점차 다양해지고 고도화되고 있다. 오늘날 이용자들은 단순한 결과 목록이 아니라, 자신의 질문 맥락을 반영한 설명형 응답을 기대하는 경향이 있다. 이러한 경향은 복잡한 정보탐색 과제에서 인간과 자동 에이전트를 비교한 대화형 검색 실험을 통해서도 확인된 바 있다(Vtyurina et al., 2017). 그러나 웹 기반 검색 환경에서 이용자의 질의는 문장형, 구어형, 단서형 등 다양한 양상으로 나타나며, 특히 동일한 의미가 서로 다른 표현으로 제시되는 경우가 많다. 이러한 상황에서 키워드 일치 중심의 검색은 질의 표현의 다양성을 충분히 반영하지 못해 검색 누락 또는 결과 부족을 초래할 수 있다. 또한 결과가 목록 중심으로 제시될 경우, 이용자는 검색 결과의 의미를 파악하고 후속 탐색을 설계하는 데 상당한 인지적 부담을 겪게 된다(Borgman, 1996).

이러한 상황에서 국립중앙도서관을 비롯한 전국의 공공도서관은 협력형 디지털 참고서비스(Collaborative Digital Reference Service, 이하 CDRS)인 '사서에게 물어보세요'를 통해 전문 사서가 검증한 신뢰도 높은 정보를 제공함으로써 이용자의 알 권리를 충족해 왔다(장혜란, 이경숙, 2014). 이 서비스는 전문 사서가 직접 단행본, 학술논문, 기사 등 신뢰도 높은 정보원을 선별하여 제시한다는 점에서 높은 공신력을 지닌다. 또한 이 서비스의 답변 결과는 '지식정보 DB'로 체계적으로 축적되고 있다. 그러나 질문 접수부터 이첩, 답변 작성 및 검토, 제공에 이르는 다단계 절차로 인해 응답 지연 문제가 지속적으로 지적되고 있다(김왕종, 이제환, 2016; 임정훈, 2025). 이와 같은 응답 지연은 학습, 연구, 업무 등 신속한 정보 확보가 필요한 상황에서 이용자의 정보탐색 활동을 제약하는 요인으로 작용할 수 있다.

최근 ChatGPT와 같은 대규모 언어 모델(Large Language Model, 이하 LLM) 기반 생성형 인공지능 서비스는 이용자의 질문에 즉시 응답하고, 자연어로 설명과 맥락을 전달함으로써 높은 편의성을 제공하고 있다. 특히 이러한 자연어 생성 기능은 이용자의 정보 이해를 촉진하는 데 기여할 수 있다(Kasneci et al., 2023). 그러나 생성형 인공지능 서비스는 '환각'(hallucination) 문제를 내포하고 있어, 정확성과 신뢰성이 핵심 가치인 도서관 참고서비스에 단독으로 적용하는 데에는 한계가 있다(Alkaissi & McFarlane, 2023; Ji et al., 2023).

이러한 문제를 해결하기 위해 전문 사서가 검증하여 축적한 '사서에게 물어보세요'의 지식정보 DB를 LLM 기술과 결합한다면, 기존 CDRS가 지닌 응답 지연과 설명 부족의 한계를 보완하면서도 LLM의 환각 문제를 통제할 수 있는 새로운 참고서비스 모델을 구축할 수 있을 것이다. 특히 최근 주목받고 있는 의미 기반 벡터 검색은 문장 수준의 의미 유사도를 활용하여 질의어와 문서 간의 어휘 불일치(lexical mismatch) 문제를 완화할 수 있다. 그러나 벡터 검색은 정확한 용어 일치가 중요한 상황에서 관련 자료를 누락하거나, 실제 관련성이 낮은 결과를 반환하는 문제가 발생할 수 있다(Thakur et al., 2021). 따라서 도서관 서비스 맥락에서는 키워드 검색의 정밀성과 벡터 검색의 유연성을 결합한 하이브리드 검색이 보다 실용적인 대안이 될 수 있다.

나아가 검색 증강 생성(Retrieval-Augmented Generation, 이하 RAG) 구조는 개별 검색 결과의 선정 근거와 맥락을 설명함으로써 이용자의 결과 해석 부담을 낮추고 참고서비스의 안내 품질을 실질적으로 향상시킬 수 있다(Lewis et al., 2020). 즉, 참고자료 자체는 지식정보 DB가 제공하는 검증된 정보원으로 한정하고, LLM은 해당 자료의 선택 근거, 활용 맥락, 관련 개념과 배경지식을 자연어로 설명하는 ‘설명 레이어’로 작동하도록 설계함으로써 신뢰성, 신속성, 설명의 충실성을 동시에 갖춘 새로운 참고서비스 모델의 토대를 마련할 수 있을 것이다.

본 연구의 목적은 국립중앙도서관 ‘사서에게 물어보세요’ 서비스에 축적된 지식정보 DB와 LLM의 자연어 생성 능력을 결합하여, 신속한 응답과 상세한 설명을 제공하는 디지털 참고서비스 모델을 설계하는 데 있다. 이를 위해 지식정보 DB를 재구조화하고, 분류체계(KDC), 키워드, 텍스트 임베딩을 결합한 하이브리드 검색과 RAG 구조를 적용하여 시스템을 구현하였다. 또한 제안 시스템의 성능 검증을 위해 동일한 지식정보 DB를 활용하는 기존 서비스인 ‘사서에게 물어보세요’와의 비교 실험을 수행하고, 이를 통해 하이브리드 검색과 RAG를 결합한 방식이 기존 키워드 매칭 방식 대비 서비스 제공 특성에서 어떤 차이를 보이는지를 실증적으로 분석하였다.

2. 이론적 배경

2.1 디지털 참고서비스와 CDRS

참고서비스(Reference Services)는 이용자의 정보요구에 따라 사서가 적절한 정보를 제공하거나 정보를 찾는 방법을 안내하는 일련의 활동을 의미한다. 전통적으로 참고서비스는 도서관 내에서 사서와 이용자가 대면하는 방식으로 이루어져 왔으나, 정보통신기술의 발달과 함께 이메일, 게시판, 채팅, 웹 폼 등을 활용한 디지털 참고서비스(Digital Reference Services, 이하 DRS)로 확장되었다. 이러한 변화는 이용자가 도서관을 방문하지 않고도 전문 사서의 도움을 받을 수 있도록 함으로써, 시간적·공간적 제약을 완화하고 정보 접근성을 향상시켰다(Lankes, 2004).

CDRS는 여러 도서관이 협력하여 온라인 기반으로 참고질문에 응답하는 서비스 형태를 의미한다. 국립중앙도서관이 운영하는 ‘사서에게 물어보세요’는 국내를 대표하는 CDRS로, 전국의 공공도서관 사서들이 참여하여 이용자 질문을 분담하고 전문적 답변을 제공한다. 이러한 구조는 다양한 지역과 주제 분야의 사서가 축적한 전문성과 경험을 공동으로 활용할 수 있으며, 생성된 답변이 데이터베이스로 축적됨에 따라 지속적인 지식 기반의 확장으로도 이어질 수 있다는 장점이 있다(장수현, 남영준, 2021).

그러나 CDRS는 질문 접수 및 이첩, 사서의 정보원 탐색, 답변 작성에 이르는 전 과정이 사서가 직접 수행하는 구조로 이루어져 있다. 이로 인해 실제 답변이 제공되기까지 상당한 시간이 소요되며, 이러한 응답 지연은 즉각적인 정보 획득을 기대하는 이용자의 서비스 활용을 제약하는 요인으로 작용할 수 있다. 또한 답변이 자료 목록 중심으로 제시되는 경우가 많아, 이용자가 제시된 정보의 의미와 활용 방법을 스스로 해석해야 하는 인지적 부담이 발생한다.

이는 참고서비스가 본래 수행해 온 안내와 설명의 기능을 충분히 구현하지 못하는 결과로 이어질 수 있다.

이처럼 CDRS는 신뢰성과 전문성이라는 강점에도 불구하고 응답의 신속성과 설명의 충실성 측면에서 한계가 있다. 다만, 이 서비스를 통해 전문 사서가 검증한 질의응답 결과물이 '지식정보 DB'로 체계적으로 축적되고 있다는 점은 주목할 만하다. 2026년 1월 3일 기준 5,898건이 저장되어 있으며, 단행본, 학위논문, 학술논문, 신문기사 등 신뢰도 높은 정보원이 포함되어 있다. 또한 한국십진분류법(KDC)을 활용하여 유강목 수준까지 분류되어 있어 체계적인 검색과 활용이 가능하다.

사서가 검증한 지식정보 DB를 LLM의 응답 생성 근거로 활용할 경우, LLM의 환각 문제를 완화하는 데 기여할 수 있다. 특히 지식정보 DB에는 다양한 질문 유형과 답변 방식을 포괄하는 데이터가 축적되어 있어, 자연어 기반 질문 분석, 주제어 추출, 의미 기반 검색 등의 AI 기술과 결합하기에 적합한 구조를 갖추고 있다. 따라서 지식정보 DB를 LLM 기반 참고서비스의 핵심 정보원으로 활용한다면, 기존 CDRS의 한계를 보완하는 새로운 디지털 참고서비스 모델을 구축할 수 있을 것이다.

2.2 RAG 및 벡터 검색

ChatGPT와 같은 생성형 AI는 대규모 언어 모델(LLM)을 기반으로 하여 인간과 유사한 자연어 응답을 생성하고, 정보의 요약, 설명, 재구성 등을 수행할 수 있다(Kasneci et al., 2023). 그러나 LLM은 확률적 모델에 기반하기 때문에 사실이 아닌 정보를 사실처럼 생성하는 환각이 발생할 수 있으며, 이는 정확한 출처와 사실 확인이 필수적인 참고서비스 영역에서 심각한 신뢰성 저하를 초래할 수 있다. 따라서 LLM을 참고서비스에 적용하기 위해서는 신뢰할 수 있는 외부 지식을 근거로 응답을 생성하도록 제어하는 구조가 필요하다.

이러한 요구에 대응하기 위해 제안된 방법이 RAG 구조이다. RAG는 이용자의 질의를 바탕으로 외부 문서 집합에서 관련 정보를 검색하고, 이를 활용하여 자연어 응답을 생성하는 방식이다. Gao et al.(2024)은 RAG를 LLM의 내재적 지식과 외부 지식 저장소를 결합하는 생성 패러다임으로 개념화하고, 외부 문서를 생성 과정에 통합함으로써 응답이 특정 정보원에 근거하도록 설계된 구조임을 설명하였다. Shuster et al.(2021)은 이러한 검색 결합 생성 방식이 대화형 언어 모델의 사실 오류와 환각을 유의미하게 감소시키며, 특히 지식 기반 대화에서 응답의 정확성과 신뢰성을 향상시킨다고 보고하였다. 본 연구에서 RAG를 적용할 경우, LLM은 '사서에게 물어보세요' 지식정보 DB에 근거하여 응답을 생성하게 되며, 이를 통해 이용자에게 정확하면서도 이해하기 쉬운 참고 응답을 제공할 수 있다.

RAG 기반 시스템에서 핵심이 되는 기술은 의미 기반 검색이다. 전통적인 키워드 검색은 용어의 정확한 일치를 기반으로 하기 때문에, 질의 표현이 다양하거나 간접적인 경우 적합한 자료를 검색하기 어렵다. 반면 벡터 검색(Vector Search)은 문장이나 문서를 임베딩 벡터로 변환하여 의미적 유사도를 계산함으로써, 표현이 다르더라도 의미적으로 유사한 정보를 검색할

수 있다(Karpukhin et al., 2020; Reimers & Gurevych, 2019). 예를 들어 '친구'라는 단어가 포함되지 않더라도 '우정'이나 '교우 관계'와 관련된 문서를 검색할 수 있다.

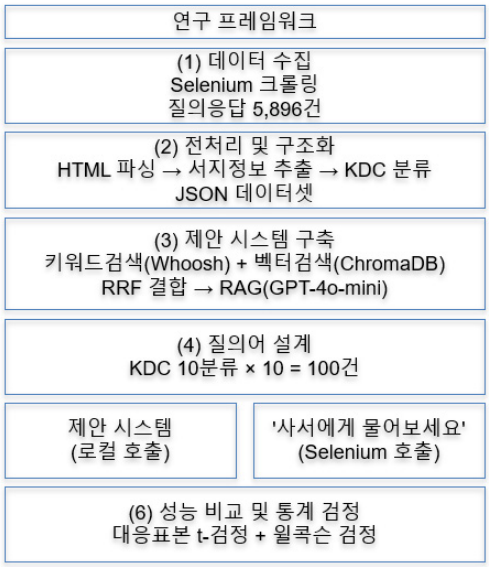
그러나 벡터 검색은 의미적 유사성에 기반하기 때문에, 도서명이나 고유명사와 같이 정확한 용어 일치가 중요한 상황에서는 부적절한 결과를 반환할 수 있다. 따라서 도서관 서비스 맥락에서는 키워드 검색의 정밀성과 벡터 검색의 유연성을 결합한 하이브리드 검색이 실용적인 대안이 될 수 있다. 하이브리드 검색은 두 방식의 결과를 결합함으로써 용어 일치와 의미적 연관성을 동시에 고려할 수 있으며, 이용자의 다양한 질문 유형에 효과적으로 대응할 수 있다.

따라서 RAG 구조와 하이브리드 검색의 결합은 검증된 정보원을 토대로 즉각적이고 상세한 응답 생성을 가능하게 하며, 기존 CDRS의 신뢰성과 전문성을 유지하면서 응답 지연과 설명 부족이라는 한계를 극복할 수 있는 기술적 기반을 제공한다.

3. 연구 방법

3.1 연구 설계

본 연구는 데이터 수집, 전처리 및 구조화, 제안 시스템 구축, 질의어 설계, 실험데이터 생성, 성능 지표 산출 및 통계 검정의 6단계 절차에 따라 수행하였다. 비교 대상은 제안 시스템과 국립중앙도서관 '사서에게 물어보세요' 서비스이며, 평가 지표는 검색 결과 건수, 검색 성공률, 응답시간으로 설정하였다. 비교 대상을 '사서에게 물어보세요' 서비스로 선정한 것은 해당 서비스가 제안 시스템과 동일한 지식정보 DB(5,896건)를 기반으로 운영되는 유일한 기존 서비스이기 때문이다. 이를 통해 동일 데이터 기반에서 키워드 매칭 방식과 하이브리드 및 RAG 방식 간 검색 성능 및 응답 특성의 차이를 실증적으로 분석하고자 하였다.



〈그림 1〉 전체 연구 프레임워크

〈그림 1〉은 본 연구의 전체 프레임워크를 나타낸다. 연구는 크게 여섯 단계로 구성된다. 첫째, Selenium 기반 웹 크롤링을 통해 '사서에게 물어보세요' 서비스의 질의응답 데이터 5,896건을 수집한다. 둘째, HTML 파싱, 서지정보 추출, KDC 분류를 거쳐 JSON 형식의 구조화된 데이터셋을 구축한다. 셋째, 키워드 검색 엔진(Whoosh, 도치색인), 벡터 검색 데이터베이스(ChromaDB, 문장 임베딩), 하이브리드 결

합(RRF), RAG 응답 생성(GPT-4o-mini) 모듈로 구성된 제안 시스템을 구축한다. 넷째, KDC 10개 대분류에서 각 10건씩 총 100건의 실험용 질의어를 설계한다. 다섯째, 동일한 질의어에 대해 제안 시스템과 '사서에게 물어보세요' 서비스를 각각 호출하여 검색 결과와 응답시간을 수집한다. 여섯째, 대응표본 t-검정과 윌콕슨 부호순위 검정을 통해 두 시스템의 성능 차이를 통계적으로 검증한다.

3.2 데이터 수집

본 연구의 데이터 수집 대상은 국립중앙도서관이 운영하는 '사서에게 물어보세요' 서비스에 게시된 참고질문과 답변 게시글이다. 수집 기준일은 2026년 1월 3일로 설정하였으며, 해당 시점까지 축적된 전체 게시글 5,898건을 대상으로 자동 수집을 수행하였다. 그 결과 총 5,896건의 게시글을 확보하였으며, 나머지 2건은 접근 제한 등의 사유로 수집되지 않았다.

〈그림 2〉는 '사서에게 물어보세요' 서비스의 참고질문 목록과 개별 질의응답 상세 화면을 나타낸 것이다. 이용자의 참고질문에 대해 전국 공공도서관 사서가 〈그림 3〉과 같이 답변을 작성하며, 각 답변에는 질문 요약, 답변 본문, 관련 서지정보, 답변 도서관, 답변일 등의 정보가 체계적으로 기록되어 있다. 본 연구는 이러한 질의응답 게시글의 전체 내용을 자동 수집하여 연구데이터로 활용하였다.

3.3 전처리 및 데이터셋 구성

전처리는 검색 인덱스의 정확성과 실험 평가의 일관성을 확보하기 위하여 다음과 같은 단계로 진행되었다. 첫째, 수집된 HTML 문서에서 답변 본문 영역을 중심으로 텍스트를 추출하고, 메뉴, 네비게이션, 광고 영역 등 분석에 불필요한 요소를 제거하였다. 둘째, 모든 텍스트 파일의 인코딩을 UTF-8로 통일하여, 검색 엔진 및 자연어 처리 과정에서 발생할 수 있는 문자 손

〈그림 2〉 '사서에게 물어보세요' 질문 목록 화면

〈그림 3〉 개별 질의응답 상세 화면 1건

상 문제를 방지하였다. 셋째, 답변 본문에 포함된 단행본, 학위논문, 학술기사, 보고서 등의 서지정보를 식별·분리하여 구조화된 항목으로 변환하였다. 넷째, 최종 데이터는 문서 식별자, 질문, 답변 본문, 서지정보, 답변 도서관, 답변일, KDC 분류기호 등의 필드를 포함하도록 JSON 형식으로 재구성하였다. 이와 같은 전처리 과정을 통해 원문 텍스트의 맥락을 보존하면서 검색 인덱스 구축과 시스템 연계를 용이하게 하는 구조화된 데이터셋을 구축하였다.

처리된 데이터의 주제 분포를 파악하기 위하여 KDC 대분류 기준으로 분석한 결과는 〈표 1〉과 같다. 데이터 분포는 '총류'와 '사회과학' 영역에 상대적으로 편중되어 있다. 이에 따라 본 연구에서는 실험용 질의어를 KDC 대분류별로 균등하게 배분하여, 특정 주제 영역에 편중된 데이터 특성이 실험 결과에 미치는 영향을 통제하였다. 데이터 분포에 비례하여 질의어를 추출할 경우, 종교(1.2%), 언어(1.5%) 등 소수 영역에서는 1~2건의 질의어만 배정되어 해당 영역의 검색 성능을 통계적으로 분석하기 어렵다. 반면 총류(43.7%) 등 다수 영역에 질의어가 편중되면 전체 성능이 특정 주제의 검색 특성에 과도하게 영향을 받을 수 있다. 따라서 본 연구에서는 층화 균등 배분(stratified equal allocation)을 채택하여 모든 대분류에서 동일한 수(10건)의 질의어로 성능을 비교함으로써 분류별 검색 성능의 차이를 균형 있게 식별할 수 있도록 하였다(〈표 1〉 참조).

3.4 실험 환경 구축

3.4.1 시스템 구성 요소

본 연구의 실험 환경은 웹 인터페이스, 검색 엔진(키워드 검색, 벡터 검색, 하이브리드 검색), 결과 융합 모듈, RAG 기반 응답 생성 모듈로 구성하였다. 이용자는 웹 기반 화면을 통해 질의어를 입력하고 필요에 따라 분류 필터를 설정하며, 시스템은 입력된 질의를 바탕으로 지식정보 DB를 검색하여 결과를 제시한다. 이때 시스템은 검색 결과 목록과 함께, 필요시 검색 근거를 바탕으로 생성된 설명형 응답을

〈표 1〉 KDC 대분류별 데이터 분포

대분류	건수	비율(%)
총류	2,575	43.7
사회과학	1,338	22.7
기술과학	561	9.5
역사	369	6.3
문학	272	4.6
예술	244	4.1
자연과학	187	3.2
철학	178	3.0
언어	89	1.5
종교	68	1.2
합계	5,896	100.0

제공하도록 설계하였다. 이를 통해 이용자는 동일한 인터페이스에서 검색된 자료와 그에 대한 설명을 함께 확인할 수 있다.

실험 환경의 기술적 구성은 〈표 2〉와 같다. 특히 국립중앙도서관 '사서에게 물어보세요' 서비스는 동적 렌더링 요소를 포함하고 있어, 정적 HTTP 요청이나 단순 HTML 파싱만으로는 검색 결과 수집이 불안정할 수 있다. 이에 본 연구에서는 Selenium(WebDriver)을 활용하여 실제 브라우저 환경에서 검색을 재현하고, 결과 수집의 안정성과 일관성을 확보하였다.

3.4.2 웹 플랫폼 구축

웹 플랫폼은 Streamlit 기반으로 구현하였다. Streamlit을 선택한 이유는 첫째, Python 기반의 검색·분석 코드와 사용자 인터페이스(UI)를 단일 런타임 환경에서 통합할 수 있다는 점, 둘째, 프로토타이핑 속도가 빠르고 수정 및 확장이 용이하다는 점, 셋째, 100개 질의에 대한 반복 실험 과정에서 동일한 인터페이스를 통해 결과를 확인하고 로그를 검증하는 데 효율적이라는 점이다. 이러한 특성은 제안 시스템의 설계, 실험, 검증을 하나의 환경에서 일관되게 수행하는 데 적합하다.

UI는 검색창, 검색 방식 선택(키워드, 벡터, 하이브리드), KDC 대분류 필터, 검색 결과 목록, 결과 상세 영역(질문, 답변 도서관, 서지정보)으로 구성하였다. 이를 통해 이용자는 자연어 질의를 입력하고, 분류체계를 기반으로 검색 범위를 조정하며, 반환된 결과를 구조적으로 확인할 수 있도록 설계하였다.

〈그림 4〉는 제안 시스템의 메인 검색 인터페이스를 보여준다. 사용자는 상단 검색창에 질의어를 입력하고, KDC 분류체계에 기반한 대분류-중분류-소분류의 3단계 필터를 통해 검색 범위를 제한할 수 있다. 검색 결과는 하이브리드 검색 방식으로 산출된 관련도 점수에 따라 정렬되며, 각 결과에는 원 질문, 답변 도서관, KDC 분류, 추천 자료의 서지정보가 함께 표시된다. 좌측 사이드바에서는 AI 답변 생성 기능과 국립중앙도서관 '사서에게 물어보세요' 서비스와의 실시간 비교 기능을 선택적으로 활성화할 수 있으며, 전체 데이터베이스의 규모(총 5,896건)와 분포를 확인할 수 있다.

3.4.3 키워드 검색 색인 구축

키워드 검색은 Whoosh 기반의 도치색인 구조로 구현하였다. 인덱스 스키마는 문서 식별

〈표 2〉 실험 환경 기술 스택 개요

구분	기술	용도
프로그래밍 언어	Python 3.9	데이터 수집·전처리·검색·평가 모듈 구현
웹 플랫폼	Streamlit	검색 UI 및 결과 확인
키워드 검색	Whoosh 2.7	도치색인 기반 전문 검색
벡터 검색	ChromaDB 0.4	임베딩 저장 및 유사도 검색
임베딩 모델	all-MiniLM-L6-v2	문장·문단 임베딩 생성
브라우저 자동화	Selenium(WebDriver)	국중 서비스 검색 결과 자동 수집
RAG 생성	GPT-4o-mini	근거 기반 설명형 응답 생성

〈그림 4〉 제안 시스템 메인화면 구성

자, 질문 텍스트, 답변 텍스트, 서지정보 텍스트, KDC 분류기호, 답변 도서관, 답변일 등의 필드로 구성하였다. 이 가운데 질문 텍스트와 서지정보는 이용자 질의와 직접적으로 연관되는 핵심 필드이므로, 다중 필드 검색이 가능하도록 설계하였다. 인덱스 구축 과정에서는 텍스트에 대해 기본적인 정규화(불필요한 기호 제거, 공백 정리)를 수행한 후 토큰화 결과를 도치색인에 저장하였다. 질의 처리 시에는 복수 필드에 대한 통합 질의를 수행하여 질문 표현과 자료정보가 동시에 반영되도록 하였다.

Whoosh를 검색 엔진으로 선택한 이유는 다음과 같다. 첫째, 별도의 서버 환경 없이 애플리케이션 내부에서 작동하므로 실험 환경의 재현성이 높다. 둘째, 도치색인 기반의 전통적인 전문검색 구조를 비교적 단순한 설정으로 구현할 수 있다. 셋째, 결과 순위 산출이 안정적이어서 벡터 검색과 결합되는 하이브리드 검색의 한 축으로 적합하다.

3.4.4 벡터 검색 구축(ChromaDB 및 임베딩)

벡터 검색은 ChromaDB 0.4.x를 활용하여 구현하였다. ChromaDB는 문장 임베딩 기반 벡터 검색을 위해 설계된 경량 벡터 데이터베이스로, Python 환경에서의 손쉬운 설치와 사용, 임베딩 모델 및 응용 프레임워크와의 높은 호환성을 특징으로 한다. 특히 컬렉션 단위의 벡터 관리, 메타데이터 저장 및 필터링, 코사인 유사도 기반 검색 기능을 기본적으로 제공하여, 비교적 단순한 실험 환경에서도 벡터 검색 시스템을 안정적으로 구현할 수 있다.

본 연구는 대규모 분산 시스템의 확장성이 아니라 하이브리드 검색 구조 자체의 효과를 분석하는 데 초점을 두고 있다. 실험데이터 규모(5,896건)와 단일 서버 기반의 프로토타입 환경을 고려할 때, FAISS, Milvus, Weaviate와 같은 대규모 분산형 벡터 데이터베이스를 적용하는 것은 시스템 복잡도를 불필요하게 증가시키고 검색 구조 자체의 효과 분석에 혼선을 줄 수 있다. 따라서 본 연구에서는 단일 서버

환경에서 안정적인 벡터 검색 기능을 제공하고, 실험 설계의 재현성과 구현의 일관성을 확보하기에 적합한 ChromaDB 0.4.x를 채택하였다.

문서 임베딩 모델로는 all-MiniLM-L6-v2를 사용하였다(Wang et al., 2020). 해당 모델은 Transformer 구조를 기반으로 문장 수준의 의미 정보를 고정 차원 벡터로 표현하도록 학습된 임베딩 모델로, 의미 유사도 검색, 정보 검색, 문서 군집화 등 검색 중심 응용 분야에서 널리 활용되고 있다(Reimers & Gurevych, 2019). all-MiniLM-L6-v2는 384차원의 비교적 낮은 벡터 차원과 경량화된 모델 구조를 갖추고 있어, 대규모 데이터셋이나 실시간 응답이 요구되는 검색 환경에서도 빠른 추론 속도와 안정적인 성능을 제공한다. 본 연구는 단일 서버 환경에서 키워드 검색과 하이브리드 검색의 성능과 응답 특성을 비교하므로 임베딩 모델의 계산 효율성과 처리 속도를 중요하게 고려하였다. 아울러 ChromaDB와의 호환성 및 풍부한 기술 문서와 시스템 구현의 재현성과 실험 설계의 일관성을 확보하는 데 유리한 점을 감안하여 이 모델을 채택하였다.

임베딩 단위는 문서 전체가 아니라 문단 단위로 설정하였다. 참고정보 답변은 서지정보와 설명이 혼합되어 상대적으로 길기 때문에, 문서 전체를 하나의 벡터로 변환할 경우 질의와 직접적으로 관련된 핵심 구간의 의미가 희석될 가능성이 있다. 이에 비해 문단 단위 임베딩은 질의와 직접적으로 연관된 의미를 보다 정확하게 포착할 수 있고, 검색 결과가 '문서 전체'가 아니라 '관련 문단 중심'으로 제공될 수 있어 RAG 기반 설명형 응답에서 근거 제시에 유리하다. 각 벡터 레코드에는 문단 텍스트와 함께 원문 문서 식별자, KDC 대분류, 답변 도서관 등의 메타데이터를 저장하여 필터링과 결과 추적이 가능하도록 하였다. 벡터 검색은 문단 단위로 유사도를 산출하되, 검색된 문단이 속한 원본 문서의 식별자를 기준으로 결과를 역추적하여 문서 단위로 최종 결과를 제시한다. 즉, 동일 문서에서 파생된 복수의 문단이 검색될 경우 해당 문단들의 유사도 점수 중 최고값을 해당 문서의 대표 점수로 사용한다. 이를 통해 벡터 검색의 결과 단위는 개별 문단이 아닌 원본 질의응답 문서(게시글)가 된다.

〈그림 5〉는 검색 결과 하단에 표시되는 관련

〈그림 5〉 검색 결과 하단에 표시되는 관련 질문 추천 기능

질문 추천 기능을 나타낸다. 시스템은 상위 첫 번째 검색 결과의 질문 텍스트를 기준으로, 의미적으로 유사한 질문을 벡터 유사도 검색을 통해 추천한다. 추천 결과에는 각 질문의 KDC 분류 정보와 유사도 점수(%)가 함께 제시되어 관련 주제로 탐색을 확장할 수 있도록 지원한다. 각 추천 질문 옆의 '검색' 버튼을 클릭하면 해당 질문을 질의어로 즉시 재검색된다. 이 기능은 초기 질의어만으로는 발견하기 어려운 관련 자료를 발굴하고, 탐색 범위를 점진적으로 확장하는 데 기여할 수 있다.

관련 질문 추천은 다음과 같은 조건과 절차에 따라 수행된다. 질문 임베딩에는 paraphrase-multilingual-mpnet-base-v2 모델을 사용하고, 유사도는 코사인 유사도로 계산한다. 검색된 후보 질문에 대해 KDC 분류 정보를 반영하여 가중치를 적용하는데, 동일한 대분류에 속할 경우 유사도에 15%, 중분류는 10%, 소분류는 5%의 가중치를 추가한다. 이후 기존 검색 결과에 포함된 문서는 제외하고, 최종적으로 상위 5개의 질문을 추천한다. 이용자가 추천 항목을 선택하면 해당 질문이 새로운 질의어로 자동 입력되어 즉시 재검색이 이루어진다. 이를 통해 의미적 유사성과 주제적 근접성을 동시에 고려한 탐색 확장이 가능하다.

3.4.5 하이브리드 검색 결과 융합

키워드 검색 결과와 벡터 검색 결과를 Reciprocal Rank Fusion(RRF) 방식으로 융합하였다(Cormack et al., 2009). RRF는 서로 다른 검색 시스템에서 산출된 결과의 순위 정보를 역수 형태로 변환하여 합산함으로써, 점수 스케일이 상이한 결과를 편향 없이 결합할 수 있는 기법이다. 이 방식은 특정 검색 엔진의 점수 체계에 의존하지 않고, 각 시스템에서 상위에 위치한 결과를 안정적으로 반영할 수 있다는 장점이 있다.

융합의 목적은 키워드 일치 기반 검색의 정밀성과 벡터 기반 의미 검색의 확장성을 동시에 확보하는 데 있다. 즉, 용어가 정확히 일치하는 문서를 우선적으로 반영하면서도 표현은 다르지만 의미적으로 유사한 질의-문서 쌍을 함께 포착하여, 자연어 질의에 보다 포괄적이고 유연하게 대응할 수 있도록 설계하였다. 최종 출력 결과는 상위 10건으로 제한하여, 제안 시스템과 '사서에게 물어보세요' 서비스 간 비교 실험에서 동일한 기준으로 검색 결과를 수집하고 성능을 평가할 수 있도록 하였다. 키워드 검색(Whoosh)은 문서 단위로, 벡터 검색(ChromaDB)은 문단 단위로 검색을 수행하지만, RRF 결합 시 두 검색 결과는 문서 식별자를 기준으로 통합된다. 따라서 최종 출력되는 상위 10건은 개별 문단이 아닌 문서(질의응답 게시글) 단위이며, 이는 비교 시스템인 '사서에게 물어보세요' 서비스의 검색 결과 단위와 동일하다.

3.4.6 RAG 응답 생성 모듈

RAG는 검색 결과의 서지정보와 핵심 문단을 컨텍스트로 활용하여 이용자에게 설명형 응답을 제공하도록 설계하였다. 생성형 모델은 GPT-4o-mini를 사용하였으며, 응답의 일관성과 재현성을 확보하기 위해 temperature 값을 낮게 설정하였다. RAG 응답은 검색 결과를 대체하는 기능이 아니라, 검색 결과의 의미를 해석하고 활용을 돕는 보조적 기능으로 정의하였다. 생성형 AI의 환각 위험을 최소화하기 위해

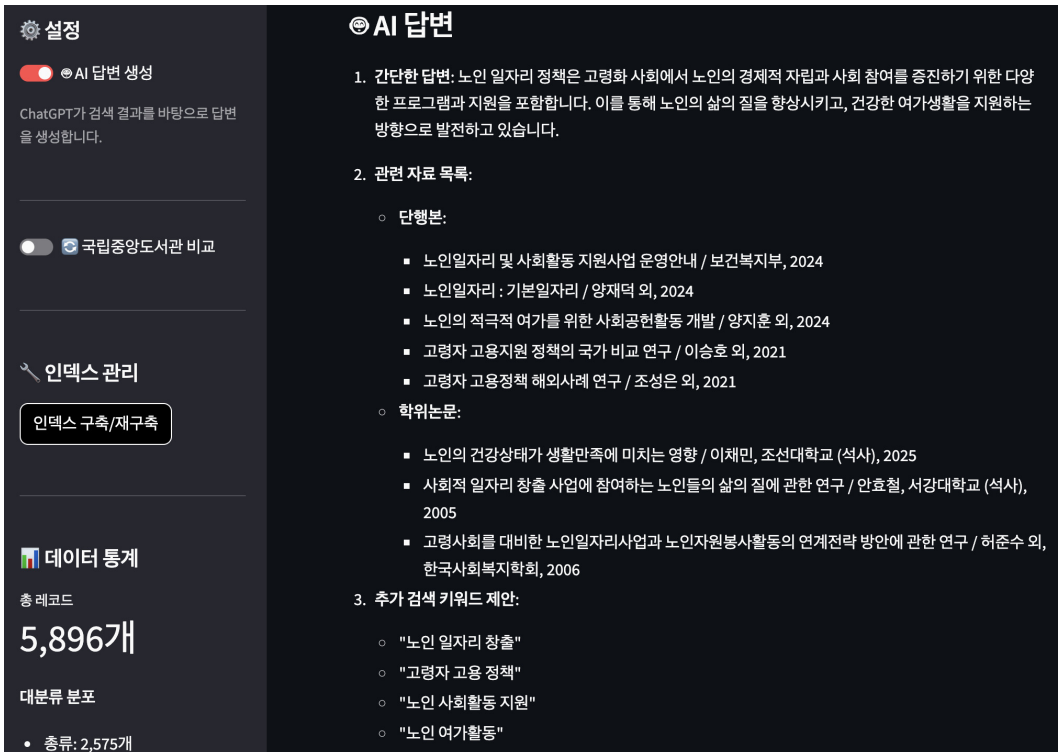
프롬프트에서 "컨텍스트로 제공된 정보만을 근거로 답변을 생성할 것"을 명시하였다.

〈그림 6〉은 RAG 기반 AI 답변 생성 기능의 작동 예시이다. 이용자가 '노인 일자리 정책'을 검색하면, 시스템은 먼저 하이브리드 검색을 통해 관련 문서를 검색한 후, 상위 5개 검색 결과의 서지정보를 컨텍스트로 활용하여 GPT-4o-mini 모델이 구조화된 답변을 생성한다. 생성된 답변은 세 단계로 구성된다. 첫째, 질문의 핵심 요지를 요약한 개요를 제시한다. 둘째, 단행본과 학위논문 등 자료 유형별로 구분된 관련 자료 목록을 제공한다. 셋째, 탐색 확장을 위한 추구 검색 키워드를 제안한다. 이러한 RAG 방식은 사서가 수행해 온 참고서비스의 설명 기능을 자동화된 형태로 보완하며, 이용자에게 즉각적이고 체계적인 정보를 제공할 수 있다.

3.5 실험데이터 생성

3.5.1 질의어 구성

실험에 사용될 질의어는 한국십진분류법(KDC) 대분류 10개 범주별로 각 10개씩 총 100개를 구성하였다. 이는 특정 주제 영역의 편중을 방지하고 두 시스템의 성능을 균형 있게 비교하기 위함이다. 질의어는 '사서에게 물어보세요' 지식정보 DB에 축적된 실제 질문 표현과 키워드를 참조하여 생성하였으며, 단일 키워드형 질의와 문장형(자연어) 질의를 혼합하였다.



〈그림 6〉 RAG 기반 AI 답변 생성

질의어 규모를 100건으로 설정한 것은 다음과 같은 이유에 근거한다. 첫째, KDC 10개 대분류에서 각 10건씩 균등하게 배분하는 층화 균등 배분 설계를 통해 분류별 성능 비교가 가능한 최소 규모를 확보하였다. 둘째, 비교 시스템에 대한 데이터 수집이 Selenium 기반 실제 웹 브라우저 호출 방식으로 이루어져 1건당 약 12~14초가 소요되며, 외부 서비스에 대한 과도한 호출은 서비스 안정성에 영향을 줄 수 있다. 셋째, 정보검색 분야의 대표적 평가 체계인 TREC(Text REtrieval Conference)에서도 평가용 질의 세트의 규모를 50~100건 수준으로 구성하는 것이 일반적이다(Voorhees & Harman, 2005). 본 연구에서는 100건의 질의에 대해 대응표본 t-검정과 윌콕슨 부호순위 검정 모두에서 $p<.001$ 수준의 강한 통계적 유의미성이 확인되어 표본 규모의 적절성이 검증되었다.

3.5.2 실제 호출 기반 검색결과 데이터 수집

본 연구의 실험데이터는 사전에 정답(골드셋)을 구축하는 방식이 아니라, 동일한 질의어에 대해 제안 시스템과 '사서에게 물어보세요' 서비스를 각각 실제로 호출하여 반환한 검색결과를 수집·기록하는 방식으로 생성하였다. 즉, 실험데이터는 두 시스템의 '실측 결과'를 포함하는 비교 로그로 구성되며, 각 질의에 대해 검색 결과 총 건수, 상위 10건의 검색 결과 목록, 응답시간, 오류 발생 여부 등의 정보가 포함된다.

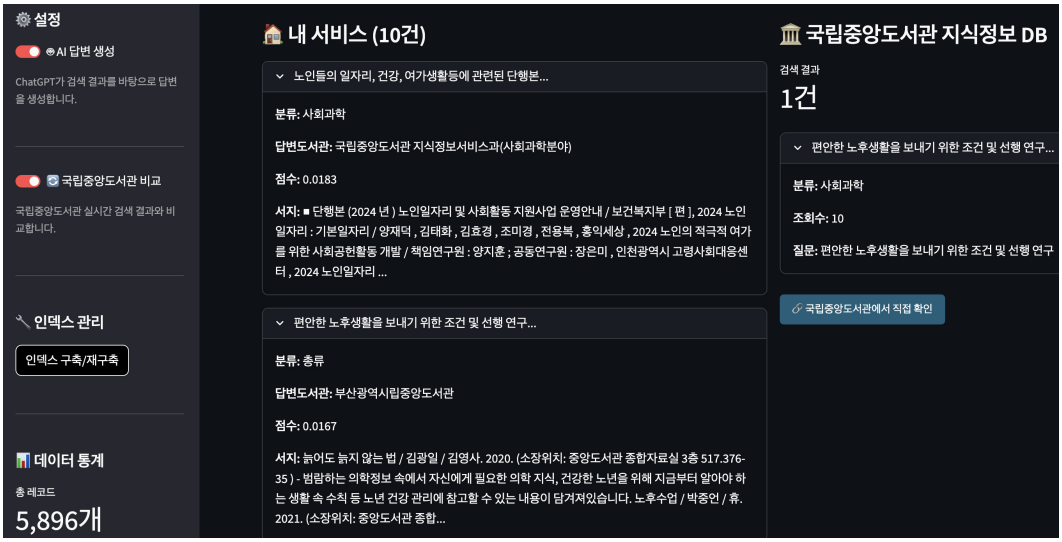
제안 시스템의 경우 모든 질의에 대해 결과 출력 상한을 10건으로 고정하였다. 이는 동일한 조건에서 결과를 수집하여 비교 가능성을 확보하고, 응답시간과 검색 성공률을 안정적으로 측정하기 위함이다.

3.5.3 '사서에게 물어보세요' 서비스 호출 자동화와 Selenium 사용

국립중앙도서관 '사서에게 물어보세요' 서비스의 검색 결과는 웹 인터페이스 기반으로 제공되며, 결과 목록이 자바스크립트 실행 이후에 완성되거나 페이지 로딩 상태에 따라 DOM 구조가 달라지는 특성을 지닌다. 이러한 환경에서는 requests와 같은 정적 HTTP 요청 방식만으로는 검색 결과를 안정적으로 수집하기 어렵다. 또한 본 연구는 100개의 질의어를 동일한 조건으로 반복 실행해야 하므로 자동화된 호출 방식이 필수적이다.

이에 본 연구는 Selenium(WebDriver)을 활용하여 실제 브라우저 환경에서 페이지를 로딩하고, 명시적 대기(Explicit Wait)를 통해 결과 영역이 완전히 렌더링된 이후 DOM에서 검색 결과 요소를 추출하는 방식을 채택하였다. 이를 통해 동적 로딩 환경에서의 안정적인 결과 수집, 반복 실행 시 파싱 일관성 확보, 대량 질의의 자동화된 처리와 로그 기록을 가능하게 하여 실험의 재현성을 확보하였다.

〈그림 7〉은 제안 시스템과 '사서에게 물어보세요' 서비스의 검색 결과를 실시간으로 비교한 예시이다. '편안한 노후생활을 보내기 위한 조건 및 선행 연구'라는 동일한 질의에 대해, 제안 시스템은 10건의 관련 결과를 반환한 반면, '사서에게 물어보세요' 서비스는 1건의 결과만을 반환하였다. 제안 시스템의 결과는 총류, 사회과학 등 다양한 KDC 분류에 걸쳐 분포하며, 각 결과에 대해 도서명, 저자, 출판연도, 소장위치 등 상세한 서지정보를 제공한다. 이러한



〈그림 7〉 제안 시스템과 '사서에게 물어보세요' 지식정보 DB의 검색 결과

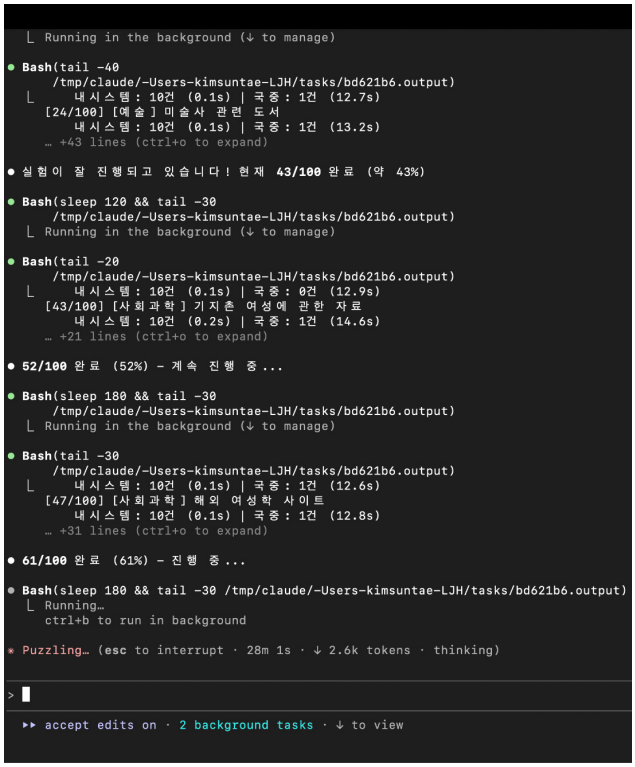
비교 기능은 제안 시스템의 검색 성능을 객관적으로 검증하고 기존 서비스 대비 개선점을 시각적으로 확인할 수 있다.

〈그림 8〉은 검색 시스템 비교 평가 실험의 자동화 스크립트 실행 화면이다. Python 기반의 평가 스크립트는 KDC 10개 대분류에서 각 10개씩 총 100개의 질의어를 무작위로 추출하고, 각 질의어에 대해 제안 시스템과 '사서에게 물어보세요' 서비스를 순차적으로 호출한다. 실험 진행 상황은 CLI(Command Line Interface) 환경에서 실시간으로 모니터링되며, 각 질의어별 검색 결과 건수와 응답시간이 즉시 출력된다. 로그에서 확인할 수 있듯이, 제안 시스템은 일관되게 10건의 결과를 0.1~0.2초 내에 반환하는 반면, '사서에게 물어보세요' 서비스는 평균 1건의 결과를 12~14초에 걸쳐 반환하였다. 실험 완료 후 결과는 JSON과 CSV 형식으로 저장되며, 통계 검정 및 분석 리포트가 자동 생성된다. 〈표 3〉은 검색 시스템 비교 평가 실험을 위한 실험데이터 생성 조건이다.

3.6 분석 지표 및 통계 검정

분석 지표는 평균 결과 건수, 검색 성공률, 평균 응답시간, 오류 건수로 설정하였다. 본 연구에서 응답시간은 시스템 내부의 알고리즘 처리 시간(processing time)이 아닌, 질의를 입력한 시점부터 검색 결과를 확인할 수 있는 시점까지의 전체 소요시간(end-to-end response time)으로 정의한다. 이는 이용자 관점에서 서비스를 이용할 때 실제로 체감하는 응답 특성을 반영하기 위함이다.

두 시스템의 차이는 대응표본 t-검정과 윌콕슨 부호순위 검정을 통해 검증하였다. 대응표본 t-검정은 차이값의 정규성을 가정하는 모수적 검정 방법이다. 그러나 본 연구의 검색 결과 건수 데이터는 제안 시스템이 모든 질의에 대해 10건을 고정적으로 반환하여 분산이 0인 반



〈그림 8〉 검색 시스템 비교 평가 실험의 자동화 스크립트 실행 화면

〈표 3〉 실험데이터 생성 조건

항목	내용
실험 설계	KDC 10개 대분류 × 각 10개 질의어 = 총 100개
무작위 시드	42(재현성 보장)
국중 검색 방식	Selenium WebDriver(Headless Chrome)
병렬 처리	순차 실행(API 과부하 방지, 1초 지연)
총 실험 시간	약 20-25분
출력 형식	JSON(상세), CSV(요약), Markdown(리포트)

면, '사서에게 물어보세요' 서비스는 0건에서 45건까지 극단적인 편차를 보여 정규성 가정을 충족하기 어렵다. 이에 정규성 가정에 의존하지 않는 비모수적 검정인 윌콕슨 부호순위 검정을 병행하여 분석 결과의 강건성(robustness)을 확보하고자 하였다. 이러한 모수적-비모수적 검정의 교차 검증은 소표본 연구에서 통계 분석의 신뢰성을 높이는 일반적인 방법론이다. 통계분석은 Python과 R 환경에서 교차 수행하였다. 1차 분석은 Python의 scipy 패키지를 이용하여 대응표본 t-검정과 윌콕슨 부호순위 검정을 실시하였으며, 분석 결과의 신뢰성을 검증하기

위해 R의 stats 패키지를 사용하여 동일한 검정을 재수행하였다. 시각화는 R의 ggplot2 패키지를 사용하여 박스플롯과 막대그래프를 생성하였다.

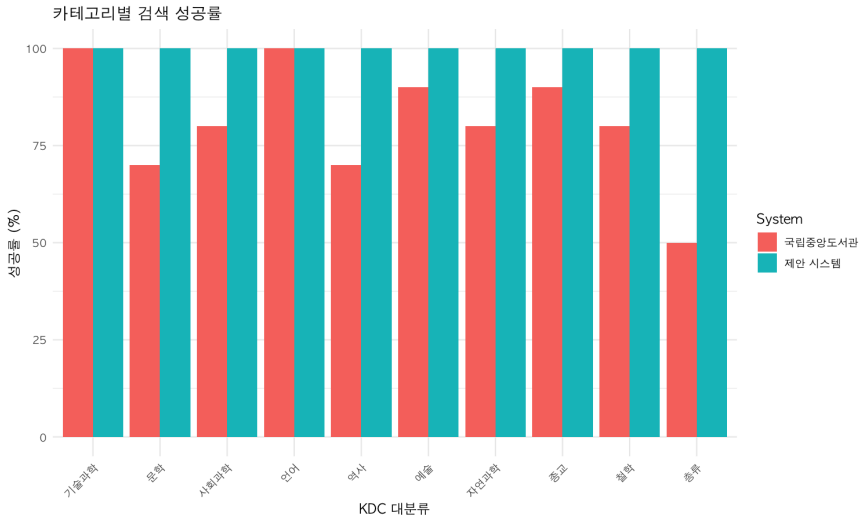
4. 연구 결과

4.1 전체 성능 비교

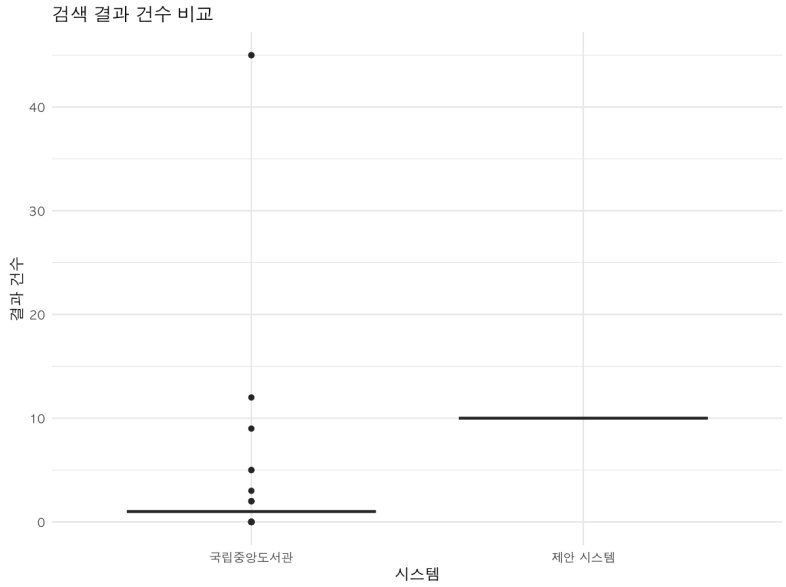
〈그림 9〉는 KDC 10개 대분류별 검색 성공률을 비교한 결과를 나타낸다. 제안 시스템은 모든 카테고리에서 100%의 검색 성공률을 달성하여 주제 영역에 관계없이 일관된 성능을 보였다. 반면 '사서에게 물어보세요' 서비스는 분야에 따라 50%에서 100%까지 큰 편차를 나타냈다. 특히 총류 영역에서 해당 서비스의 성공률은 50%로 가장 낮게 나타났는데, 이는 문헌정보학, 서지학 등 전문 용어가 포함된 질의에 대해 검색 매칭이 제한적으로 이루어졌기 때문으로 해석된다. 반면 기술과학과 언어 분야에서는 두 시스템 모두 100%의 성공률을 기록하였다.

〈그림 10〉은 두 시스템의 검색 결과 건수 분포를 박스 플롯으로 비교한 것이다. 제안 시스템은 모든 질의에 대해 일관되게 10건의 결과를 반환하여 박스 플롯이 단일 선으로 표현되었다(중앙값=10, 표준편차=0). 반면 '사서에게 물어보세요' 서비스는 중앙값이 1건으로, 대부분의 질의에서 매우 적은 결과를 반환하였으며, 일부 질의에서는 45건까지 반환되는 극단적인 이상치가 관찰되었다. 해당 서비스의 결과 분포는 0~1건에 집중된 오른쪽 꼬리 분포(right-skewed distribution)를 보여, 특정 질의어에서만 높은 검색 결과를 반환하는 불균일한 성능 특성을 보였다.

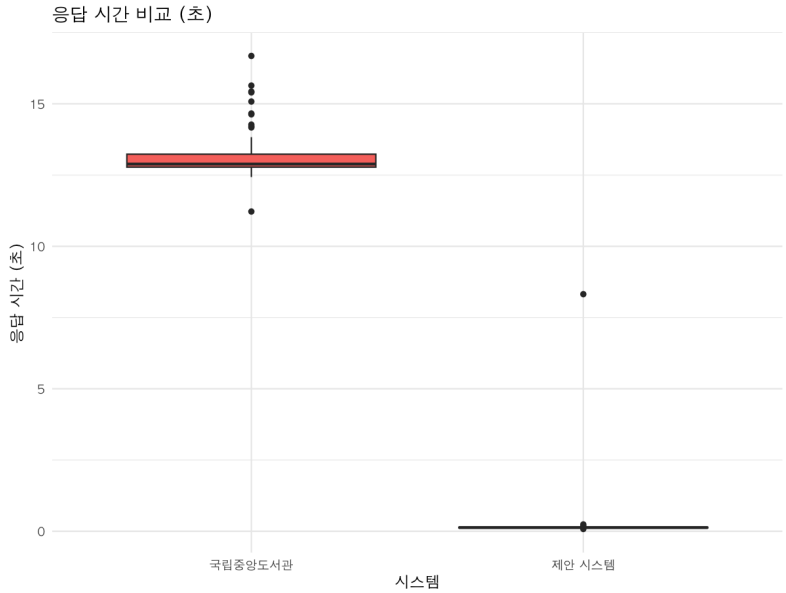
〈그림 11〉은 두 시스템의 응답시간 분포를 박스플롯으로 비교한 결과이다. 제안 시스템은



〈그림 9〉 KDC 10개 대분류별 검색 성공률을 비교한 결과



〈그림 10〉 제안시스템과 ‘사서에게 물어보세요’ 서비스의 검색 결과 건수 분포



〈그림 11〉 제안시스템과 ‘사서에게 물어보세요’ 서비스의 응답시간 분포 박스플롯

평균 0.21초의 응답시간을 기록하여 거의 즉각적인 검색 결과를 제공하였으며, 분포 역시 0초 근처에서 매우 좁은 범위로 나타났다. 반면 ‘사서에게 물어보세요’ 서비스는 평균 13.12초로,

응답시간이 12~14초 범위에 집중되었고 일부 질의에서는 17초에 이르는 이상치가 관찰되었다. 이러한 차이는 제안 시스템이 로컬 인덱스 기반의 검색 구조를 사용하는 반면, '사서에게 물어보세요' 서비스는 웹 환경에서의 실시간 질의 처리와 페이지 렌더링 과정을 포함하는 구조에 의존하기 때문으로 해석된다.

〈표 4〉는 100개 질의에 대해 두 시스템의 전체 성능을 비교한 결과이다. 제안 시스템의 결과 건수 표준편차가 0.00으로 나타난 것은, 모든 질의에 대해 상위 10건을 반환하도록 출력 상한을 고정했기 때문이다. 반면, '사서에게 물어보세요' 서비스의 표준편차 6.33은 질의별 결과 수의 편차가 크다는 것을 의미한다.

〈표 4〉의 결과는 다음과 같은 의미가 있다. 첫째, 결과 건수 측면에서 제안 시스템은 모든 질의에 대해 동일한 결과 제시 능력을 유지하였다. 이는 이용자가 질의어를 세밀하게 조정하지 않더라도 최소한의 탐색 후보군을 안정적으로 확보할 수 있음을 시사한다. 제안 시스템이 모든 질의에 대해 10건의 결과를 반환하는 것은 하이브리드 검색의 구조적 특성에 기인한다. 키워드 일치 결과가 부족하더라도 벡터 검색의 의미적 유사도를 통해 관련 문서를 탐색하므로, 키워드 불일치에 따른 미탐(zero result)을 구조적으로 해소한다. 반면 '사서에게 물어보세요' 서비스의 평균 결과 건수 2.01은 다수의 질의에서 결과가 제한적이거나 0건일 가능성이 크다. 이는 이용자가 원하는 결과를 확보하기 위해 질의를 재구성해야 하는 부담으로 이어질 수 있다.

둘째, '사서에게 물어보세요' 서비스의 검색 성공률 81.0%는 100개 질의 중 19개에서 검색 결과를 확보하지 못했음을 나타낸다. 이는 검색 구조의 한계일 수도 있고, 웹 렌더링이나 세션 처리와 같은 호출 환경의 영향일 가능성도 있다. 그러나 본 연구는 동일 조건에서 실제 호출을 반복하여 측정하였으므로, 이용자 경험의 관점에서는 '일부 질의에서 결과가 제공되지 않는 현상' 자체가 중요한 성능 지표로 해석된다. '사서에게 물어보세요' 서비스에서 검색 결과가 반환되지 않은 19건의 질의를 분석한 결과, 이들 질의 대부분은 지식정보 DB에 관련 답변이 존재함에도 질의어의 표현 방식이 DB에 축적된 질문과 상이하여 키워드 매칭에 실패한 사례로 파악되었다. 예를 들어, '노인복지시설 운영 기준'이라는 질의에 대해 해당 서비스에서는 결과가 반환되지 않았으나, 제안 시스템에서는 '고령자 복지시설', '노인요양시설 관련 법규' 등 의미적으로 유사한 질의응답 문서를 검색하였다. 또한 '한국 현대 미술의 흐름'과 같은 포괄적 질의에서도 해당 서비스는 정

〈표 4〉 실험데이터 생성 조건

지표	제안 시스템	'사서에게 물어보세요' 서비스
평균 결과 건수	10.00	2.01
표준편차	0.00	6.33
검색 성공률	100.0%	81.0%
평균 응답시간	0.21초	13.12초
오류 건수	0건	0건

확한 키워드 일치가 이루어지지 않아 결과를 반환하지 못한 반면, 제안 시스템은 벡터 검색의 의미적 유사도를 활용하여 관련 문서를 제시하였다. 다만, 제안 시스템이 반환하는 10건의 결과가 모두 질의 의도에 적합한 것은 아니며, 상위 결과의 적합도(relevance) 평가는 향후 과제로 남겨둔다.

셋째, 응답시간의 차이는 서비스의 실질적 활용 가능성과 관련된 요인 중 하나이다. 제안 시스템의 평균 0.21초는 대화형 탐색(질의-수정-재질의)이 가능한 수준인 반면, ‘사서에게 물어보세요’ 서비스의 평균 13.12초는 반복 탐색의 비용을 크게 증가시킨다. 다만, 이러한 응답시간의 차이에는 검색 알고리즘의 효율성뿐 아니라 시스템 아키텍처, 네트워크 조건, 페이지 렌더링 방식 등 물리적 환경을 포함한 복합적 요인이 작용하므로, 시스템 자체의 성능 차이로 단정하기 어렵다. 이러한 한계를 고려하여 본 연구의 응답시간 비교는 참고 수준의 관찰 결과로 해석하는 것이 적절하며, 향후 통제된 환경에서의 추가 검증이 필요하다.

4.2 대분류별 성능 비교

〈표 5〉는 KDC 대분류별로 10개씩 총 100개의 질의를 적용하여, 각 시스템의 평균 결과 건수와 검색 성공률을 비교한 결과이다.

제안 시스템은 모든 대분류에서 결과 건수와 성공률이 동일하게 나타났으며, 이는 상위 10건 출력 정책과 인덱스 기반 처리 구조의 안정성을 반영한다. 반면 ‘사서에게 물어보세요’ 서비스는 대분류에 따라 결과 건수와 성공률의 변동 폭이 크게 나타났다. 〈표 5〉에 따르면 주제 영역별 성능 편차가 뚜렷하며 특히 문학 분야는 평균 결과 건수 0.70, 성공률 70%로 낮은 수준을 보였다. 이는 표현의 다양성이 큰 분야에서 키워드 일치 기반 검색이 상대적으로 취약할 수 있음을 보여준다.

총류는 평균 결과 건수가 4.90으로 비교적 높게 나타났음에도 성공률은 50%로 가장 낮았다. 이는 총류에 속한 질의가 범용적이거나 모호한 표현을 포함할 가능성이 높아, 검색 결과가 제시되더라도 의도에 부합하지 않는 경우가 많기

〈표 5〉 대분류별 성능 비교

대분류	질의 수	제안 시스템 평균	‘사서에게 물어보세요’ 평균	제안 시스템 성공률	‘사서에게 물어보세요’ 성공률
기술과학	10	10.00	1.00	100%	100%
문학	10	10.00	0.70	100%	70%
사회과학	10	10.00	1.20	100%	80%
언어	10	10.00	1.10	100%	100%
역사	10	10.00	5.10	100%	70%
예술	10	10.00	1.00	100%	90%
자연과학	10	10.00	2.20	100%	80%
종교	10	10.00	2.00	100%	90%
철학	10	10.00	0.90	100%	80%
총류	10	10.00	4.90	100%	50%

때문으로 해석된다. 이러한 특성은 이용자에게 반복적인 질의 수정과 재탐색의 부담을 증가시킬 수 있다. 반면 제안 시스템은 모든 대분류에서 100%의 성공률을 보였다. 이에 따라 향후 평가에서는 결과 제공 여부 자체보다 상위 결과의 적합도나 설명형 응답의 질과 같은 질적 지표를 핵심 기준으로 포함할 필요가 있다.

4.3 통계 검정

결과 건수에 대해 대응표본 t-검정과 윌콕슨 부호순위 검정을 수행한 결과, 두 시스템 간의 차이는 모두 통계적으로 유의미한 것으로 확인되었다. 이는 관측된 성능 차이가 우연에 의해 발생했을 가능성이 낮으며, 제안 시스템과 '사서에게 물어보세요' 서비스 간의 구조적 차이 즉, 로컬 인덱스 기반의 하이브리드 검색 방식과 키워드 매칭 기반의 기존 검색 방식의 차이에 의한 것으로 해석된다.

본 연구의 통계 분석은 Python과 R 환경에서 교차 검증되었으며, 두 환경에서 동일한 통계량과 유의확률이 산출되어 분석 결과의 재현성과 신뢰성이 확인되었다.

5. 논의 및 결론

본 연구는 국립중앙도서관 '사서에게 물어보세요' 지식정보 DB를 기반으로 하이브리드 검색과 RAG를 결합한 디지털 참고서비스 모델을 설계하였으며, 이를 바탕으로 100개 질의에 대해 제안 시스템과 기존 서비스 간의 비교 실험을 실제 호출 방식으로 수행하였다. 그 결과, 제안 시스템은 응답시간, 검색 성공률, 결과 산출의 일관성 측면에서 기존 서비스보다 전반적으로 우수한 결과를 보였다. 다만, 본 비교는 두 시스템의 전반적 우열을 판단하기 위한 것이 아니라, 동일 지식정보 DB 기반에서 키워드 매칭 방식과 하이브리드 검색 및 RAG 방식 간 서비스 제공 특성의 차이를 실증적으로 확인하기 위한 것이다. 연구의 논의점은 다음과 같다.

첫째, 제안 시스템은 로컬 인덱스 기반 구조와 간결한 처리 흐름을 통해 평균 0.21초의 응답시간을 기록한 반면, '사서에게 물어보세요' 서비스는 평균 13.12초를 기록하였다. 다만 이 차이는 시스템 아키텍처, 네트워크 조건, 페이지 렌더링 방식 등 물리적 환경을 포함한 복합적 요인에 의한 것으로, 향후 통제된 환경에서의 추가 검증이 필요하다.

둘째, '사서에게 물어보세요' 서비스에서 관찰된 검색 결과 건수와 성공률의 변동에는 질의어 표현 방식, 주제 범주의 모호성, 웹 기반 호출 환경 등 복합적인 요인이 영향을 미쳤을 가능성이 있다. 특히 해당 서비스는 동적 렌더링 구조를 갖고 있어, 결과 요소가 완전히 생성되기 이전에 파싱이 수행될 경우 결과 수집에 실패할

〈표 6〉 통계 검정 요약

비교 항목	검정 방법	통계량	결론
결과 건수	대응표본 t-검정	$t = 12.5496$, $p < .001$	유의미
결과 건수	윌콕슨 부호순위 검정	$W = 201.0000$, $p < .001$	유의미

수 있다. 본 연구에서는 이러한 환경적 요인을 통제하기 위해 Selenium 기반의 브라우저 자동화를 적용하여 실제 이용자와 동일한 실행 환경에서 결과를 수집하였으며, 이를 통해 비교 평가의 일관성을 확보하고자 하였다.

셋째, 하이브리드 검색과 RAG의 결합은 결과 목록뿐만 아니라 ‘왜 이 결과가 제시되었는가’에 대한 설명까지 제공할 수 있다. 참고서비스에서 사서의 핵심 역할은 자료를 나열하는 것이 아니라, 이용자의 질문 의도를 해석하고 적합한 정보원을 선정하여 그 이유와 활용 맥락을 안내하는 데 있다. 본 연구에서 구현한 RAG 방식은 검색 결과를 근거로 요약, 추천, 추가 키워드 제안을 수행함으로써 이러한 사서의 안내 기능을 디지털 환경에서 부분적으로 구현하였다. 다만 생성형 모델의 환각 위험을 관리하기 위해서는 검색 맥락에 근거한 응답 생성 원칙, 출처 노출, 사용자 피드백 수집과 같은 운영적 장치가 필수적이며, 이는 향후 실제 서비스 적용 단계에서 체계적으로 설계되어야 할 것이다.

본 연구의 결론을 통한 제언은 다음과 같다. 첫째, 후속 연구에서는 적합도 중심의 평가 지표를 도입할 필요가 있다. 본 연구는 결과 건수, 성공률을 중심으로 비교하였으나, 향후 전문가 판단이나 사용자 평가를 기반으로 정확률, nDCG 등 순위 기반 지표를 적용하여 검색 품질의 실질적 우수성을 검증할 필요가 있다. 특히 제안 시스템이 항상 10건의 결과를 반환한다는 구조적 특성을 고려할 때, 반환된 결과의 적합도를 평가하는 것이 시스템의 실질적 유용성을 판단하는 핵심 과제이다. 둘째, 한국어 특화 임베딩 모델과 형태소 기반 질의 처리를 결합하여 검색 품질을 더욱 고도화하는 방안이 모색되어야 한다. 셋째, 앞서 논의한 RAG 응답의 신뢰성 확보를 위해, 근거 기반 생성 원칙의 구체적인 기준 수립, 출처의 단계별 제시 방식, 사용자 피드백을 반영한 응답 품질 개선 체계 등을 포함한 운영 모델을 체계적으로 설계해야 할 것이다. 넷째, 실제 서비스 환경에서의 동시 접속, 네트워크 지연, 서버 부하 등을 고려한 배포·운영 실험을 수행하여 외부 요인이 성능에 미치는 영향을 추가적으로 검증할 필요가 있다. 다섯째, 향후 연구에서는 API 기반 자동화 등을 통해 질의어 규모를 확대하고, 다양한 난이도와 유형의 질의를 포함하여 연구 결과의 일반화 가능성을 높일 필요가 있다.

본 연구는 범용 임베딩 모델인 all-MiniLM-L6-v2를 사용하였다. 이 모델은 다국어 및 영어 중심 데이터로 학습되어 한국어 참고질의 및 응답 데이터의 미세한 의미 차이나 도메인 특성을 충분히 반영하지 못할 수 있다. 이러한 한계로 인해 본 연구의 결과는 하이브리드 검색 구조의 가능성과 성능 특성을 검증하는 데 의의가 있으나, 임베딩 모델 자체의 성능 최적화를 일반화하기에는 제한적이다. 또한 두 시스템의 응답시간 비교는 물리적 환경 차이가 체감 속도에 영향을 미칠 수 있어, 향후 통제된 환경에서의 추가 연구가 필요하다. 이러한 한계를 바탕으로 향후 연구에서는 한국어 특화 임베딩 모델이나 도메인 적응 파인튜닝을 적용하여 검색 적합도와 의미 구분 능력을 향상시키는 방향으로 연구를 확장할 수 있을 것이다.

참 고 문 헌

김왕종, 이제환 (2016). 한국 도서관계의 '협력형' 디지털참고서비스(CDRS): 문제점과 개선안. 한국도서관·정보학회지, 47(4), 69-91. https://doi.org/10.16981/kliss.47.4.201612.69

임정훈 (2025). 디지털 참고서비스 비교 분석: '사서에게 물어보세요'와 생성형 AI 서비스 사례를 중심으로. 한국비블리아학회지, 36(3), 105-127. https://doi.org/10.14699/KBIBLIA.2025.36.3.57336

장혜란, 이경숙 (2014). 협동 디지털참고서비스의 질문 분석: 국립중앙도서관의 '사서에게 물어보세요'를 중심으로. 정보관리학회지, 31(4), 7-28. https://doi.org/10.3743/KOSIM.2014.31.4.007

장수현, 남영준 (2021). 협력형 디지털 참고서비스(CDRS) 지식정보 DB 내용분석 연구. 한국비블리아학회지, 32(2), 101-123. https://doi.org/10.14699/kbiblia.2021.32.2.101

Alkaissi, H. & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus, 15(2), Article e35179. https://doi.org/10.7759/cureus.35179

Borgman, C. L. (1996). Why are online catalogs still hard to use? Journal of the American Society for Information Science, 47(7), 493-503.

Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 758-759. https://doi.org/10.1145/1571941.1572114

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. arXiv. https://arxiv.org/abs/2312.10997

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248, 1-38. https://doi.org/10.1145/3571730

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 6769-6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J.,

Widmann, S., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Lankes, R. D. (2004). The digital reference research agenda. Journal of the American Society for Information Science and Technology, 55(4), 301-311.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474. https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 3982-3992. https://doi.org/10.18653/v1/D19-1410

Shuster, K., Komeili, M., Poff, S., Chen, K., Ram, A., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. arXiv. https://arxiv.org/abs/2104.07567

Thakur, N., Reimers, N., Ranade, A., Bhatt, A., & Gurevych, I. (2021). BEIR: a heterogeneous benchmark for zero-shot evaluation of information retrieval models. arXiv. https://doi.org/10.48550/arXiv.2104.08663

Voorhees, E. M. & Harman, D. K. eds. (2005). TREC: experiment and evaluation in information retrieval. Cambridge, MA: MIT Press.

Vtyurina, A., Savenkov, D., Agichtein, E., & Clarke, C. L. A. (2017). Exploring conversational search with humans, assistants, and creative writing. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 4553-4561. https://doi.org/10.1145/3025453.3025654

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In Advances in Neural Information Processing Systems, 33, 5776-5788.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Chang, Hye-Rhan & Yi, Kyung-Suk (2014). Question analysis of the collaborative digital reference service at the National Library of Korea. Journal of the Korean Society for

Information Management, 31(4), 7-28. https://doi.org/10.3743/KOSIM.2014.31.4.007

Jang, Su-Hyun & Nam, Young-Joon (2021). Content analysis of collaborative digital reference service knowledge information database. Journal of the Korean Biblia Society for Library and Information Science, 32(2), 101-123. https://doi.org/10.14699/kbiblia.2021.32.2.101

Kim, WangJong & Lee, Jae-Whoan (2016). Development and management of CDRS in Korean library community. Journal of Korean Library and Information Science Society, 47(4), 69-91. https://doi.org/10.16981/kliss.47.4.201612.69

Lim, Jeonghoon (2025). A comparative analysis of digital reference services: focusing on "Ask a Librarian" and generative AI services. Journal of the Korean Biblia Society for Library and Information Science, 36(3), 105-127. https://doi.org/10.14699/KBIBLIA.2025.36.3.57336