JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 5, pp. 67-74, May 2026
https://doi.org/10.9708/jksci.2026.31.05.067

A Study on a Retrieval-Augmented Generation-Based for University Information

Yong-Min Cho*, Sung-Jin Kim**

*Student, School of AI·BigData, MyongJi College, Seoul, Korea
**Professor, Dept. of AI·BigData, MyongJi College, Seoul, Korea

[Abstract]

The decentralized information structure of existing university websites often hinders the information retrieval efficiency of Generation Z, who are accustomed to conversational search. To address this limitation, this study proposes a Retrieval-Augmented Generation (RAG)-based academic chatbot system designed to efficiently provide decentralized academic information on university websites, reflecting the preferences of Generation Z for natural language-based conversational search. The research methodology involved collecting scattered academic data and converting it into structured text, then applying the Reciprocal Rank Fusion (RRF) algorithm to a hybrid search method that combines keyword and semantic searches to enhance retrieval accuracy. Additionally, context-aware querying and multilingual support functions were integrated into the system design. Through experimentation, optimal search conditions were derived to verify stable response generation capabilities, and a control layer using system prompts was established to effectively mitigate hallucinations in the generative model. In conclusion, this system is expected to improve information accessibility for students and alleviate the burden of repetitive administrative tasks for faculty and staff , thereby contributing to the overall quality improvement and operational efficiency of university administrative services.

▸ **Key words**: Bachelor's chatbot, Z-Generation, Hybrid Retrieval, Reciprocal Rank Fusion, Retrieval-Augmented Generation

[요 약]

기존 대학 홈페이지의 분산된 정보 구조는 대화형 검색에 익숙한 Z세대의 정보 탐색 효율성을 저하시키는 한계가 있다. 이에 본 연구는 대학 홈페이지 내 분산된 학사 정보를 효율적으로 제공하기 위해, 자연어 기반 대화형 검색을 선호하는 Z세대의 특성을 반영한 RAG 기반 학사 챗봇 시스템을 제안한다. 연구 방법으로는 분산된 학사 데이터를 수집하여 구조화된 텍스트로 변환하고, 키워드 검색과 의미 검색을 결합한 하이브리드 검색 방식에 RRF 알고리즘을 적용하여 검색의 정확도를 높였으며, 맥락 기반 질의 및 다국어 지원 기능을 통합 설계하였다. 실험을 통해 최적의 검색 조건을 도출하여 안정적인 응답 생성 능력을 검증하였고, 시스템 프롬프트를 통한 제어 계층을 구축하여 생성형 모델의 환각 현상을 효과적으로 제어하였다. 결론적으로 본 시스템은 학생의 정보 접근 편의성을 증진하고 교직원의 반복적인 행정 업무 부담을 경감함으로써 대학 행정 서비스의 전반적인 질적 향상과 운영 효율화에 기여할 것으로 기대된다.

▸ **주제어**: 학사 챗봇, Z세대, 하이브리드 검색, 상호 순위 융합, 검색 증강 생성

• First Author: Yong-Min Cho, Corresponding Author: Sung-Jin Kim
*Yong-Min Cho (whdydals0802@mjc.ac.kr), School of AI·BigData, MyongJi College
**Sung-Jin Kim (ict214548@mjc.ac.kr), Dept. of AI·BigData, MyongJi College
• Received: 2026. 04. 10, Revised: 2026. 04. 30, Accepted: 2026. 05. 10.

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

I. Introduction

최근 대학 홈페이지는 학사 정보, 편의시설, 장학금 등 다양하고 방대한 정보를 제공하고 있다. 그러나 정보는 홈페이지 내 여러 경로에 흩어져 있다. 사용자가 필요한 정보를 획득하기 위해 사이트의 계층 구조를 파악하고 탐색해야 하는 인지적 절차를 거쳐야 하는 문제가 존재한다. 실제로 세대별 검색 방식의 변화를 분석한 연구에 따르면, 현재 대학 생활에 주류가 되는 Z세대는 키워드 나열이나 계층 구조 기반의 검색 방식보다 자연어 문장으로 질문하여 검색하는 '대화형 검색(Conversational Search)'을 선호하는 경향을 나타낸다. Z세대는 약 43.5% 이상이 검색 시 질문 형태의 긴 문장을 사용하며, 이는 다른 세대와 비교했을 때, 높은 편차를 보인다. 이를 통해 전통적인 정보 접근 방식은 즉각적인 답변에 익숙해진 디지털 네이티브 세대, 특히 Z세대에게는 정보 탐색의 효율성을 낮추는 주요 원인으로 작용한다[1].

이러한 문제점은 학생뿐만 아니라 대학 교직원의 업무 효율성을 낮추는 요인이 된다. 교직원의 업무는 상당 부분 매 학기 단순·반복적인 학사 문의(장학금 일정, 수강 신청 방법, 서류 발급 방법 등)를 대응하는 데 소모되고 있다. 이는 교직원의 업무적 피로도를 높이고 행정 자원의 낭비를 일으킨다. 연구에 따르면 AI 기반의 자동화된 응대 시스템은 조직 내 반복적인 업무를 효율적으로 처리하여 업무의 질적 변화를 일으킬 수 있다[2]. AI 챗봇이 단순 질의응답 업무를 처리하고 교직원은 학생 상담, 교육 프로그램 기획, 학사 정책 수립과 같은 고부가가치 업무에 역량을 집중할 수 있는 환경을 마련할 수 있다[3].

본 연구에서는 대학 홈페이지 내 여러 경로에 분산되어 있는 학사 정보를 수집하여 저장하는 벡터 데이터베이스(Vector Database)를 설계하고, 이를 참조하여 답변하는 검색 증강 생성(Retrieval-Augmented Generation, RAG) 기반의 학사 챗봇 시스템을 연구했다. 사용자의 질문을, 관련 학사 정보를 벡터 검색을 통해 찾고, 이를 바탕으로 답변을 생성함으로써 기존의 계층 구조를 직접 파고들어야 했던 정보 탐색 방식의 한계를 보완한다.

결과적으로 본 연구는 질문 기반의 정보 탐색에 익숙한 디지털 네이티브 세대의 학사 정보 접근성을 높이고, 반복적인 행정 문의를 챗봇을 통해 자동화하여 보다 주요한 고부가가치의 업무에 집중할 수 있는 환경을 마련한다.

본 논문은 총 4장으로 구성되며, 각 장의 내용은 다음과 같다.

제Ⅱ장에서는 국내 대학의 챗봇 시스템 도입 현황과 사례를 살펴보고 이를 분석한다. 또한 기존 룰 기반 챗봇과 LLM 기반 연구의 동향 및 한계점을 검토하고, 환각 현상을 방지하면서 정확한 학사 정보를 제공하기 위한 RAG 기반 학사 챗봇을 본 연구의 방향으로 제시한다.

제Ⅲ장에서는 RAG 기반 학사 챗봇의 데이터 파이프라인과 시스템 아키텍처 설계 방안을 상세히 설명한다. 데이터 전처리 과정뿐만 아니라, 하이브리드 검색, 문맥 인식 기반 쿼리 처리, 다국어 지원, 토큰 사용량 관리 등 최적의 응답 생성을 위한 세부 모듈들을 구체적으로 다룬다.

마지막으로 제Ⅳ장에서는 연구 내용을 종합하여 결론을 도출하고, 제안하는 RAG 기반 학사 챗봇 시스템이 학사 행정 및 사용자 편의성에 미치는 기대 효과를 설명한다. 또한 본 연구의 한계점을 제시하고, 이를 보완하기 위한 향후 연구 방향을 제안한다.

II. Preliminaries

1. Related research and trends

대학 학사 행정은 챗봇 시스템을 통해 행정 업무 효율화와 학생 및 교직원의 정보 접근성 향상을 목적으로 점진적으로 도입되었다. 초기에는 정형화된 학사 정보 제공을 중심으로 한 스크립트 기반 또는 룰 기반의 챗봇이 주를 이뤘으며, 최근에는 대규모 언어 모델(Large Language Models, LLM)의 발전에 따라 자연어 이해 및 생성 능력을 활용한 AI 챗봇으로 발전하는 추세를 보인다.

1.1 Script/Rule-based Academic Chatbot

국내 다수 대학에서는 학사 일정, 수강 신청, 성적, 졸업 요건 등 반복적인 질의에 대응하기 위해 스크립트 기반과 FAQ 중심의 챗봇을 도입하였다. 기존 사례로는 성균관대 'KINGO봇'과 명지대 '마루봇'이 있다[4][5]. 이러한 시스템은 사전에 정의된 질문-응답의 쌍으로 이루어지거나 키워드 매칭 규칙에 기반을 두어 안정적인 정보 제공이 가능하다는 장점을 지닌다. 그러나 입력된 질의가 사전에 정의된 범위를 벗어나거나, 줄임말이 포함될 경우 대응이 어렵다는 한계를 지닌다.

1.2 LLM-based Academic Chatbot

최근 생성형 인공지능의 발전과 함께, 일부 연구에서는 LLM을 활용하여 자연어 질의에 유연하게 답변하는 학사 챗봇을 시도하고 있다[6]. 이러한 접근은 기존 스크립트 기

반 시스템과는 다르게 사용자의 질문 의도를 이해하고 자연스러운 문장 형태의 응답을 생성할 수 있다는 장점을 가진다. 하지만 학사 행정 데이터를 학습하는 방식은 매년 달라지는 정책과 일정을 학습해야 한다는 문제와 실제 존재하지 않는 규정이나 부정확한 정보를 생성하는 '환각(Hallucinations)' 문제가 발생할 가능성도 존재한다[7].

환각 현상은 모델이 학습 데이터에 존재하지 않는 사실을 임의로 생성하거나, 서로 다른 정보를 논리적으로 결합하여 왜곡된 결과를 생성하는 것을 의미한다. 이러한 환각은 모델이 제공된 맥락에 반하는 정보를 생성하는 '내재적 환각(Intrinsic Hallucinations)'과 모델이 학습 데이터에 존재하지 않는 내용을 임의로 조작하거나 사실 관계를 왜곡하여 답변하는 '외재적 환각(Extrinsic Hallucination)'으로 구분된다[8]. 이러한 현상의 주요 원인은 크게 두 가지로 분석된다.

첫째, LLM의 '확률적 생성'이다. 모델은 사실 관계의 정확성보다는 문장의 확률적 완결성을 우선시하기 때문에, 정보가 불확실한 상황에서도 그럴듯한 답변을 생성하려는 경향을 보인다[9].

둘째, '지식의 단절' 문제이다. 모델은 사전 학습이 완료된 시점까지의 데이터만을 보유하고 있으며, 학습 이후에 발생한 최신 정보나 실시간으로 변동되는 데이터에 대해서 환각을 일으킬 가능성이 높다[10].

결과적으로 환각 현상은 사용자가 시스템으로부터 얻은 정보의 신뢰도를 저하시키는 결정적 요인이 된다. 특히 높은 정확성이 요구되는 정보 탐색 과정에서 발생하는 환각은 디지털 네이티브 세대가 기대하는 '즉각적이고 정확한 답변'이라는 가치를 훼손하며, 오히려 잘못된 정보 수정에 추가적인 인지적 비용을 발생시키는 역효과를 낳는다. 따라서 LLM의 높은 언어 이해 능력을 활용하되, 생성되는 정보의 사실적 근거를 확보하여 환각을 제어할 수 있는 기술적 장치가 요구된다.

1.3 RAG-based Academic Chatbot

LLM의 높은 언어 이해 능력을 활용하고, 사실 기반의 답변을 생성하는 기술적 장치로 Retrieval-Augmented Generation(RAG)이 주목받고 있다. Fig 1과 같이 RAG는 사용자의 질의에 대해 LLM이 응답을 생성하기에 앞서, 관련 정보를 검색(retrieval)하고, 해당 결과를 기반으로 응답을 생성하는 방식이다. 최근 국내 대학은 내부 문서를 기반으로 한 RAG 구조의 챗봇 시스템에 관한 연구 및 적용 사례가 점차 증가하고 있다[11][12].

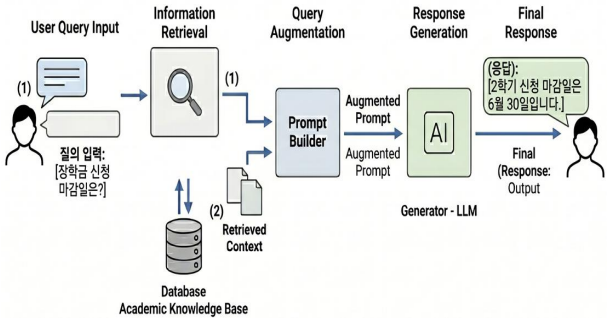

Fig. 1. Academic RAG Chatbot flow

이러한 접근은 LLM의 자연어 이해 및 생성 능력을 유지하면서도, 응답의 근거를 실제 학사 문서에 두어 정보의 정확성과 최신성을 확보할 수 있다는 점에서 학사 행정 도메인에 적합한 구조라고 평가된다.

2. A Study on the Direction

앞의 설명과 같이 국내 대학의 학사 챗봇은 점진적으로 발전해 왔다. 그러나 학사 행정 도메인의 특성상, 정보의 정확성과 최신성이 요구되며, 다수의 규정과 문서가 분산되어 존재한다는 점에서 LLM 단독 학습 기반 접근에는 구조적 한계가 존재한다. 이에 따라 국내 여러 대학에서는 RAG 기반 학사 챗봇 구조를 도입한 사례가 점차 증가하고 있다. 이러한 접근은 학사 행정과 관련된 실제 운영 문서를 검색 기반으로 참조하여, LLM 단독 학습 방식의 한계를 보완한다. 그러나 단순히 RAG 기반의 구조를 도입한다고 해서 문제가 해결되지 않는다. 학사 행정 문서는 서술형 텍스트뿐만 아니라 표, 목록, 혼합 구조 문단 등 정형-비정형 요소가 복합적으로 포함되어 있기 때문이다. 그리하여 본 연구에서는 복합적으로 포함된 학사 행정 문서를 정확하게 추출하고 이를 기반으로 답변하는 RAG 기반의 학사 챗봇 시스템을 제안한다.

III. The Proposed Scheme

1. Data ingestion pipeline

본 논문에서는 명지전문대학에 적용될 학사 챗봇 시스템을 Fig 2와 같은 흐름으로 설계하였다. 학교 홈페이지와 학칙 및 학사 내규 문서를 수집하고, 참조하여 답변하는 RAG 구조의 학사 챗봇 시스템을 아래와 같이 연구하였다.

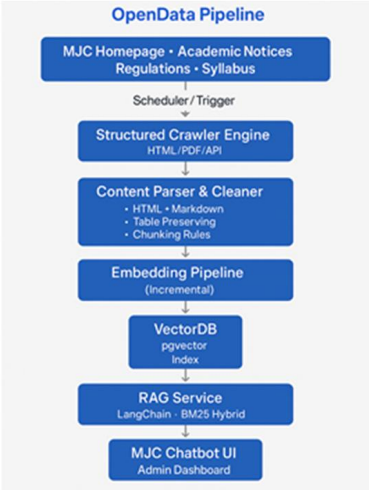

Fig. 2. data-pipeline flow

1.1 Data Collection

본 연구에서는 데이터를 두 가지 방식으로 수집하였다. 첫째, 대학에서 제공하는 PDF 및 HWP 형식의 학칙 및 학사 내규 문서이다. 둘째, 학사 홈페이지에서 제공되는 학과 소개, 학사 일정, 공지 사항 등 교내 웹 페이지에 분산된 데이터이다.

대학 학과 소개 및 학사 일정, 공지 사항 등은 학과 홈페이지, 생활 서비스 등 다양한 소스에 분산되어 있다. 이러한 웹 페이지들은 서로 다른 DOM(Document Object Model) 구조로 되어 있어 정형화된 크롤러(Crawler)를 적용하기 어렵고, 실제 정보와 무관한 내비게이션 바, 이미지 등 불필요한 노이즈 데이터가 다수 포함되어 있어 검색 성능을 저하하는 요인이 된다. 이에 본 연구에서는 구글 크롬 확장 프로그램인 PrintFriendly를 활용하여 데이터를 수집하였다. PrintFriendly는 웹 페이지의 불필요한 이미지와 UI 요소를 제거하고 텍스트 위주의 핵심 콘텐츠만을 추출하여 정제된 PDF 파일로 변환하는 기능을 제공한다. 이를 통해 복잡한 웹 구조로 인한 크롤링 누락 문제를 방지하고, LLM이 처리하기 용이한 고품질의 텍스트 데이터를 확보하였다.

1.2 Data Preprocessing and Conversion

수집된 PDF 데이터를 벡터 데이터베이스에 임베딩하기 위해서는 LLM이 이해할 수 있는 텍스트 포맷으로 변환하는 과정이 필요하다. 기존의 단순 텍스트 추출 방식은 문서의 구조적 정보(헤더, 표, 리스트 등)를 소실할 위험이 있다. 따라서 본 논문에서는 PyMuPDF4LLM 라이브러리를 사용하여 PDF 문서를 마크다운(Markdown) 형식으로 변환하였다[13].

문서를 페이지 단위의 청크(Chunk)로 분할함과 동시에, 문서의 계층적 구조를 마크다운 문법으로 보존하였다. 이렇게 변환된 각 페이지 데이터는 Lang Chain 프레임워크의 Document 객체로 매핑된다. 이러한 과정은 문서의 내용뿐만 아니라 페이지 번호, 소스 파일명 등의 메타데이터를 함께 관리할 수 있게 하여, 추후 검색 결과의 출처를 명시하는 데 활용된다.

1.3 Text Splitting

전처리 과정을 거쳐 Document 객체로 변환된 데이터는 대규모 언어 모델(LLM)의 컨텍스트 윈도우(Context Window) 제한을 고려하고, 검색의 정밀도를 높이기 위해 적절한 크기로 분할되어야 한다. 문맥의 의미적 연결성을 유지하면서도 검색 효율을 극대화하기 위해 재귀적 문자 분할기(RecursiveCharacterTextSplitter)를 채택하였다.

재귀적 분할 방식은 문단, 문장, 단어 순으로 구분자를 시도하며 텍스트를 분할하므로, 문서의 구조적 단위를 최대한 보존할 수 있다는 장점이 있다[14]. 본 연구에서 설정한 세부 파라미터는 다음과 같다.

Chunk Size(300): 개별 텍스트 조각의 최대 길이를 300자로 제한하였다. 이는 학사 규정이나 공지 사항의 개별 항목이 너무 길어져 정보의 손실이 발생하는 것을 방지하기 위함이다.

Chunk Overlap(100): 분할된 청크 간에 100자의 중첩 영역을 설정하였다. 이는 텍스트 분할 지점에서 발생할 수 있는 의미 단절을 최소화하고, 인접한 청크 간의 문맥적 연속성을 확보하기 위한 장치이다.

1.4 Embedding and Vectorization

분할된 텍스트 청크를 벡터 공간으로 투영하여 의미론적 유사도 검색이 가능하도록 임베딩 과정을 수행한다. 본 시스템은 한국어의 언어적 특성과 학사 용어의 복잡성을 고려하여, 한국어 성능이 검증된 nlpai-lab/KURE-v1 임베딩 모델로 선정하였다[15]. 수집된 모든 텍스트 청크는 LangChain의 임베딩 파이프라인을 통해 고차원 벡터 공간의 점은 아래 수식 (1)과 같이 변환된다.

$$v_i = f_{embed}(t_i) \in R^d \quad (1)$$

생성된 벡터 데이터는 이후 검색을 위해 벡터 데이터베이스에 저장된다.

2. System Architecture

본 연구에서 제안하는 전체 시스템 아키텍처는 Fig 3과 같다.

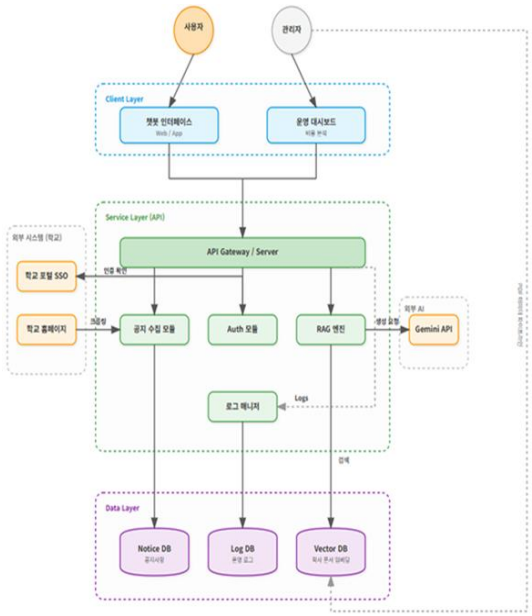

Fig. 3. overall system architecture

위의 Fig 3은 본 연구에서 제안하는 전체 시스템은 LLM의 환각 현상을 최소화하면서 학사 도메인에 특화된 정보를 제공하기 위해 RAG 기반으로 설계되었다. 본 시스템은 크게 Client Layer와 Service Layer 및 Data Layer로 구성하였다.

Client Layer는 대상에 따라 사용자(학생 및 교직원)와 관리자로 이원화된 인터페이스를 제공하고 사용자와 시스템 간의 접점을 제공한다. 사용자는 '챗봇 인터페이스'를 통해 질문하고 관리자는 '운영 대시보드'를 통해 챗봇 사용량을 모니터링하여 LLM API 사용에 따른 비용을 분석한다.

Service Layer는 시스템의 중추 역할을 하는 계층으로, 시스템의 제어 및 데이터 처리 파이프라인을 담당한다. 해당 레이어의 핵심 모듈은 'RAG Engine'으로 사용자의 질의를 벡터화하여 저장된 벡터 DB 안에 가장 유사한 문서를 검색하고, 검색된 문맥과 질의를 결합한 프롬프트를 'Gemini API'에 전송하여 문서 기반의 최종 답변을 생성한다.

Data Layer는 데이터 성격에 따라 학사와 관련된 비정형 문서를 데이터 파이프라인을 통해 저장한 벡터 DB와 공지 사항, 로그 등을 저장한 관계형 DB로 나눠진다. 이렇게 설계된 시스템 아키텍처를 바탕으로 RAG 기반의 학사 챗봇을 구현하였다.

2.1 Hybrid Retrieval

본 연구에서는 학사 행정 문서의 특성을 고려하여 키워드(희소) 검색과 의미 유사도(밀집) 검색을 결합한 하이브리드 검색(Hybrid Retrieval)을 채택하였다. 키워드 검색은 정확한 키워드 매칭에 강점을 가지나, 사용자의 표현이 문서 용어와 일치하지 않는 경우 의미적 대응에 한계가 존재한다. 반면 의미 유사도 검색은 문맥과 의미적 유사성을 반영하여 용어가 일치하지 않아도 의미가 같다면 응답이 가능하다. 그러나 정확한 용어가 요구되는 질의에서는 응답률 저하가 발생할 수 있다[16].

본 연구에서는 학교 홈페이지 내에 흩어져 있는 각 정보가 같은 의미임에도 다른 단어를 쓰거나, 정확한 용어를 요구하는 정보들이 혼합된 특성을 가지므로, 하이브리드 검색 구조를 적용하였다.

2.2 Langchain Ensemble Retriever

본 절에서는 2.1절에서 설명한 하이브리드 구조를 적용하기 위해 LangChain의 Ensemble Retriever를 활용하였다. Ensemble Retriever는 서로 다른 특성을 갖는 복수의 Retriever를 병렬적으로 수행한 뒤, 각 Retriever가 산출한 검색 결과를 Reciprocal Rank Fusion(RRF) 알고리즘을 통해 병합하는 구조를 가진다[17].

RRF는 각 개별 검색 알고리즘이 산출하는 점수(Score) 대신, 결과 리스트에서의 상대적인 순위(Rank)만을 이용해 최종 점수를 계산하는 방식이다. 이는 서로 다른 Retriever가 산출하는 평가 척도를 별도의 정규화 과정 없이도 병합할 수 있게 된다. 이에 활용되는 RRF는 수식 (2)와 같다.

$$score(d) = \sum_{r \epsilon R} \frac{1}{k + r(d)} \qquad (2)$$

여기서 R은 검색 결과 병합에 참여하는 Retriever들의 집합을 의미하며, 키워드 기반의 검색 BM25 Retriever와 의미 유사도 기반의 검색 Dense Retriever가 이에 해당된다. $r(d)$는 이들 각 Retriever가 특정 문서 d에 대해 독립적으로 부여한 상대적인 순위 정보로, 이를 활용함으로써 서로 다른 평가 척도를 가진 검색 모델 간의 결과를

별도의 정규화 과정 없이 병합할 수 있다. 상수 k는 순위 권 하위에 위치한 문서가 전체 점수 합산에 끼치는 영향력을 조절하는 평활화 인자 역할을 수행한다. 이러한 연산 과정을 거쳐 최종 검색 순위를 결정한다.

2.3 Parent document

본 연구에서는 검색된 문서 중 상위 3개(top_k=3)만을 생성 모델의 입력 컨텍스트로 사용하였다. 이는 제한된 컨텍스트 윈도우 내에서 질문과 가장 관련성이 높은 정보를 기반으로 답변을 생성하기 위함이다. 참조 문서 수가 과도할 경우, 질문과 관련이 없는 정보가 함께 포함되어 응답 정확도가 저하될 수 있다. 반대로 참조 문서 수가 적을 경우에는 질문에 필요한 정보가 누락되어 잘못된 답변이 생성될 가능성이 존재한다.

Table 1. Top_k Experiment

top_k	response rate	note
1	0.66	Low information content, degraded quality
3	0.91	Stable performance
5	0.89	Stable performance
7	0.72	Decreased consistency in candidate documents

Table 1은 설정값의 객관성을 확보하기 위해 각 파라미터별로 동일한 질문을 반복 실행한 결과이다. 이를 통해 본 연구에서는 상위 3개의('top_k=3')의 문서가 가장 적합하다는 사실을 검증하였다.

2.4 Context-aware Query Processing

학사 행정 질문은 이전 질문의 맥락을 전제로 이어지는 경우가 존재한다. 예를 들어, "1학년이 휴학할 수 있어?"라고 질문한 후, "접수 기간은 언제야?"와 같은 질문은 앞선 대화의 맥락을 고려하지 않는다면 정확한 해석이 어렵다. 이러한 특징으로 인해 본 연구에서는 맥락 기반의 질의 처리 구조를 도입하였다. 본 시스템은 최근 대화 내용을 하나의 문서로 보고 답변을 생성하게 된다.

이러한 구조를 통해 Fig 4와 같이 이전 대화의 맥락을 전제로 이어지는 실제 학사 상담 환경에 가까운 대화를 지원할 수 있게 된다.

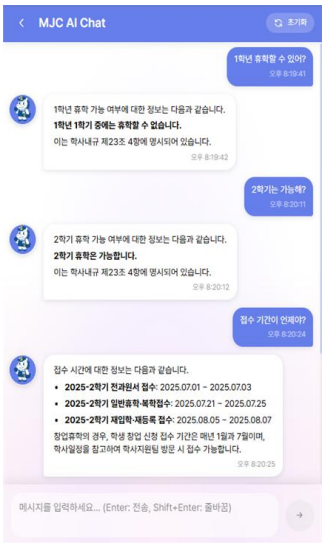

Fig. 4. Context-based QA

2.5 Multilingual Interaction Support

기존 학사 정보 시스템은 대부분 한국어 사용자 중심으로 설계되어 있다. 이는 외국인 학생의 정보 접근성을 제한하는 문제가 된다. 본 연구에서는 이러한 사용자 특성을 고려하여 언어 감지(lang_detect) 기반의 다국어 지원 기능을 설계하였다[18].

본 연구는 입력된 질의를 단순히 번역하여 처리하는 방식이 아닌, 먼저 언어 감지를 통해 사용자의 질의 언어를 식별하고, LLM 모델을 활용하여 질의의 의미를 기존 한국어 기반 학사 행정 문서 검색에 적합한 형태로 재구성한다. 이후 검색은 한국어로 임베딩 된 벡터 데이터베이스에서 수행되며, 응답 생성 단계에서만 사용자 언어에 맞게 자연어 출력을 생성한다. 언어 감지를 통한 다국어 답변에 대한 아키텍처는 아래 Fig 5와 같다.

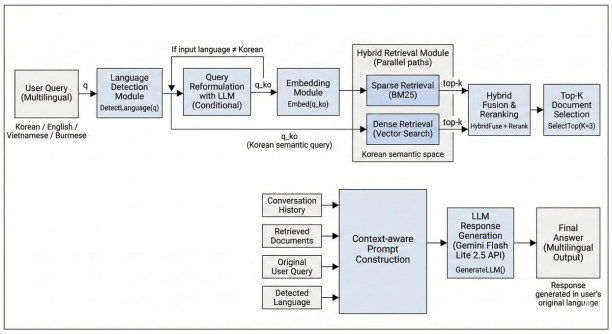

Fig. 5. Multilingual Architecture

2.6 Response Generation

응답 생성 단계에서는 검색 모듈을 통해 확보된 상위 문서를 기반으로 생성형 모델이 자연어 응답을 생성한다. 본 연구에서는 생성형 모델로 'gemini-flash-lite 2.5'를 API 형태로 호출하여 사용하였다[19].

2.7 Token Usage Management

본 연구의 학사 챗봇은 생성형 AI 모델을 API 방식으로 연동하여 구동된다. 이때 사용자의 질의와 모델의 응답 과정에서 소비되는 토큰(Token)은 모두 시스템 운영 비용과 직결된다. 즉, 이를 체계적으로 관리하는 기능이 필수적이다. 이에 관리자 페이지를 Fig 6과 같이 기간별 요청 및 응답에 따른 토큰 사용량을 시각화하여 제공한다. 관리자는 시각화된 데이터를 통해 챗봇의 이용 현황을 파악하고, 이를 안정적인 서비스 운영을 위한 객관적인 지표로 활용할 수 있다.

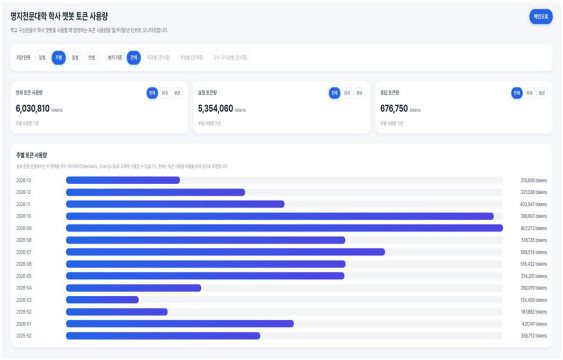

Fig. 6. Admin Page UI

본 연구에서는 앞서 설명한 하이브리드 검색, 맥락 기반의 응답, 언어 감지를 통한 다국어 지원과 같은 개별 모듈로 구현된 기능들을, 최종 응답 단계에서는 시스템 프롬프트(System Prompt)를 통해 통합적으로 제어한다. 즉, 시스템 프롬프트는 단순히 모델에게 역할 부여, 응답 형식 지정뿐만 아니라, 참조해야 할 정보와 범위를 제어하는 일종에 제어 계층으로써 활용된다. 구체적으로, 시스템 프롬프트에는 "검색을 통해 선택된 상위 문서만을 근거로 응답을 생성할 것","이전 대화 내용을 기반으로 맥락이 자연스럽게 이어지도록 응답할 것","언어 감지 결과에 맞는 답변을 응답할 것"과 같은 제약이 포함된다. 이를 통해 생성형 모델은 독립적인 추론 주체가 아닌, 사전에 정의된 시스템 흐름을 따르는 응답 생성기로 작동한다. 이러한 제약은 생성형 모델이 검색 결과에 포함되지 않은 내용을 임의로 생성하는 것을 방지한다.

IV. Conclusions

1. Conclusion

본 연구에서는 대학 생활의 주류인 Z세대의 정보 소비 패턴을 반영하여, 명지전문대학의 파편화된 학사 정보를 통합 제공하는 RAG 기반 학사 챗봇 시스템을 연구하였다. 연구 과정에서 데이터의 일관성 확보와 구조적 문제의 한계를 PrintFriendly와 PyMuPDF4LLM을 활용하여 노이즈가 제거된 마크다운 형식의 데이터셋을 구축할 수 있었다. 기술적으로는 하이브리드 검색과 맥락 기반의 질의 처리 기능을 도입하여 실제 학사 행정과 관련된 문의와 유사한 연속성 있는 대화 환경을 처리할 수 있도록 설계하였다. 언어 감지(lang_detect)를 통한 다국어 지원으로 외국인 학생의 접근성을 확장하였다. 이러한 기능들을 시스템 프롬프트를 통해 제어하여 환각 현상을 최소화하고 검색된 근거에 기반한 학사 챗봇 시스템을 구축하였다.

2. Expected Effects

본 연구를 통해 구축된 RAG 기반 학사 챗봇 시스템은 현재 대학 생활에 주류가 되는 Z세대의 '대화형 검색'을 선호하는 경향을 반영하였다. 학생은 홈페이지 내에서 직접 정보를 탐색하는 번거로움을 겪지 않고도 "휴학 절차에 대해 알려줘"와 같이 자연어 문장으로 원하는 정보를 탐색하여 효율적인 정보 탐색이 가능해질 것이라 기대한다.

매 학기 반복되는 단순 반복적인 학사 문의를 챗봇이 응대함으로써 행정 업무의 과부하를 해소한다. 이를 통해 교직원은 보다 고부가가치인 업무에 역량을 집중할 수 있는 환경이 마련되어 업무 생산성이 향상될 것이라 기대된다.

3. Limitations and Future Research Directions

본 연구에서는 일부 학사 데이터(학사 일정, 학과 소개 등)를 수집하는 과정에서 'PrintFriendly'를 활용한 수동 PDF 변환 방식에 의존하였다. 이로 인해 학사 일정이나 학과 정보의 변경 사항을 실시간으로 반영하기 어렵다는 한계가 존재한다. 또한, 벡터 데이터베이스 내에서 변경이 필요한 문서만을 식별하여 갱신하는 데 어려움이 있어, 단일 정보의 수정에도 전체 데이터를 삭제 및 재삽입해야 하는 관리적 비효율성이 발생한다.

향후 연구에서는 이러한 수동 수집 방식의 한계를 극복하기 위해, 비정형 웹 구조에서도 핵심 콘텐츠를 동적으로 추출할 수 있는 지능형 크롤링 파이프라인 구축이 필요하다. 또한, 문서의 변경 여부를 해시(Hash) 값으로 비교하

여 벡터 데이터베이스 내에서 변동이 발생한 청크만을 선택적으로 삭제 및 재삽입하는 방안을 도입한다면, 데이터 최신성과 운영 효율성을 동시에 향상시킬 수 있을 것으로 기대된다.

REFERENCES

[1] WebFX, “Conversational Search: How Younger Generations Are Leading the AI-Powered Search Revolution,” WebFX Research Report, 2025.

[2] Al-Jaf, Kanaan, Cemil Öz, Hoger Mahmud, and Tarik A. Rashid, “Leveraging Chatbots for Effective Educational Administration: A Systematic Review,” Preprints.org, preprint version 1, online 7 October 2024. DOI: 10.20944/preprints202410.0238.v1

[3] Brynjolfsson, E., Li, D., and Raymond, L. R., “Generative AI at Work,” National Bureau of Economic Research (NBER), Working Paper No. 31161, 2023. DOI: 10.3386/w31161

[4] MyongjiUniversity ‘MaruBot’ Retrievedfrom https://chatbot.mju.ac.kr/

[5] Sungyunkwan University ‘KingoBot’ Retrieved from https://kingo.skku.edu/chat

[6] Yoonji Nam, TaeWoong Seo, Gyeongcheol Shin, Sangji Lee, JaeEun Im, “NOVI : Chatbot System for University Novice with BERT and LLMs,” CoRR, abs/2409.06192, Sep. 10 2024. DOI: 10.48550/ARXIV.2409.06192

[7] Zhang, Y., Li, H., and Liu, X., “Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models,” arXiv preprint, 2023. DOI: 10.1162/coli.a.16

[8] Ji, Z., Lee, N., Frieske, R., et al., “A Survey of Hallucination in Natural Language Generation,” ACM Computing Surveys, Vol. 55, No. 12, pp. 1–38, December 2023. DOI: 10.1145/3571730

[9] Zhao, W. X., Zhou, K., Li, J., et al., “A Survey of Large Language Models,” ACM Computing Surveys, Vol. 56, No. 3, pp. 1–45, 2023. DOI: 10.1145/3581783

[10] Park, Jongjin, “Development of Chatbot Using Knowledge Graph and AI Agent,” The International Journal of Internet, Broadcasting and Communication, Vol. 17, No. 1, pp. 307–315, January 2025. DOI: 10.7236/IJIBC.2025.17.1.307

[11] Sookmyung women’s University ‘AI noonsong’ https://www.sookmyung.ac.kr

[12] Konkuk University ‘AI KU’ https://www.konkuk.ac.kr

[13] Artifex, Using PyMuPDF as a Data Feeder in LLM / RAG Applications, https://pypi.org/project/pymupdf4llm/

[14] LangChainReference, langchain-text-splitters, https://reference.langchain.com/python/langchain_text_splitters

[15] Nlpai-lab, KURE-v1, https://huggingface.co/nlpai-lab/KURE-v1

[16] Abraham Itzhak Weinberg, “Hybrid Dense-Sparse Retrieval for High-Recall Information Retrieval,” January 2026. DOI:10.13140/RG.2.2.23909.46562

[17] LangChainReference, EnsembleRetriever, https://reference.langchain.com/v0.3/python/langchain/retrievers/langchain.retrievers.ensemble.EnsembleRetriever.html

[18] Pypi, langdetect 1.0.9, https://pypi.org/project/langdetect/

[19] Google Cloud, Gemini 2.5 Flash-Lite https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash-lite

Authors

Yong-Min Cho is currently studying Artificial Intelligence and Big Data at Myongji College from 2025 to the present. His research interests in the field of data science and AI.

Sung-Jin Kim receved as B.S., M.S. in Computer Science from Halla University 2013, 2015 and 2021 Ph.D. degree data science part in multimedia engineering at GangNeung-Wonju National University in Wonju, Korea. He currently holds assistant professor with the Department of Artificial Intelligence and Big Data at Myongji College. His currently research interests include artificial intelligence and data science.