ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
https://www.jiisonline.org
J Intell Inform Syst 2025 September: 31(3): 211~225 http://dx.doi.org/10.13088/jiis.2025.31.3.211

SLLM을 활용한 RAG기술 기반의 교육기관 문서 질의응답 기법 활용 방안*

이현우
동국대학교 핀테크블록체인학과
(soul8383@dongguk.edu)

김경재
동국대학교 경영정보학과
(kjkim@dongguk.edu)

이영섭
동국대학교 통계학과
(yung@dongguk.edur)

최근 대규모 언어모델(Large Language Model, LLM)은 자연어 처리의 비약적인 발전으로 다양한 교육 현장에서 활용되고 있다. 특히 검색 증강 생성(Retrieval-Augmented Generation, RAG) 기술은 대규모 언어모델이 보유하지 않은 최신 정보나 특정 도메인 지식을 외부 문서에서 실시간으로 검색하여 활용함으로써, 더욱 정확하고 신뢰할 수 있는 답변을 생성할 수 있게 한다. 그러나 대규모 언어모델을 직접 사용하는 방식은 높은 비용 부담과 개인정보보호 문제로 인해 교육기관에서 활용하는데 제약이 따른다. 또한 교육기관의 문서들은 입시 모집요강이나 학사제도 안내서처럼 도메인에 특화 되어 있으며, 복잡한 표나 이미지가 포함된 구조화된 정보가 풍부하다는 특징이 있다. 기존 일반적인 대규모 언어모델 기반 챗봇은 몇 가지 한계를 가진다. 첫째, 전형일정이나 모집단위와 같이 셀이 병합되고 복잡한 표 형태로 정리된 정보를 정확하게 추출하는 능력이 미흡하다. 둘째, 장·절·하위 절로 구성된 긴 문서에서 필요한 단락만 정확히 찾아내지 못하는 단점이 있다. 또한 매 질의마다 수백 또는 수천개의 토큰을 재처리해 비용이 많이 발생하는 한계를 가진다. 이러한 문제점을 해결하기 위해 본 연구에서는 소규모 언어모델(Small Language Model, SLLM)과 RAG를 결합한 방법을 제안한다. 섹션 기반 문서 파싱 기법을 통해 문서를 체계적으로 분석하고, 복잡한 표 형태의 데이터를 소규모 언어모델이 정확하게 처리할 수 있는 방법을 제시함으로써, 유료모델에 대한 경제적, 기술적 대안을 마련하고자 한다.

주제어 : SLLM, RAG, 챗봇, 문서 질의응답 시스템

논문접수일 : 2025년 8월 25일 논문수정일 : 2025년 9월 17일 게재확정일 : 2025년 9월 19일
원고유형 : Regular Track 교신저자 : 이영섭

1. 개요

최근 GPT, Claude, Gemini 등 대규모 언어모델의 급속한 발전은 자연어 처리(Natural Language Processing, NLP) 분야에 있어 혁신적 성과를 이루어내고 있으며, 이는 교육 분야 뿐만 아니라 다양한 분야에서 활용되고 있다. 대규모 언어모델은 텍스트 요약, 지식 검색 등 여러 응용 분야에서 뛰어난 성능을 보여주고 있지만, 현실적으로 이러한 모델을 교육기관에서 질의응답 시스템에 활용하기에는 많은 제약과 한계가 존재한다.

우선, 대규모 언어모델을 실시간으로 활용할 때 발생하는 높은 API 사용 비용의 단점이 있다. 일반적인 대규모 언어모델의 경우 매 질문마다 긴 문서의 내용을 전부 다시 분석하여 처리하는 방식으로 운영되기 때문에, 매 다른 질의마다 처리

* 본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(No.2021R1A2C1007095)과 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업(IITP-2025-RS-2020-II201789)의 연구결과로 수행되었음.

해야 하는 토큰 수가 수백에서 수천 단위에 이르고 이는 결국 API 호출 비용 증가 및 응답 속도 지연으로 이어진다. 그리고 교육기관이 보유한 입시 모집요강, 학사제도 안내 등 도메인 특화 문서들은 일반적인 문서와는 다르게 문서의 구조도가 복잡한 경우가 많다. 특히 이러한 교육기관 문서들은 대개 대분류, 중분류, 소분류 등의 계층적 구조를 가지며, 전형 일정·모집 인원·성적 반영 비율과 같은 핵심적인 정보들이 표나 이미지 형태로 구조화되어 있는 경우가 흔하다. 그러나 일반적인 대규모 언어모델 기반의 챗봇은 복잡한 구조를 가진 문서 내에서 사용자의 질문에 정확히 대응하거나, 복잡한 표 형식으로 구성된 데이터에서 특정 값을 정확하게 추출하는 데 어려움이 있다. 예컨대 모집요강 같은 문서에서 지원 자격이나 모집 인원과 같은 정보가 복잡한 표 형식으로 제시되어 있는 경우, 상용 대규모 언어모델조차 내용을 정확히 해석하지 못하는 경우가 존재한다.

이러한 한계를 극복하기 위해 최근에는 소규모 언어모델(Small Language Model, SLLM)과 정보검색 기술을 결합한 질의응답 시스템 개발이 활발히 이루어지고 있다. 소규모 언어모델은 특정 도메인이나 기관의 데이터로 쉽고 빠르게 파인튜닝할 수 있어, 각 교육기관이 보유한 고유한 문서와 지식을 정밀하게 반영하는 문서 기반 질의응답 시스템을 구축할 수 있다는 점에서 그 필요성이 대두되고 있다. 실제로 특정 전문지식에 대한 질의응답 시스템의 연구개발을 통해, 사용자에게 빠르고 정확하게 답변을 제공할 수 있음을 확인하였다(Kim & Yu, 2023). 또한, 소규모 언어모델은 기존의 대규모 언어모델 대비 상대적으로 적은 수억개에서 수십억개 정도의 파라미터를 가진 경량화된 언어모델로, 적은 하드웨어 자원으로도 빠른 추론 속도와 효율성을 제공할 수 있는 장점이 있다. 즉, 소규모 언어모델은 기존의 대규모 언어모델 대비 구축과 운영 비용이 저렴하여 예산과 인프라가 제한적인 교육기관에서도 현실적으로 도입이 가능한 대안으로 평가받고 있다. Woo et al.(2025)은 소규모 언어모델을 기반으로 한 파인튜닝을 수행하여 효과적인 답변을 생성하는 Q&A 시스템을 제안하였다. 이와 함께 윤요섭과 안현철(2024)은 RAPTOR(Recursive Abstractive Processing for Tree-Organized Retrieval) 기법을 활용한 질의응답 시스템을 개발하여 계층적 정보 검색의 효율성을 입증하였다. 특히 주목할 만한 점은 검색 증강 생성(Retrieval-Augmented Generation, RAG) 기법의 활용이다.

RAG 기술은 2020년 Facebook AI Research에서 처음 제안된 하이브리드 자연어처리 기법으로(Lewis et al., 2020), 사전 훈련된 언어 모델의 매개변수에 저장된 지식만으로는 한계가 있는 문제를 해결하기 위해 외부 지식 베이스에서 관련 정보를 동적으로 검색하여 생성 과정에 활용하는 접근법이다. 문서 기반 질의응답 시스템은 방대한 문서에서 질문에 답할 수 있는 관련된 정보를 검색하여 텍스트를 생성해야 하는데, RAG 기술은 이러한 정보 검색(Retrieval)과 텍스트 생성(Generation)을 결합한 효과적인 해결책을 제공한다. Xiong et al.(2024)의 연구에 따르면, RAG 기술을 적용함으로써 GPT-3.5와 Mixtral과 같은 중간 규모 언어 모델의 성능을 GPT-4 수준까지 향상시킬 수 있는 것으로 보고되었다. 이와 관련하여 정천수(2023)는 대규모 언어모델 애플리케이션 아키텍처를 활용한 생성형 AI 서비스 구현 방안을 제시하면서, RAG 기술과 LangChain 프레임워크를 기반으로 한 실용적인 구현 방법론을 제안하였다. 이는 상대적으로 작은 모델에서도 외부 지식베이스를 효과적으로 활용할 경우

대규모 언어 모델에 준하는 성능을 달성할 수 있음을 시사한다.

최근 RAG 기술은 교육 현장에서도 다양한 형태로 응용되고 있으며, 특히 질의응답 시스템이나 과제 피드백 도구로서의 활용이 활발하다. 학습자가 질문을 입력하면 RAG 기술이 교과서, 학습 자료, 논문, 온라인 백과사전 등의 외부 지식 기반에서 관련 정보를 검색하고, 이를 바탕으로 응답을 생성함으로써, 교사나 교수자가 제공하는 일방적인 정답 외에도 다양한 관점과 심층 정보를 자동으로 제공할 수 있다는 점에서 주목받고 있다(윤여찬, 김수균, 2025).

본 연구에서는 상대적으로 규모가 작고 효율적인 소규모 언어모델과 RAG 기술을 결합한 교육기관 맞춤형 문서 질의응답 시스템을 제안한다. 이를 위하여 문서의 구조를 적극적으로 활용할 수 있는 섹션 기반의 계층적 문서 파싱(Document Parsing) 전략을 적용한다. 이는 단순히 문서를 일정 크기의 청크(chunk)로 무작위 분할하는 기존의 방식과 달리, 문서 내 장·절의 계층 구조를 명시적으로 파싱하여 단락으로 정보 단위를 형성하고, 이를 벡터 데이터베이스에 저장하고 인덱싱함으로써 검색 정확도를 높인다. 추가적으로 교육기관 문서에서 자주 나타나는 표 형태의 데이터를 보다 효율적으로 추출하고 처리하기 위해, 상용화된 멀티모달 언어모델(Multimodal LLM)을 활용한 새로운 접근 방식을 연구하였다. 이 연구에서는 표를 Markdown과 HTML 형식으로 각각 변환하여 모델의 입력으로 활용하는 두 가지 방법을 비교분석하였다. 두 방식 각각의 정확도, 응답 속도, API 호출 비용을 측정한다. 이를 통해 다양한 현실적인 문서 형태에 대한 최적의 표 처리 방식을 제안한다.

2. 이론적 배경

2.1 소규모 언어모델

소규모 언어모델은 자연어를 처리하고 이해하며 생성할 수 있는 인공지능 언어 모델로, 대규모 언어모델에 비해 작은 규모와 범위를 갖는 것이 특징이다. 일반적으로 수억에서 수십억개의 파라미터를 가진 트랜스포머 기반 디코더 전용 언어 모델로 정의되며, 일부 연구에서는 10억 개 미만의 파라미터를 가진 모델을 소규모 언어모델으로 분류하기도 한다.

소규모 언어모델은 대규모 언어모델과 비교할 때 다음과 같은 주요 특징을 가진다. 첫째, 더 적은 연산 자원과 비용으로 운영이 가능하여 경량화된 환경에서의 활용이 용이하다. 둘째, 특정 작업에 특화된 데이터로 훈련되어 대규모 언어모델보다 정밀하게 튜닝되는 경우가 많다. 셋째, 중소 규모 조직에서도 대규모 언어모델 기술을 쉽게 활용할 수 있는 접근성을 제공하는 것을 목적으로 한다.

본 연구에서는 소규모 언어모델 중 하나인 Mixtral-8×7B모델을 채택하였다. 해당 모델은 Mistral AI에서 개발한 MoE(Mixture of Experts) 아키텍처를 기반으로 한 오픈소스 언어 모델로, Apache 2.0 라이선스 하에 자유롭게 활용이 가능하다. Mixtral-8×7B의 구조적 특징은 총 8개의 전문가 네트워크 중 입력에 따라 최적의 2개만을 선택적으로 활성화하여 추론을 수행한다는 점이다. 이로 인해 전체 파라미터 수는 47B에 달하지만, 실제 추론 시에는 13B 수준의 활성 파라미터만 사용되어, LLaMA 2 70B에 비해 약 5배 낮은 연산 자원으로 더 빠르고 효율적인 처리가 가능하다. 실제 성능 면에서도 Mixtral-8×7B는 GPT-3.5와 유사하거나 일부 벤치마크에서는 이를 상회하며,

특히 ARC(Abstraction and Reasoning Corpus) 벤치마크에서 84.4%의 정확도를 기록하며 오픈소스 모델 중 최고 수준의 성능을 보였다.

또한 본 연구에서는 Mistral AI가 개발한 또 다른 소규모 언어모델인 Mixtral-7B 모델도 선택하였다. 이 모델은 전문가 선택 방식 없이 단일 전문가 모델로 구성되어 있으며, 7B 규모의 파라미터를 통해 경량성과 성능의 균형을 추구한 모델이다. Mixtral-7B는 특히 짧은 지연 시간과 우수한 단일 토큰 생성 성능으로 인해 제한된 자원 환경에서의 활용 가능성이 높으며, 다양한 벤치마크에서 GPT-3.5에 근접한 성능을 보여주고 있다.

이들 Mixtral 모델들은 한국어 지원이 가능하며, 교육기관과 같은 제한된 연산 자원을 가진 환경에서도 독립적인 운영이 용이하다는 장점을 가지고 있다. 특히 클라우드나 GPU 인프라가 부족한 기관에서도 로컬 서버 또는 경량 클러스터 환경에서 실용적인 성능을 발휘할 수 있다는 점에서 높은 활용 가능성을 가진다.

2.2 문서 파싱

문서 파싱이란 PDF, 스캔된 이미지, 차트, 표 등과 같이 비구조화된 문서로부터 프로그램이 처리할 수 있는 구조화된 형태로 데이터를 변환하는 과정을 의미한다. 최근 대규모 언어모델 기반의 응용 프로그램이 활발히 도입됨에 따라, 원본 문서의 내용을 완전하고 정확히 보존하기 위한 고도의 문서 파싱 기술이 요구되고 있다. 특히, 텍스트뿐만 아니라 이미지, 표, 헤더와 같은 문서의 다양한 구조적 요소를 정밀하게 인식하여 구조화된 형태로 변환하는 작업은 대규모 언어모델의 성능과 정확성 향상에 필수적인 전처리 과정으로 자리잡고 있다.

본 연구에서는 문서 파싱을 위해 Upstage Document Parsing API를 채택하였다. 이 API는 복잡한 문서 레이아웃에서도 뛰어난 구조 인식 정확도를 보여주며, 특히 표와 차트의 구조적 정보를 정확하게 보존하는 데 강점을 보인다. 또한 다양한 형식의 문서를 구조화된 형태로 변환하여 원문 의미의 손실 없이 효율적인 정보 추출이 가능하며, 페이지당 평균 0.6초의 빠른 처리 속도와 함께 TEDS-S(Structural Tree Edit Distance-based Similarity) 기준 94% 이상의 표 인식 정확도를 기록하는 등 우수한 성능을 입증하고 있다.

3. 소규모 언어모델을 활용한 RAG 기술 기반의 문서 질의응답 시스템

본 연구에서는 소규모 언어모델을 활용한 RAG 기술 기반의 질의응답 시스템의 성능을 종합적으로 평가하기 위한 연구 대상으로 대학 입시모집요강을 선정하였다. 이러한 선택은 다음과 같은 특성을 고려한 것이다.

첫째, 입시 모집요강은 20페이지 이상의 방대한 분량으로 구성되어 있으며, 여러 개의 다양한 전형 구조로 되어 있다. 또한 학과별, 전형별로 서로 다른 지원자격 및 평가기준을 제시하고 있어 매우 복잡한 정보구조를 가지고 있다.

둘째, 이러한 문서들은 명확한 장·절·항 구분된 체계적인 문서 구조를 갖추고 있다. 동시에 모집인원표, 전형일정표, 수능 반영비율표, 교과 반영비율표 등 다양한 형태의 복잡한 표들이 포함되어 있다. 특히 표 기반 질의는 표를 해석하여 다수의 셀에 걸친 정보를 검색, 추론, 통합해 답변해야 하는 복합적인 작업이므로 일반적인 텍스트 질의응답보다 난이도가 높다.

셋째, 입시 모집요강은 법적·행정적 근거 문서이므로 요약적 제시나 불완전·부정확한 응답이 허용되지 않는다. 예컨대 “학생부 교과 반영비율은 70%”와 같이 단편적 수치를 제시하면, 실제로는 학과·전형·지원계열별로 크게 달라질 수 있다. 수험생이 이 정보를 그대로 활용할 경우 지원 자격 미충족·전형 선택 오류로 이어질 위험이 크다. 따라서 질의응답 시스템은 정확한 수치와 조건을 그대로 인용함과 동시에 해당 근거(페이지, 장·절, 표·셀 좌표)를 함께 제시해야만 서비스 품질 기준을 충족할 수 있다.

넷째, 모집요강의 계층 구조상 동일 질의에 대해 복수의 답변이 공존할 수 있다. 예를 들어 “사회과학계열 전형일정”을 묻는 질문은 수시전형에서 학생부전형, 논술전형 등 여러 전형표에 흩어져 기술되어 있으며, 각 전형마다 원서접수·면접·합격자 발표 날짜가 상이하다. 이에 따라 질의응답 시스템은 전형을 계층별로 구분 태깅으로 관련 섹션을 모두 검색하고, 다중 결과를 병렬로 제시하며, 각 결과마다 식별 가능한 근거를 제시해야 한다.

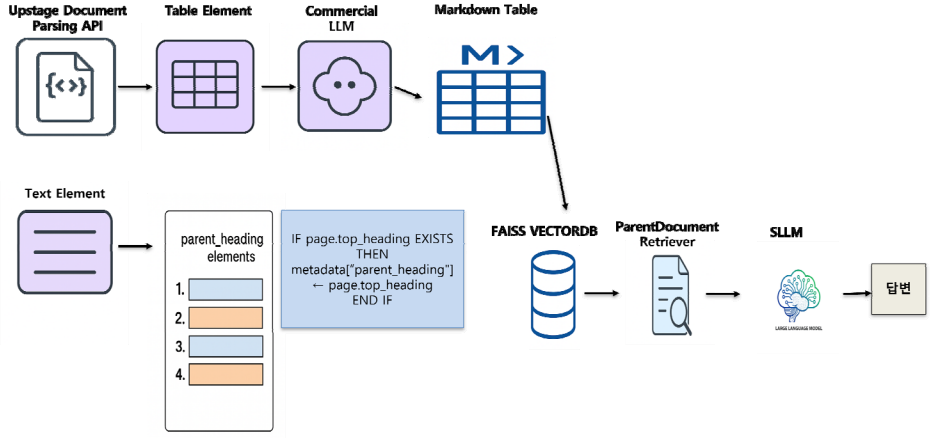
마지막으로, 입시 모집요강은 연간 입시상담 문의가 발생하는 실제적인 문서로 실제 활용 가능성이 높아 실용적 가치가 크다고 판단된다.

<그림 1>은 입시모집요강 문서에 대한 소규모 언어모델을 활용한 RAG 기술 기반의 질의응답 시스템에 대한 전체 그림으로, 각 단계별 설명은 다음과 같다.

3.1 문서 파싱

본 연구에서 제안하는 문서 파싱은 다음의 순차적 처리 단계를 통해 구현되었다. 첫 번째 단계에서는 PDF, 이미지 등 다양한 형태의 원본 문서를 Upstage Document Analysis API에 전송하여 문서의 레이아웃과 구성 요소를 분석한다. 이후 API로부터 수신된 JSON 형태의 분석 결과를 파싱하여 문서 내 각 요소의 위치, 유형, 속성 정보를 추출한다.

두 번째 단계에서는 추출된 문서 요소들을 텍스트, 이미지, 테이블의 세 가지 주요 카테고리로 분류한다. 이러한 분류 과정은 각 요소의 특성에 맞는 최적화된 후처리 방법을 적용하기 위해 필수적이다. 특히 테이블 요소의 경우, 테이블 영역을 원본 문서에서 이미지 형태로 크롭한 후, 이를 멀티모달 언어모델에 입력하여 Markdown 형태의



〈그림 1〉 문서 질의응답 시스템

구조화된 데이터로 변환한다.

최종 단계에서는 각각 처리된 텍스트, 이미지, 테이블 요소들을 문서의 원본 구조와 순서를 유지하면서 통합하여 완전히 문단의 구조를 유지한 구조화된 형태의 문서 데이터를 생성한다. 이러한 체계적인 워크플로우를 통해 복잡한 레이아웃을 가진 문서에서도 일관되고 정확한 정보 추출이 가능하다.

3.2 텍스트 청킹(Text Chunking)

문서의 텍스트 청킹 과정에서는 두 가지 주요 전략을 비교 적용하였다. 첫째, 계층적 구조 기반 청크 분할 전략을 적용하였다. 입시모집요강의 장, 절, 항으로 구성된 명확한 계층 구조를 파악하여 각 장을 하나의 의미적 단위로 하는 청크를 구성하였다. 이를 통해 문서의 원본 구조적 정보를 유지하면서도 의미적 연관성을 보장할 수 있었다. 또한 페이지 맨 상단에 헤더 정보가 존재하면 청킹된 문서의 meta 정보에 해당 정보를 별도로 관리하여, 사용자의 질의가 주어지면, 벡터화된 문서 정보와의 유사도 검사를 수행하여 0.9 이상의 임계값을 만족하는 경우 해당 문서의 내용을 검색 대상으로 포함시켰다. 사용자의 질의는 본문의 구체적인 내용을 물어보는 경우도 있지만, 헤더 정보를 기반으로 큰 틀에서 물어보는 경우도 있기 때문에, 특정 질의가 입력되었을 때 해당 질의와 각 청크의 헤더 정보 간의 의미적 유사도를 계산하여, 유사도가 높은 헤더를 가진 청크를 우선적으로 검색 결과에 포함시키는 방식을 적용하였다. 이는 문서 내 특정 섹션에 대한 정확한 정보 접근을 가능하게 한다.

둘째, 본 연구에서 제안하는 방법의 성능을 객관적으로 평가하기 위해 널리 사용되는 전통적인 접근법인 Sliding Window Chunking 방식을 비교 대상으로 선정하였다. 이 방법은 문서를 고정된 크기의 청크로 분할하되, 인접한 청크 간에 일정 비율의 중복을 두어 문맥의 연속성을 보장하는 방식이다. 구체적으로, 청크 크기는 1,000 토큰으로, 중첩 윈도우는 200 토큰으로 설정하여 인접 청크 간 20%의 중복률을 유지하도록 구성하였다. 이러한 매개변수 설정은 기존 연구에서 효과적인 성능을 보인 것으로 보고된 값들을 참고하여 결정하였다.

Sliding Window Chunking 방식의 주요 특징은 다음과 같다. 첫째, 문서의 구조나 의미적 경계와 무관하게 기계적으로 분할을 수행한다. 둘째, 청크 간 중복 영역을 통해 경계에서 발생할 수 있는 정보 손실을 최소화한다. 셋째, 구현이 단순하고 모든 종류의 문서에 일관되게 적용할 수 있다는 범용성을 갖는다. 그러나 이 방법은 문서의 논리적 구조를 고려하지 않기 때문에 문단이나 섹션의 중간에서 분할이 발생할 수 있으며, 이로 인해 의미적 완결성이 떨어질 수 있다는 한계가 있다. 본 연구에서는 이러한 전통적 방법과의 비교를 통해 구조 기반 청킹 방법의 효과를 검증하고자 한다.

3.3 RAG 시스템 구성

3.3.1 임베딩 모델(Embedding Model)

문서 및 쿼리의 의미적 표현을 위해 Upstage의 Solar-Embedding-1-Large-Passage 모델을 채택하였다. 이 모델은 4,096차원의 임베딩 벡터를 생성하며, 특히 긴 문서를 처리하는데 최적화되어 있어 문서 검색 및 RAG 분야 연구에서 널리 활용되고 있다(Lee et al., 2024).

3.3.2 검색기(Retriever)

문서의 맥락 유지 및 정밀한 검색 결과를 제공하기 위해 Parent Document Retriever를 사용하였다.

이는 LangChain에서 제공하는 고급 검색 전략으로, 문서를 두 단계로 나누어 처리하는 방식이다. 먼저 원본 문서를 큰 단위의 부모 문서(parent documents)로 분할하고, 이를 다시 작은 자식 청크(child chunks)로 세분화한다. 검색 시에는 작은 자식 청크를 통해 정확한 매칭을 수행하지만, 실제 반환되는 것은 해당 청크가 속한 부모 문서이다. 이러한 계층적 접근 방식은 검색의 정밀도를 높이면서도 충분한 맥락 정보를 제공할 수 있어, 생성 모델이 보다 완전하고 일관성 있는 답변을 생성할 수 있도록 한다.

3.3.3 **벡터 데이터베이스**(Vector Database)

임베딩된 문서 벡터의 효율적 저장 및 유사도 기반 검색을 위해 FAISS(Facebook AI Similarity Search)를 사용하였다. FAISS는 대규모 벡터 데이터에 대해 신속하고 정확한 유사도 검색을 지원하는 고성능 벡터 데이터베이스로, 근사 최근접 이웃(Approximate Nearest Neighbor) 알고리즘을 통해 검색 속도와 정확도 간의 최적의 균형을 제공한다. 이는 많은 양의 교육 자료를 실시간으로 처리하고 검색하는 데 적합하다.

본 연구에서는 Inner Product(IP) 인덱스를 채택하여 벡터 간 유사도를 측정한다. IP 인덱스는 두 벡터의 내적(dot product)을 통해 유사도를 계산하는 방식으로, 정규화된 임베딩 벡터에서 코사인 유사도와 동일한 결과를 제공한다. 이는 특히 Solar-Embedding 모델에서 생성된 정규화된 벡터들 간의 의미적 유사성을 효과적으로 측정할 수 있으며, 계산 복잡도가 낮아 대용량 문서 컬렉션에서도 빠른 검색 성능을 보장한다.

특히 문서 질의응답 시스템에서 IP 인덱스는 질문과 문서 간의 의미적 관련성을 정확하게 포착하여 높은 정밀도의 문서 검색을 가능하게 한다. 사용자의 질문이 다양한 형태로 표현되더라도 의미적으로 관련된 문서를 효과적으로 식별할 수 있으며, 검색 결과의 순위 매김에서도 우수한 성능을 보인다.

3.3.4 **소규모 언어모델**

본 연구에서는 소규모 언어모델로 먼저 약 87억 개의 매개변수를 보유한 Mixtral-8×7B 모델을 사용하였다. 해당 모델은 Hugging Face API를 통해 클라우드 기반으로 모델을 호출하여 사용였으며, 로컬 환경에서 직접 구동할 경우 약 90-100GB의 VRAM 또는 시스템 RAM이 요구된다.

두 번째로는 매개변수가 7억 개로 Mistral-7B를 사용하였는데, 이 모델은 경량화된 구조 덕분에 상대적으로 빠른 추론 속도와 효율적인 메모리 사용을 특징으로 한다. 일반적인 질의응답 태스크에서 우수한 성능을 보이며, 실시간 서비스 환경에 적합하다. 로컬 환경에서 구동 시 약 14-16GB의 VRAM 또는 시스템 RAM이 필요하여 상대적으로 접근 가능한 하드웨어 요구사항을 가진다. 이는 일반적인 GPU 서버나 고성능 워크스테이션에서도 운영이 가능한 수준이다. 정천수(2024)의 연구에서는 Mistral-7B를 활용한 실제 보험 금융 도메인 대규모 언어모델을 생성하는 방법 및 구현 절차를 제시하고 있으며, Mistral-7B의 실용적 활용 가능성을 보여주었다.

4. 실험 및 결과

4.1 데이터셋

본 연구의 데이터셋은 두 가지 핵심 요소로 구성하였다. 첫 번째는 병합된 셀을 포함한 복합

테이블에서 모집정원, 전형별 인원 등 정확한 숫자 정보 추출하는 질문을 구성하였다. 두 번째는 합격자 발표 일정, 서류 제출 기한, 유의사항 등의 완전한 문단 내용을 그대로 제공하는 정보 정확성 평가용 질문을 구성하였다.

4.2 평가지표

본 연구에서는 소규모 언어모델이 생성한 자연어 답변의 품질을 객관적으로 평가하기 위해 사전 학습된 다국어 언어 모델 기반의 평가 지표인 BERTScore를 활용하였다(Zhang et al., 2020).

BERTScore는 기존의 어휘 기반 평가 지표인 BLEU(Bilingual Evaluation Understudy), ROUGE (Recall-Oriented Understudy for Gisting Evaluation) 등과 차별화되는 핵심 특징을 갖는다(Papineni et al., 2002; Maples, 2017). 전통적인 지표들이 단순한 n-gram 중첩 정도를 통해 표면적 유사성만을 측정하는 반면, BERTScore는 BERT의 문맥적 임베딩을 활용하여 생성 문장과 참조 문장 간의 의미적 유사성을 깊이 있게 분석한다. 이러한 접근 방식은 동일한 의미를 다른 표현으로 나타낸 문장들 간의 유사성을 보다 정확하게 포착할 수 있다.

BERTScore의 계산 과정은 다음과 같다. 먼저 생성 문장과 참조 문장의 각 토큰을 BERT 모델을 통해 고차원 벡터 공간의 임베딩으로 변환한다. 이후 각 토큰별 임베딩 간의 코사인 유사도를 계산하여 토큰 수준에서의 의미적 대응 관계를 파악한다. 이를 바탕으로 정밀도, 재현율, 그리고 이 둘의 조화평균인 F1 점수를 산출한다.

정밀도는 생성된 문장의 각 토큰이 참조 문장의 토큰들과 얼마나 잘 일치하는지를 나타내며, 재현율은 참조 문장의 각 토큰이 생성 문장의 토큰들과 얼마나 잘 일치하는지를 나타낸다. F1 점수는 정밀도와 재현율의 조화 평균으로, 두 지표를 종합적으로 고려한 의미적 유사성을 나타낸다.

BERTScore의 점수 범위는 0에서 1 사이이며, 값이 높을수록 생성된 답변과 참조 문장 간의 의미적 유사성이 높음을 의미한다. 본 연구에서는 각 소규모 언어모델이 생성한 답변과 해당 질문에 대한 정답 문장 간의 BERTScore의 정밀도, 재현율, F1 점수를 산출하여 대규모 언어모델별 답변 생성 성능을 정량적으로 비교·분석하였다.

4.3 실험설계

4.3.1 표 형식 변환 기반 질의응답 실험

본 연구에서는 표 이미지 데이터의 구조화 방식이 질의응답 성능에 미치는 영향을 분석하기 위해 다음과 같은 실험을 설계하였다.

실험에는 복잡한 병합 셀 구조를 포함한 실제 입학 전형 안내 PDF에서 추출한 표 이미지를 사용하였다. 분석에 사용된 표 이미지 구조는 <그림 2>와 같다.

대학	모집단위		계열	재외국민 (일반)	전교육과정 국외이수자	북한이탈주민
불교	불교학부			-		
	문화유산학과★			1		
문과	국어국문・문예창작학부	국어국문학/문예창작학 /뉴미디어한국어문화	인문	2		
	영어영문학부	영어문학/영어통번역학		2		
	일본학과			1		
	중어중문학과			2		
	철학과			-		
이과	사학과		자연	1	모집단위별 약간 명	모집단위별 약간 명
	수학과			1		
	화학과			1		
	통계학과			1		
	물리학과			-		
법과	법학과		인문	2		
사회과학	정치행정학부	정치외교학		2		
		행정학		2		
		북한학		-		
	경제학과			1		
	국제통상학과			2		
	사회・언론정보학부	사회학		-		
		미디어커뮤니케이션학		2		
	식품산업관리학과			-		
	광고홍보학과			2		
	사회복지학과			-		
경찰사법	경찰행정학부	경찰학/산업보안/범죄과학/교정학		3	3	-
경영	경영학과			3		
	회계학과			1		
	경영정보학과			1		
네이오 시스템	바이오환경과학과			-		
	생명과학과			-		
	식품생명공학과			1		
	의생명공학과			1		
공과	전자전기공학부		자연	3	모집단위별 약간 명	모집단위별 약간 명
	정보통신공학과			1		
	건설환경공학과			-		
	화공생물공학과			2		
	기계로봇에너지공학과			2		
	건축공학부	건축공학/건축학		-		
	산업시스템공학과			1		
	에너지신소재공학과			1		
첨단융합	컴퓨터AI학부★			4		
	시스템반도체학부			1		
사범	국어교육과		인문	1	-	-
	수학교육과		자연	1	-	-
예술	영화영상학과		예체능	2	약간 명	약간 명
계				54	-	

〈그림 2〉 복잡한 표 이미지

표 데이터의 구조화를 위해 상용 대규모 언어 모델 기반 서비스인 ChatGPT-4o 및 Claude Sonnet4.0, Gemini 2.5 Pro 모델을 활용하여 표 이미지를 HTML과 Markdown 형식으로 각각 변환하였다. 이를 통해 동일한 표 데이터에 대한 서로 다른 구조적 표현을 확보하였다.

변환된 HTML 및 Markdown 형식의 표 데이터를 Mistral-7B와 Mixtral-8×7B, ChatGPT-4o 모델에 입력하여 질의응답을 수행하였다. 모든 실험에서 대규모 언어모델은 별도의 데이터 전처리나 후처리 과정 없이 구조화된 표 데이터를 직접 입력받았다. 동일한 질문 세트를 사용하여 각 포맷별 응답을 생성하고 성능을 비교하였다. 생성된 답변의 정확성은 문서에 정의된 표준 정답과의 비교를 통해 수동 평가하였으며, 각 포맷의 효율성 분석을 위해 토큰 사용량을 확인하고 비교하였다. <표 1>은 변환된 표 데이터를 바탕으로 소규모 언어모델이 답변을 생성할 때 사용하는 시스템 프롬프트를 나타낸다.

〈표 1〉 문서 내 테이블 정보 인식을 위한 시스템 프롬프트

Task	System Prompt
Table	You are a helpful AI assistant. Follow these rules exactly: 1. Please review the HTML OR MARKDOWN document and provide an accurate response. 2. Please keep the "–" symbol displayed exactly as "–."

4.3.2 계층 기반 문서 파싱과 고정 크기 청킹 파싱 비교 실험

계층 기반 파싱과 랜덤 청킹 방식의 문서 처리 성능을 세 가지 대규모 언어 모델(Mistral-7B, Mixtral-8×7B, ChatGPT-4o)을 사용하여 비교 분석하였다. 실험의 신뢰성을 위해 모든 모델에 동일한 하이퍼파라미터를 적용하였으며, 그 세부 사항은 <표 2>에 제시되어 있다. 사전에 준비된 표준 질의셋을 모든 실험에 일관되게 적용하여 공정한 비교를 수행하였다. 또한 <표 3>은 소규모 언어모델이 답변을 생성할 때 사용하는 시스템 프롬프트를 나타낸다.

〈표 2〉 대규모 언어모델 생성 하이퍼파라미터

Hyper Parameter	Value	Description
temperature	0.1	Controls randomness; lower values yield more deterministic output
max_new_tokens	2000	Limits the number of tokens generated in the output
top_p	0.3	Nucleus sampling; restricts choices to top cumulative probability
do_sample	True	Enables sampling-based decoding instead of greedy

〈표 3〉 문서 내 정보검색을 위한 시스템 프롬프트

Task	System Prompt
Contents	You are a helpful AI assistant. Follow these strict rules: <Strong Rules> 1. DO NOT add any explanations or interpretations. 2. DO NOT modify or rephrase the original text. </Strong Rules>

4.4 실험 결과 및 성능 비교

4.4.1 표 형식 변환 기반 질의응답 결과

상용 대규모 언어모델 모델인 ChatGPT-4o와 Claude Sonnet4.0, Gemini 2.5 Pro 모델을 이용하여 병합된 셀이 포함된 복잡한 표를 HTML 및 Markdown 형태로 변환한 후 Mistral-7B와 Mixtral-8×7B 그리고 ChatGPT-4o 모델에 입력하여 질의응답을

수행한 결과, HTML로 변환된 포맷과 Markdown 포맷으로 변환한 경우 모두 정확한 답변을 도출하였다. 반면, 고정 크기 청킹 기법이나 Python 프로그램을 이용한 직접 변환 방법을 적용한 경우에는 병합 셀 정보가 분절되어 표의 구조적 일관성이 손실되었다. 이로 인해 동일한 언어 모델들(Mistral-7B, Mixtral-8×7B, ChatGPT-4o)에서 정확한 답변을 얻지 못하는 결과가 나타났다.

〈표 4〉 모델별 API 가격 및 평균 응답 시간 비교

Model	Input cost (per 1K tokens)	Output cost (per 1K tokens)	Average response latency
GPT-4o	$0.005	$0.015	~0.32 sec
Claude 4.0 Sonnet	$0.003	$0.015	~0.28 sec
Gemini 2.5 Pro	$0.00125–$0.00250	$0.010–$0.015	~a few to several dozen seconds

<표 4>에서와 같이 상용 대규모 언어모델들은 입력 토큰과 출력 토큰에 대해 차별화된 가격 체계를 적용하고 있다. GPT-4o의 경우 입력 토큰 1K당 $0.005, 출력 토큰 1K당 $0.015로 책정되어 있으며, Claude 4.0 Sonnet은 입력 토큰 1K당 $0.003, 출력 토큰 1K당 $0.015의 비용 구조를 보인다. 특히 주목할 점은 Gemini Pro가 입력 토큰에 대해 컨텍스트의 크기에 따라 $0.00125-$0.0025의 가변적 가격 정책을 채택하고 있다는 것이다.

또한 토큰 분석 결과, Markdown 포맷은 HTML에 비해 평균적으로 약 50% 적은 토큰 수를 사용한다. 이는 50%의 비용 절감 효과를 의미하며, 긴 표나 복잡한 데이터 구조를 다룰 때 소규모 언어모델의 컨텍스트 윈도우 초과 문제를 방지하는 데 유리한 특성을 가진다.

4.4.2 계층 기반 문서 파싱과 고정 크기 청킹 파싱 비교 결과

본 연구에서는 두 가지 주요 파싱 방식에 대해 다양한 모델의 성능을 평가하였다. BERTScore 점수를 기준으로 한 정량적 평가 결과는 다음과 같다.

<표 5>에서 확인할 수 있듯이, 문서 파싱 전략에 따라 소규모 언어모델의 성능 차이를 보였다. 계층 기반 문서 파싱은 모든 모델에서 Fixed-size Chunking 방식보다 높은 BERTScore를 기록했다. 동일한 파싱 방식에서 Mixtral-8×7B가 Mistral-7B보다 항상 더 높은 점수를 기록하여 모델 크기 및 구조(MoE 아키텍처)의 효과를 입증했다.

가장 높은 성능은 ChatGPT-4o와 계층 기반 문서 파싱 조합으로 Precision 0.923, Recall 0.924, F1 0.920의 지표를 기록하여 최고 수준의 응답 일치율을 보였다.

이러한 실험 결과는 문서의 구조적 특성을 반영한 계층 기반 파싱 방법론이 복잡한 문서 이해 과제에서 상당한 효과를 발휘함을 실증적으로 입증한다. 계층 기반 파싱은 문서의 논리적 구조와 의미적 맥락을 체계적으로 활용함으로써 소규모 언어모델의 문서 이해 능력을 향상시키는 것으로 분석된다. 특히 주목할 점은 실험에 사용된 모든 모델에서 일관된 성능 개선이 관찰되었다는 것으로, 이는 파싱 전략이 모델 성능에 미치는 영향의 중요성을 강조하는 결과로 해석할 수 있다.

한편, 모델별 특성에 따른 세부적인 차이점도 관찰되었다. Mixtral-8×7B의 경우 시스템 프롬프트를 사용하지 않은 상황에서는 문서의 내용을 직접 인용하는 경향을 보였으나, 줄임말과 같은 표현을 서술형으로 변환하거나 다른 표현으로 재구성하는 경우가 발견되었다. 반면 Mistral-7B는 긴 문서의 경우 시스템 프롬프트를 적용하더라도

내용을 요약하는 경향을 보였다. 이러한 모델별 행동 패턴의 차이는 각 모델의 고유한 처리 방식과 용량 한계가 파싱 결과에 미치는 영향을 시사한다.

〈표 5〉 파싱 방식 및 소규모 언어모델에 따른 BERTScore 비교

Parsing Type	Model	BERTScore		
		Precision	Recall	F1
Fixed-size Chunking Parsing	Mistral-7B	0.847	0.868	0.855
	Mixtral-8×7B	0.852	0.868	0.859
Hierarchy-based Document Parsing	Mistral-7B	0.871	0.881	0.874
	Mixtral-8×7B	0.890	0.893	0.891
	ChatGPT-4o	0.923	0.924	0.920

5. 결론

본 연구는 대규모 언어 모델의 복잡한 문서 이해 성능을 향상시키기 위한 계층 기반 문서 파싱 방법론의 효과를 실증적으로 검증하였다. 실험 결과, ChatGPT-4o와 계층 기반 문서 파싱을 결합한 접근법이 Precision 0.923, Recall 0.924, F1 0.920의 최고 성능을 달성하였으며, 특히 주목할 점은 Mixtral-8×7B 모델이 ChatGPT-4o와 유사한 수준의 우수한 성능을 보여주었다는 것이다.

본 연구는 세 가지 주요 기여를 제시한다. 첫째, 문서의 구조적 특성을 체계적으로 반영한 계층 기반 파싱 방법론을 제안하고, 이것이 복잡한 문서 이해 과제에서 효과적임을 실증적으로 검증하였다. 둘째, 문서의 논리적 구조와 의미적 맥락을 통합적으로 활용하여 대규모 언어 모델의 문서 이해 성능을 향상시킬 수 있음을 확인하였다. 셋째, 다양한 모델에서 일관된 성능 개선을 관찰함으로써 제안된 파싱 전략의 범용성과 모델 성능에 대한 긍정적 영향을 규명하였다. 특히 Mixtral-8×7B와 같은 오픈소스 모델도 적절한 파싱 전략과 결합될 때 상용 모델과 경쟁할 수 있는 수준의 성능을 달성할 수 있음을 확인하였다. 실제 시스템 구현 과정에서는 병합된 셀이 포함된 복잡한 표 구조의 경우, Claude의 Sonnet과 같은 특정 상용 대규모 언어모델을 활용해 Markdown 형식으로 변환하여 문서를 관리하는 방식이 HTML 형식에 비해 토큰 수를 줄여 비용을 절감하면서도, 구조적 정보 손실을 최소화하고 일관된 데이터 처리를 가능하게 함을 확인하였다. 또한 주목할 만한 성과는 상용 대규모 언어모델의 높은 운영 비용 문제에 대한 효과적인 해결책을 제시한 점으로, 소규모 언어모델을 활용한 문서 질의응답 시스템이 상용 대규모 언어모델과 거의 동등한 수준의 성능을 달성할 수 있음을 본 연구를 통해 입증하였다. 도메인별 특화 학습을 통해 최적화된 소규모 언어모델 기반 시스템은 비용 효율성과 성능을 동시에 만족시키는 혁신적인 대안으로 평가되었으며, 이는 실용적인 문서 질의응답 시스템 구축에 있어 중요한 돌파구를 제공한다. 또한 모호한 질문으로 다수의 관련 내용이 검색되는 경우, 질문자에게 구체적인 선택 옵션을 제공하여 정확한 답변을 도출할 수 있는 인터랙션 메커니즘의 필요성을 확인하였으며, 이는 사용자 만족도와 시스템 실용성 향상에 핵심적인 요소로 작용하였다. 다만 문서마다 상이한 형식으로 인해 일관된 파싱 알고리즘을 적용하기 어려우며, 각 문서의 특성에 따라 복잡한 커스터마이징이 필요하다는 한계점이 발견되었다.

그리고 본 연구의 한계점을 보완하고 보다 강건한 시스템을 구축하기 위해 다음과 같은 추가 연구가 필요하다. 우선, 서로 다른 교육기관의 문서나 PubMedQA와 같은 공개 데이터셋을 활용한 교차 검증을 통해 시스템의 일반화 성능을 평가할 필요가 있다. 이를 통해 특정 도메인에 국한되지 않는 범용적 적용 가능성을 확인할 수 있을 것이다. 둘째, 현재의 평가 체계를 보완하기 위해 EM(Exact Match), F1-score 등의 정량적 지표를 추가로 도입하여 시스템 성능을 다각도로 분석할 필요가 있다. 아울러 질문 유형별, 난이도별로 세분화된 분석을 수행하고, 정답 산출 과정의 추론 경로에 대한 정성적 평가 또한 병행되어야 한다. 셋째, 실제 사용 환경에서의 효용성을 검증하기 위해 Task completion time, SUS(System Usability Scale) 점수 등을 포함한 체계적인 사용자 연구 설계가 필요하다. 이를 통해 시스템의 실용성과 사용자 만족도를 종합적으로 평가할 수 있다. 마지막으로 Dense와 Sparse 검색 방식을 결합한 하이브리드 검색 전략을 탐색하고, BM25와 같은 전통적인 검색 기법을 기준선으로 포함하여 다양한 검색 기법의 성능을 비교 분석하는 연구가 요구된다. 이는 문서 검색의 최적의 검색 전략 수립에 중요한 기여를 할 것으로 기대된다.

참고문헌(References)

우민식, 시종욱, 김성영. (2025). 검색 증강 생성 기법과 소규모 대형 언어 모델을 활용한 문서 요약 기반 챗봇 시스템. *한국정보기술학회 논문지*, 23(1), 13-20. http://doi.org/10.14801/jkiit.2025.23.1.13

윤요섭, 안현철. (2024). RAPTOR를 활용한 LLM 기반 한국어 뉴스 질의응답 시스템 개발. *지능정보연구*, 30(4), 205-222.

윤여찬, 김수균. (2025). 생성형 AI를 위한 Retrieval-Augmented Generation (RAG) 기술 동향 및 전망. *컴퓨터교육학회 논문지*, 28(2), 69-80. http://doi.org/10.32431/kace.2025.28.2.007

정천수. (2023). LLM 애플리케이션 아키텍처를 활용한 생성형 AI 서비스 구현: RAG 모델과 LangChain 프레임워크 기반. *지능정보연구*, 29(4), 129-164.

정천수. (2024). 도메인 특화 LLM: Mistral 7B를 활용한 금융 업무분야 파인튜닝 및 활용 방법. *지능정보연구*, 30(1), 93-120.

Kim, H. B., Y. G. Yu, (2023). “Development of a Regulatory Q&A System for KAERI Utilizing Document Search Algorithms and Large Language Model”. *Journal of Korea Society of Industrial Information Systems*, Vol. 28, No. 5 (2023), 31 – 39.

Lee, J. H., Cha, H. S., Hwangbo, Y., & Cheon, W. J. (2024). Enhancing Large Language Model Reliability: Minimizing Hallucinations with Dual Retrieval-Augmented Generation Based on the Latest Diabetes Guidelines. *Journal of Personalized Medicine*, 14(12), 1131. http://doi.org/10.3390/jpm14121131

Lewis, Patrick., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., 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://doi.org/10.48550/arXiv.2005.11401

Maples, S. (2017). The ROUGE AR: A Proposed Extension to the ROUGE Evaluation Metric

for Abstractive Text Summarization. Symbolic Systems Department, Stanford University. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2761938.pdf

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311-318. https://doi.org/10.3115/1073083.1073135

Xiong, G., Jin, Q., Lu, Z., & Zhang, A. (2024) Benchmarking retrieval-augmented generation for medicine, *Findings of the Association for Computational Linguistics: ACL*, 6233–6251. https://doi.org/10.18653/v1/2024.findings-acl.372

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. arXiv. https://arxiv.org/abs/1904.09675

Abstract

Small Language Model – Based RAG for Document Question Answering in Education*

Hyun-Woo Lee** · Kyoung-jae Kim*** · Yung-Seop Lee****

Recent advances in large language models (LLMs) have significantly accelerated the adoption of natural language technologies in educational settings. Retrieval-Augmented Generation (RAG) further enhances the capabilities of LLMs by dynamically retrieving up-to-date or domain-specific knowledge from external sources, enabling more accurate and trustworthy responses. Nevertheless, the direct deployment of LLMs in educational institutions is often limited by high inference costs and privacy concerns. Moreover, institutional documents such as admission guidelines or academic regulations typically contain domain-specific terminology, deeply hierarchical structures, and highly formatted content including merged-cell tables and embedded images. Conventional LLM-based chatbots struggle with accurately extracting data from complex tabular layouts, effectively identifying relevant paragraphs within long, multi-level documents, and incurring substantial computational costs for each query. To address these limitations, this study proposes a framework that combines a Small Language Model (SLLM) with RAG. We introduce a section-based document-parsing strategy that systematically decomposes institutional documents and a table-aware encoding pipeline that allows the SLLM to handle intricate tabular data accurately. The proposed approach offers a cost-effective and privacy-conscious alternative to commercial LLM services while maintaining high response accuracy for document question-answering tasks in educational domains.

Key Words : SLLM, RAG, Chatbot, Document Question Answering

Received : July 18, 2025 Revised : September 17, 2025 Accepted : September 19, 2025

Corresponding Author : Yung-Seop Lee

* This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2021R1A2C1007095), and by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2025-RS-2020-II201789) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

** Department of Fintech and Blockchain, Major in AI and Big Data, Dongguk University
*** Department of MIS, Business School, Dongguk University_Soeul
**** Corresponding Author: Yung-Seop Lee
Department of Statistics, College of Science, Dongguk University
30, Pildong-ro 1-gil, Jung-gu, Seoul 04620, Korea
Tel: +82-2-2260-3218, E-mail: yung@dongguk.edu

저 자 소 개

이현우
동국대학교에서 경영학 석사학위를 취득했으며 현재는 AI·빅데이터전공 박사과정에 재학 중이다. 주 관심사는 딥러닝 및 머신러닝 기법을 활용한 예측 모델링이다. 최근에는 Retrieval-Augmented Generation(RAG) 기반 LLM의 성능 최적화 및 실제 교육·비즈니스 분야 적용 연구를 활발히 진행하고 있다.

김경재
현재 동국대학교 경영대학 경영정보학과 교수로 재직 중이다. KAIST에서 경영정보시스템을 전공으로 박사학위를 취득하였으며, 연구 관심분야는 비즈니스 애널리틱스, CRM, 추천기술, 빅데이터 분석 등이다.

이영섭
연세대학교에서 응용통계학과학사, Iowa주립대학교에서 통계학전공으로 석사학위를, Rutgers대학에서 통계학전공으로 박사학위를 취득하였으며, 현재 동국대학교 통계학과 교수로 재직하고 있다. 주 관심사는 응용통계학, 빅데이터분석, 통계적 인공지능 연구이다.