대한한의진단학회지 제30권 제1호 (2026년 10월)
The Journal of the Society of Korean Medicine Diagnostics. 2026;30(1):357-369.
[원저 Original Article]

RAG 기반 한의 생성형 AI 시스템(KMGPT) 개발에 대한 연구

서수민[1], 김안나[1], 김영은[1], 손미주[2], 장윤지[1], 김상균[1]*
[1]한국한의학연구원 한의AI응용연구실
[2]한국한의학연구원 근거중심한의학연구실

Development of a Retrieval-Augmented Generation Based Generative AI System for Korean Medicine (KMGPT)

Sumin Seo[1], Anna Kim[1], Young-Eun Kim[1], Mi Ju Son[2], Yunji Jang, Sang-Kyun Kim[1]*
[1]Korean Medicine Applied AI Research Department, Korea Institute of Oriental Medicine
[2]Evidence-Based Korean Medicine Research Department, Korea Institute of Oriental Medicine

Abstract

This study aims to develop KMGPT, a generative AI system for Korean medicine that integrates large language models (LLMs) with retrieval-augmented generation (RAG) based on Korean medicine literature. General-purpose LLMs are trained mainly on Western biomedical data. They show limited understanding of Korean medicine concepts, classical Chinese texts and domain terminology, and may generate unsupported statements (hallucinations). To address these limitations, we constructed a knowledge base integrating classical Korean medicine texts, a Korean medicine ontology derived from textbooks, research articles (OASIS), a Korean medicine terminology dictionary and clinical practice guidelines of Korean medicine. The texts were cleaned, segmented into semantic chunks and annotated with bibliographic metadata, then embedded and stored in a vector database. The RAG pipeline consists of query processing, hybrid retrieval, re-ranking and answer generation with explicit source citation. KMGPT was implemented as a private on-premise web-based service displaying retrieved evidence alongside each answer. We examined the system through representative query scenarios covering prescriptions and clinical guidelines, terminology and pattern differentiation, herbal formulas and medicinal herbs with research literature, and clinical guideline-based evidence retrieval. KMGPT produced answers grounded in the retrieved literature with source citations, whereas a general-purpose LLM gave answers without verifiable sources in the selected cases. KMGPT offers a foundation for evidence-based generative AI in Korean medicine research, education and clinical decision support. The study is limited to qualitative case review without quantitative evaluation of retrieval or answer quality. Systematic assessment using retrieval metrics and answer quality benchmarks is needed in future work.

Key words : Korean Medicine; Retrieval-Augmented Generation; Large Language Model; Generative AI; Knowledge Base; Ontology

---

* 교신저자 : Sang-Kyun Kim. Korean Medicine Applied AI Research Department, Korea Institute of Oriental Medicine. 1672 Yuseong-daero, Yuseong-gu, Daejeon, 34054, Republic of Korea.
Tel: +82-42-868-9526, E-mail: skkim@kiom.re.kr
· 원고접수일 : 2026.09.30 / 심사완료일 : 2026.10.15 / 게재결정일 : 2026.10.21.

I. 서론

최근 거대언어모델(Large Language Model, 이하 LLM)을 중심으로 한 생성형 인공지능(Artificial Intelligence, 이하 AI) 기술이 빠르게 발전하면서 의료 분야의 활용 연구가 활발히 이루어지고 있다. LLM은 대규모 텍스트 학습을 통해 다양한 언어 이해와 생성 능력을 보이며[1-2], 의학 지식 질의응답, 임상 문서 요약, 의학교육 등에서 활용 가능성이 보고되었다[3]. 국내 한의학 분야에서도 교육용 AI와 진료수행시험(CPX) 연습 챗봇이 개발되고, 한의사 국가시험 문항을 이용한 LLM의 한의학 지식 평가가 이루어졌다[4-6].

그러나 범용 LLM을 한의학 도메인에 적용하는 데에는 분명한 한계가 있다. 첫째, 대부분의 LLM은 영어권과 서양의학 중심 데이터로 학습되어 한의학 지식의 밀도가 낮다[4]. 둘째, 한의학 지식의 상당 부분은 한문 또는 국한문 혼용 고전 문헌에 기록되어 있고, 변증, 본초, 경혈 등 고유한 개념 체계를 갖는다. 셋째, LLM은 그럴듯하지만 사실과 다른 내용을 생성하는 환각(hallucination) 현상을 보이며[7], 검증 가능한 출처를 일관되게 제시하지 못할 수 있다. 근거의 출처를 확인할 수 없는 응답은 연구·교육·임상 현장에서 신뢰하기 어렵다. 넷째, LLM은 주로 온라인에 공개된 데이터로 학습되므로, 병원의 전자의무기록(Electronic Medical Record, 이하 EMR)이나 기관 내부에 문헌 데이터처럼 온라인 접근이 제한된 자료는 답변 생성에 직접 활용하기 못하여 올바른 답변을 제시하기 어렵다.

LLM에 도메인 지식을 반영하는 방법에는 파인튜닝(fine-tuning)이나 사전학습, 외부 지식을 검색해 제공하는 검색 증강 생성(Retrieval-Augmented Generation, 이하 RAG)이 있다[8]. RAG는 질의와 관련된 문헌을 지식베이스에서 검색하여 LLM에 함께 제공하므로, 검색된 문헌에 근거한 답변을 생성하고 그 출처를 명시할 수 있다[9]. 새 문헌의 추가와 재색인을 통해 지식베이스를 갱신할 수 있어, 모델 파라미터의 재학습 없이도 새로운 자료를 활용할 수 있다[4]. 다만 구축과 운영의 부담과 응답 품질은 자료의 규모 및 검색 및 생성 구성에 따라 달라진다. 의료 분야에서는 임상지침과 의학 문헌을 활용한 RAG의 성능 및 근거 활용 특성을 평가하는 연구가 보고되었다[10-11].

한의학 분야의 관련 연구는 범용 LLM의 한의학 지식 평가, 도메인 특화 모델 개발, 문헌 검색을 결합한 시스템 개발로 확장되고 있다. Jang 등은 GPT-4의 한의사 국가시험 정답률이 66.18%의 정답률로 합격 기준(60%)을 넘었으나, 한의학 전문지식이 필요한 문항의 정답률은 55.4%로 그렇지 않은 문항(82.4%)보다 크게 낮았고, 침구의학 등 한국 한의학에 특화된 과목에서 특히 취약하였다고 보고하였다[6]. 중의학 분야에서는 약 2GB의 중의학 말뭉치로 추가 사전학습과 지도 파인튜닝을 수행한 특화 모델 Qibo가 개발되었다[12]. TCMBench는 중의학 성능이 여전히 미흡하고, 도메인 특화 파인튜닝 모델이 오히려 분석 문장의 품질이 낮아 기본 능력이 손상될 수 있음을 지적하였다[13]. 최근에는 중의학 고전 문헌으로 구축한 지식그래프와 그래프 기반 RAG(GraphRAG)를 결합하여 모델 파인튜닝 없이 문헌 검색과 진단 질의응답을 지원하는 OpenTCM이 제안되는 등[14], 외부 문헌 지식을 검색해 활용하는 접근이 주목받고 있다. 그러나 한의학의 다양한 지식원을 통합하고 답변의 근거 출처를 제시하는 RAG 기반 생성형 AI 시스템에 관한 연구는 아직 부족하다.

본 연구는 한의학 문헌 데이터를 기반으로 RAG 기반 한의 생성형 AI 시스템 KMGPT (Korean Medicine GPT)를 개발하고, 대표 질의 사례를 통해 한의학 연구·교육 및 임상 분야

에서의 활용 가능성을 제시하고자 한다. 본 시스템은 한의 고문헌, 한의 온톨로지, 논문, 용어사전, 임상진료지침을 단일 지식베이스로 통합하고 동의어 처리를 파이프라인에 반영하였으며, 기관 내부 자료의 보안을 고려한 프라이빗 환경에서 구현하였다는 점에서 기존 연구와 구별된다.

II. 연구 방법

1. 연구 설계 개요

본 연구는 한의 문헌 기반 생성형 AI 시스템의 설계와 구현에 초점을 둔 개발 연구이다. 세부 목표는 한의 문헌을 통합한 지식베이스 구축, 한의학 용어 처리를 포함한 RAG 파이프라인을 설계, 답변과 근거 문헌을 함께 제시하는 웹시스템 구현 및 대표 사례 검토로 설정하였다. 연구는 문헌 수집, 전처리 및 구조화, 임베딩과 벡터 DB 구축, RAG 파이프라인 설계, 시스템 구현, 사례 기반 정성 검토 순으로 수행하였다(Figure 1). 비교 검토는 선정된 사례의 응답 특성을 기술하는 탐색적 절차로 한정하였다.

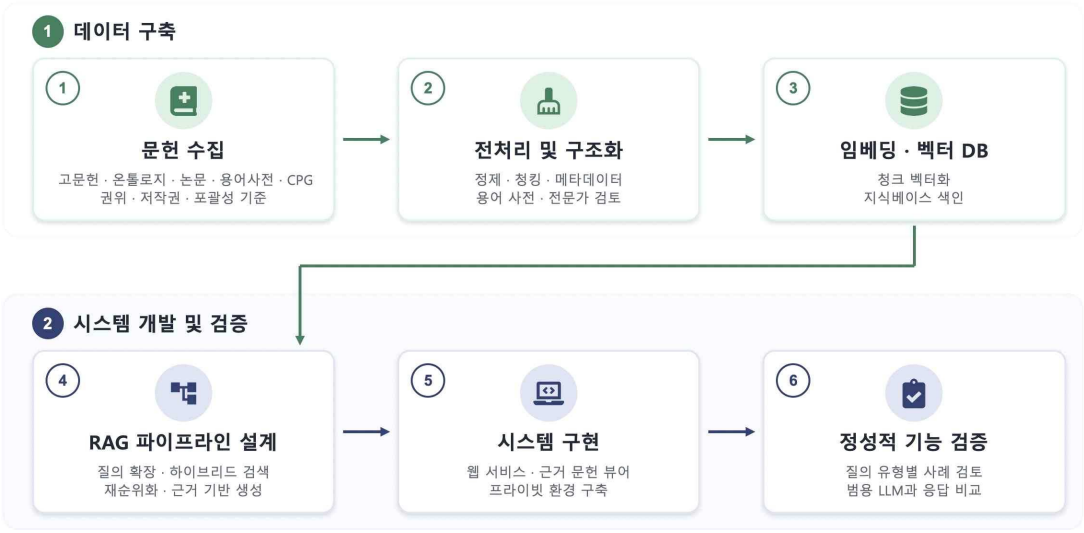


Figure 1. Study Design for KMGPT

2. 데이터 수집 및 지식베이스 구성

한의 지식베이스는 한의 고문헌, 한의 온톨로지[15], 한의 논문(OASIS)[16], 한의용어사전[15,17], 한의임상진료지침[18]으로 구성하였다. 각 데이터원의 형태와 규모는 Table 1과 같다.

Table 1. Configuration of KM Knowledge Base

데이터원	형태	규모 / 특징
한의 고문헌	텍스트	82권의 한의 고문헌 데이터
한의 온톨로지	텍스트	한의대 교과서에서 약재·처방·병증·침구경혈 데이터를 추출 및 가공하여 온톨로지 모델링한 약 100만 트리플 규모의 RDF 데이터. RDF를 텍스트로 변환하여 사용
한의 논문(OASIS)	텍스트	한의 분야 대표 논문 검색 사이트. 343,489건의 논문 데이터
한의용어사전	텍스트	병증·변증·주치·효능 등 용어에 대한 28,146 건의 일반인 대상 설명 데이터
한의임상진료지침	텍스트	78개의 임상진료지침 데이터

3. 데이터 전처리 및 구조화

수집 자료는 검색과 출처 확인에 적합한 단위로 정제하였다. 임상진료지침과 논문의 PDF에서 텍스트를 추출하고 머리말·쪽번호 등 불필요한 요소를 제거하였다. 온톨로지는 관계 정보를 텍스트로 변환하여 다른 문헌 자료와 함께 임베딩하였다. 본 연구에서 온톨로지는 관계 지식을 담은 검색 자료로 활용하였으며, 별도의 그래프 탐색이나 그래프 기반 추론을 구현한 것으로 해석하지 않았다. 논문은 단락 단위, 진료지침은 권고사항·절 단위, 용어사전은 표제어 단위로 분할하였다. 각 청크에는 데이터원, 문헌명 또는 표제어, 문헌 내 위치 정보를 부여하여 자료원별 검색 필터링과 출처 표기에 사용하였다. 동일 대상을 가리키는 한글·한자 표기와 별칭은 동의어 사전으로 연결하고 한의 표준 용어 체계와 매핑하였다. 경력 15년 이상의 한의사 3인이 정제 결과와 메타데이터를 검토하였다. 의견이 엇갈린 항목은 원문을 재확인한 뒤 합의하였다.

4. RAG 파이프라인 설계

KMGPT의 RAG 파이프라인은 지식베이스 색인, 질의 처리, 검색 및 재순위화, 답변 생성, 출처 제시의 다섯 가지 기능적 요소로 구성하였다(Figure 2). 지식베이스 색인은 질의 처리에 앞서 수행하며, 실시간 질의에서는 검색된 청크를 생성 모델에 전달하도록 구성하였다. 각 단계의 실제 구현은 에이전트 루프 방식으로 실행된다.

1) 임베딩과 벡터 DB: 전처리된 청크는 Qwen3-Embedding-8B[19]로 벡터화하고 벡터 DB Milvus[20]에 저장하였다. 임베딩 모델은 한글과 한자가 혼재된 자료의 검색에 사용할 수 있는 다국어 모델로 선정하였다.

2) 질의 처리: 사용자 질의는 용어 사전을 이용하여 동의어와 한자 표기로 확장하고, 필요한 경우 LLM을 이용하여 검색에 적합한 형태로 재작성하였다. 이러한 처리는 표현 차이에 따른 검색 누락을 줄이기 위한 설계 요소로 사용하였다.

3) 검색 및 재순위화: 처방명·약재명 등 표기 일치가 중요한 질의와 서술형 질의를 함께 처리하기 위해 키워드 기반 검색(BM25)[21]과 벡터 유사도 검색[22]을 결합한 하이브리드 검색을 적용하였다. 두 검색 결과는 Reciprocal Rank Fusion(RRF)[23]로 결합하여 상

위 10개 후보를 구성하였다. 이후 교차 bge-reranker-v2-m3[24]로 재순위화하여 최종 5개를 선별하였다. 사용자가 문헌 유형(예: 진료지침, 논문)을 지정한 경우에는 각 청크에 부여한 데이터원 정보를 이용하여 해당 유형의 문헌으로 검색 범위를 제한하였다.

4) 답변 생성: 선별된 청크와 사용자 질의를 프롬프트에 결합하고 내부 서버의 Qwen3.8-27B[25] 모델로 답변을 생성하였다. 프롬프트에는 검색된 문헌을 근거로 답변할 것, 문헌에 근거한 문장에 출처 번호를 부여하고 참고문헌 목록을 제시할 것, 근거가 부족하면 해당 정보를 찾지 못했다고 밝히고 추측을 제한할 것을 지시하였다.

5) 출처 제시: 답변의 출처 번호를 선택하면 근거 청크의 원문과 문헌명 및 문헌 내 위치를 표시하도록 구현하였다. 출처제시와 근거 부족시 응답 제한 지시는 사용자의 근거 확인을 지원하기 위한 설계이다.

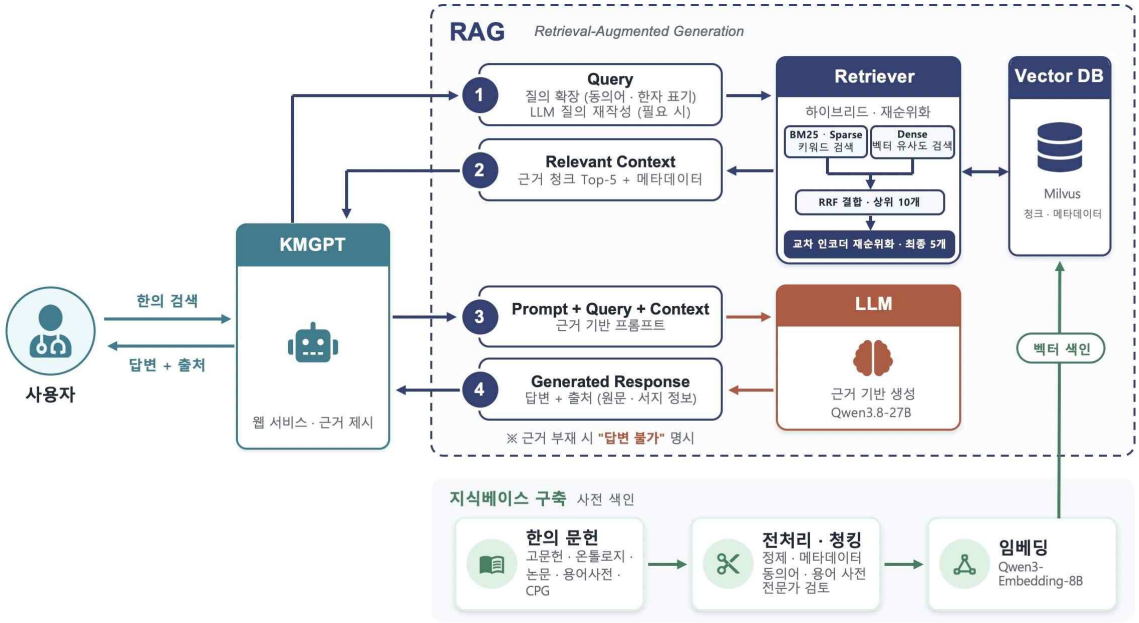


Figure 2. System Architecture of KMGPT

5. 시스템 구현 환경 및 구성

서버는 Python 웹 프레임워크 FastAPI, 웹 화면은 Next.js로 구현하였다. 검색과 생성은 자체 에이전트 루프로 연결하였으며, 이 루프에서 LLM이 함수 호출을 통해 검색 도구를 사용하고 반환된 결과를 근거로 답변하도록 구성하였다. 이 에이전트 루프는 사전에 구축된 색인을 기반으로 질의 처리, 검색·재순위화 및 답변 생성을 수행하되, 검색 도구 호출 여부와 추가 검색을 결정하는 구조이다. 따라서 연구방법에서 제시한 다섯 가지 기능적 구성요소는 시스템의 처리 흐름을 나타내며, 색인은 지식베이스 구축 과정에서 사전에 수행하고 나머지 단계는 사용자 질의와 검색 결과에 따라 연계하여 실행한다. 벡터 검색에는 Milvus 클라이언트인 pymilvus를 사용하였다.

시스템은 질의응답 모듈, 근거 문헌 뷰어, 대화 이력 모듈, 관리자 모듈로 구성하였다. 관

리자 모듈은 지식베이스, 용어 사전, 모델, 사용자 관리를 담당한다. 지식베이스와 사용자 질의 기록을 외부로 유출되지 않도록 내부 서버에서 처리하여 프라이빗 환경에 구축하고, 사용자 인증을 거쳐 접근하도록 하였다. 이러한 구성은 내부 자료의 외부 전송을 제한하기 위한 운영 구조이며, 별도의 보안성 평가 결과를 의미하지 않는다.

내부 서버는 NVIDIA RTX PRO 6000 Blackwell GPU(96GB VRAM), Intel Core Ultra 9 CPU, 그리고 128GB 시스템 메모리를 탑재한 환경으로 구성하였다. 생성 모델과 임베딩 모델을 내부 서버에서 운영하여, 병원과 같이 제한된 전산 환경에서도 외부 클라우드의 추론 API에 의존하지 않는 구성을 구현하였다.

6. 기능 검증 방법

본 연구는 검색과 생성 성능의 정량 평가 대신 대표 질의 사례에 대한 정성적 검토를 수행하였다. 질의는 한의 지식베이스 통합 검색의 특성을 살펴보기 위해 처방 및 임상진료지침(알레르기 비염에서 소청룡탕 질의), 한의학 용어·변증 비교(불면(不眠)의 심비양허(心脾兩虛)·심신불교(心腎不交)·담화요심(痰火擾心) 비교 질의), 처방·본초·연구문헌 연계(황련해독탕(黃連解毒湯) 구성 약재 및 효능·연구문헌 질의), 임상진료지침·근거문헌 검색(기능성 소화불량의 한약 치료 권고사항 및 근거문헌 질의)의 네 유형으로 구성하였다(Table 2).

한의사 3인은 각 응답을 근거 적합성, 한의학적 타당성, 출처 일치 여부의 관점에서 검토하였다. 근거 적합성은 검색된 자료가 질의와 답변 내용을 뒷받침하는지, 한의학적 타당성은 용어·변증·처방 설명이 적절한지, 출처 일치 여부는 답변에 연결된 문헌과 원문 발췌물이 일치하는지를 살펴보는 기준으로 구분하였다. 검토 결과는 사례별로 서술하였다.

또한 동일한 질의를 KMGPT 지식베이스와 연계하지 않은 범용 LLM에 입력하여 근거 제시 방식과 응답 특성을 탐색적으로 비교하였다. 이 비교는 선정된 사례에서 관찰한 응답을 기술하기 위한 것으로, 일반적인 모델 성능의 우열이나 RAG의 독립적 효과를 검증하는 설계는 아니었다.

III. 결과

1. 지식베이스 구축 결과

한의 고문헌, 한의 온톨로지, 한의 논문, 한의용어사전, 한의임상진료지침을 가공하여 검색 가능한 지식베이스를 구축하였다(Table 1). 온톨로지의 관계 정보는 텍스트로 변환하였고, 문헌 자료와 용어 설명은 각각의 의미 단위로 분할하였다. 청크에는 자료원과 문헌 식별·위치 정보를 연결하여 검색 결과에서 원문 근거를 확인할 수 있도록 하였다.

구축된 지식베이스는 처방과 약재의 관계 정보, 용어 설명, 연구문헌 및 지침 내용을 동일한 질의 과정에서 검색할 수 있는 자료 기반을 제공하였다.

2. RAG 파이프라인 구현 결과

사용자 질의는 질의 확장, 하이브리드 검색, RRF 결합, 재순위화, 근거 결합, 답변 생성 순으로 처리되었다(Figure 3). “알레르기 비염에서 소청룡탕에 대해 설명해줘”와 같은 질의에 대해 용어 확장과 문헌 검색을 수행하고, 선별된 청크를 근거로 출처 번호가 포함된 답변을 생성하도록 구현하였다. Figure 3은 이 처리 절차를 설명한 도식이다.

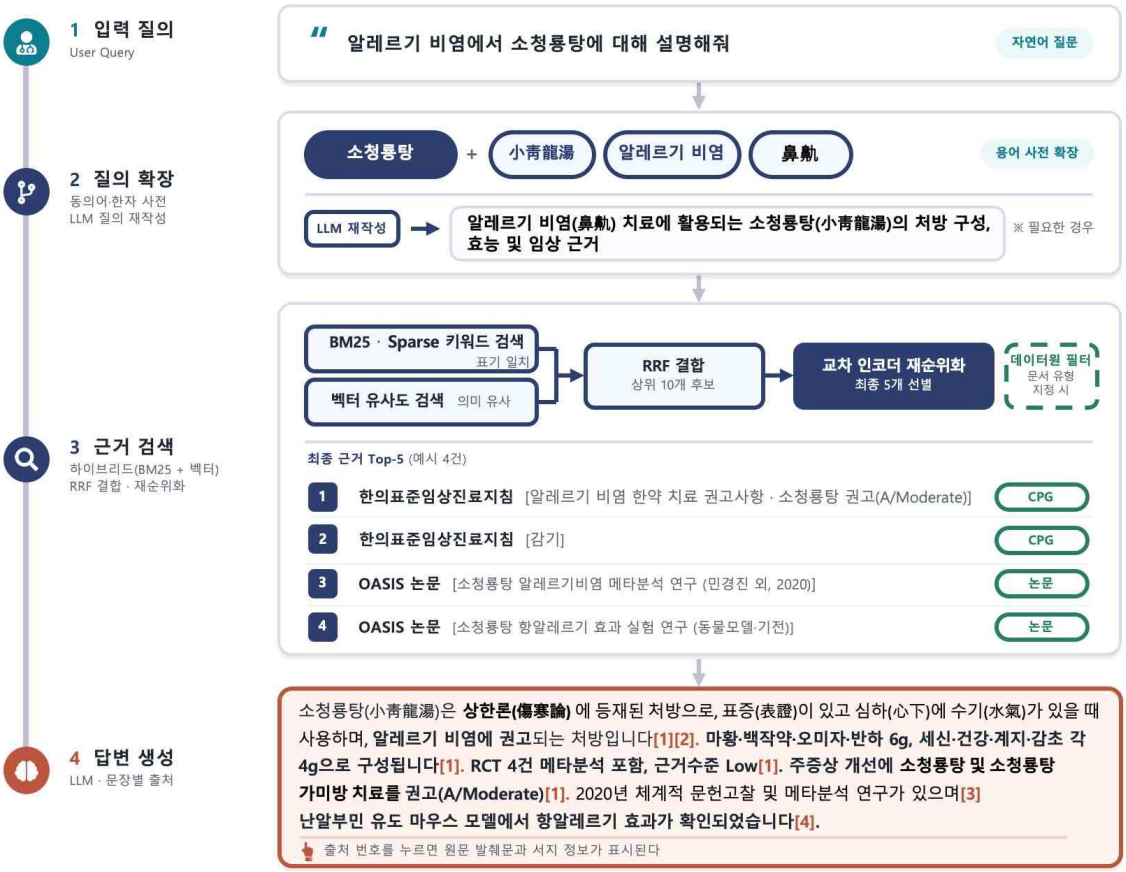


Figure 3. Schematic Example of the Query Processing Workflow in KMGPT

3. KMGPT 시스템 구현

사용자 화면은 질의창, 답변 영역, 근거 문헌 뷰어, 채팅 목록 영역으로 구성하였다(Figure 4). 사용자는 자연어로 질문하고, 필요할 경우 데이터원을 선택하여 검색 범위를 좁힐 수 있다. 답변의 출처 정보 버튼을 선택하면 원문 발췌물과 문헌 식별 정보 및 문헌 내 위치가 표시된다(Figure 5). 이 기능은 사용자가 답변과 검색 근거를 대조할 수 있도록 지원한다.

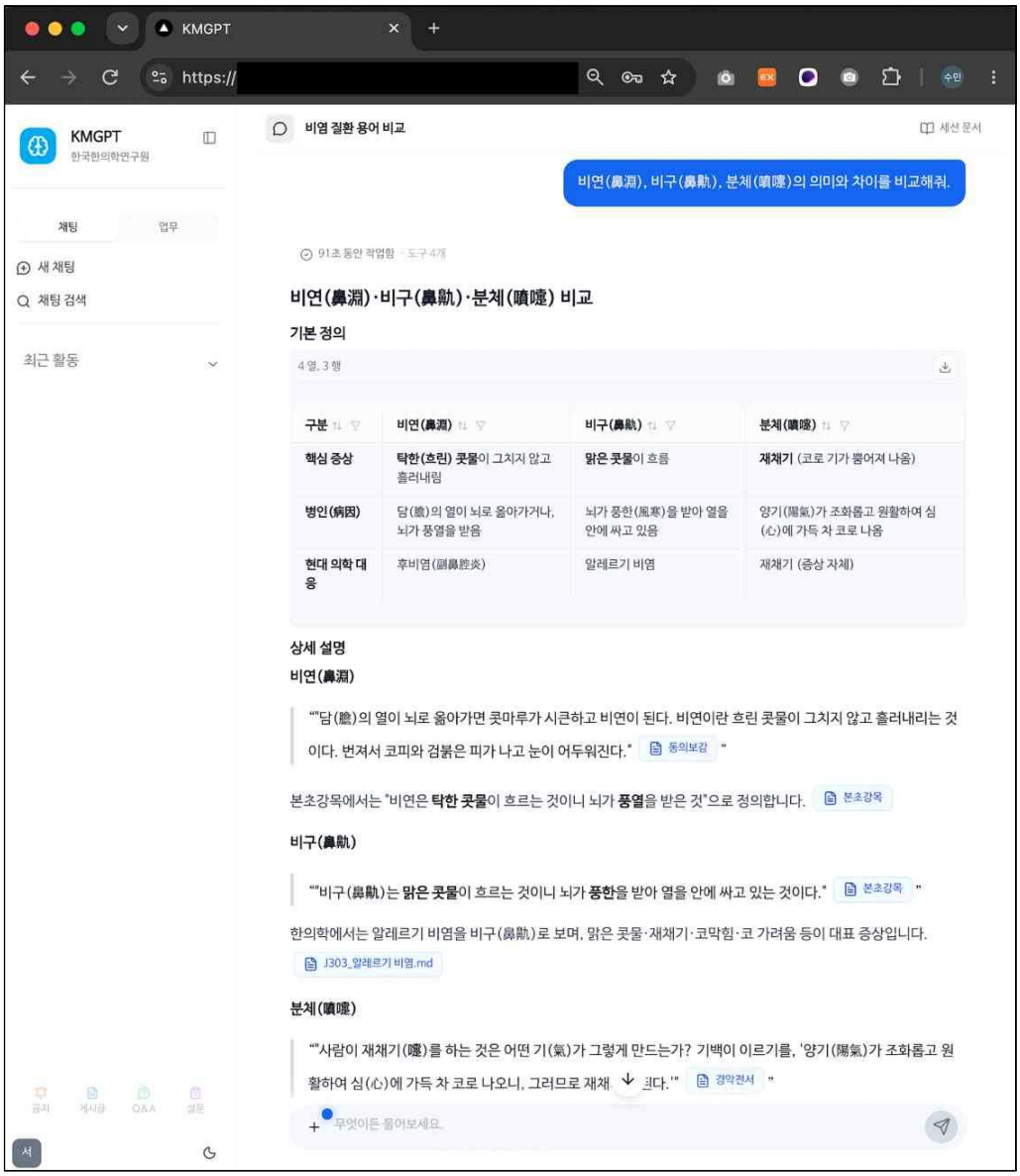


Figure 4. KMGPT User Interface and an Example Response to a Question



Figure 5. Enlarged Source-Detail View Showing a Retrieved Passage and Document Information

4. 질의응답 활용 사례

질의 유형별 대표 활용 사례는 Table 2와 같다. 한의사 3인의 정성적 검토 결과, KMGPT는 처방 구성과 용량, 한의학적 변증 개념, 관련 연구문헌 및 임상진료지침의 권고사항을 출처와 함께 제시하였다.

처방 및 임상진료지침 사례에서는 알레르기 비염의 소청룡탕 질의에 대해 구성 약재별 용량, 임상적 활용 및 진료지침 출처를 통합적으로 제시하였다. 한의학 용어·변증 비교 사례에서는 불면(不眠)의 심비양허(心脾兩虛)·심신불교(心腎不交)·담화요심(痰火擾心)을 주요 증상과 대표 처방을 중심으로 비교하고 문헌 근거를 연결하였다. 처방·본초·연구문헌 연계 사례에서는 황련해독탕(黃連解毒湯)의 구성 약재와 각 약재의 효능을 설명하고 관련 연구문헌을 제시하였다. 기능성 소화불량 사례에서는 한약 치료의 권고사항, 주요 처방 및 관련 근거문헌을 함께 제시하였다.

Table 2. Representative queries and information presented in KMGPT responses

질의 유형	대표 질의	응답에 포함된 정보
처방 및 임상진료지침	알레르기 비염에서 소청룡탕에 대해 설명해줘	구성 약재, 용량, 임상적 활용, 진료지침 출처
한의학 용어·변증 비교	불면에서 심비양허, 심신불교, 담화요심의 차이를 주요 증상과 대표 처방을 중심으로 비교하고 문헌 근거를 제시해줘	변증별 증상과 처방, 관련 문헌 근거
처방·본초·연구문헌 연계	황련해독탕의 구성 약재와 각 약재의 한의학적 효능을 설명하고, 황련해독탕과 관련된 연구문헌을 출처와 함께 제시해줘	구성 약재와 효능, 처방, 관련 연구문헌
임상진료지침 및 근거문헌 검색	기능성 소화불량의 한약 치료에 대해 권고사항과 주요 처방과 관련 근거문헌을 함께 제시해줘	임상진료지침 권고사항, 주요 처방 및 관련 근거문헌을 연결하고 출처 추적 가능

이상의 네 가지 사례에서 KMGPT는 처방 구성, 변증 개념, 연구문헌, 진료지침 권고사항을 각각의 근거 출처와 함께 제시하였다.

5. 선정된 질의 사례에서 응답 특성 비교

비교 사례에서는 근거 제시 방식의 차이가 관찰되었다(Table 3). 범용 LLM 응답에는 출처가 제시되지 않거나 웹페이지·문서 전체 수준의 출처가 제시된 사례가 있었으며, KMGPT는 검색한 청크의 원문 발췌물과 문헌 정보를 확인할 수 있도록 하였으며 근거가 부족한 질의에는 답변할 수 없음을 알렸다. 선정된 사례의 서술적 비교이며 오류율, 정확도 또는 모델 간 성능 우위를 의미하지 않는다.

Table 3. Exploratory Observations on Response Characteristics in Selected Cases

비교 항목	범용 LLM의 관찰 사례	KMGPT의 관찰 사례
근거 제시	출처가 없거나 문서 전체 수준의 출처가 제시된 사례	문장별 출처와 원문 발췌물 제시
용어 및 개념 표현	서양의학 용어로 대체하거나 한의학 개념을 단순화하여 설명한 사례	한의 고문헌, 한의 용어사전 및 온톨로지 기반으로 한자·한글 표기와 문헌 기반 설명 제시
근거 부족 시 응답	추정적 설명이 포함된 사례	근거 부족을 알리고 답변을 제한한 사례

IV. 고찰

본 연구는 한의 고문헌, 한의 온톨로지, 한의 논문, 한의용어사전, 한의임상진료지침을 통합한 한의 지식베이스와 RAG를 결합하여 KMGPT를 개발하였다. 주요 기여는 서로 다른 형태의 한의 문헌을 검색 가능한 자료로 구성하고, 생성 답변에서 문헌의 원문 근거를 확인할 수 있는 프라이빗 웹 시스템을 구현한 데 있다. 네 가지 대표 사례는 처방·변증·연구문헌·진료지침에 걸친 복합적 질의에서 시스템이 근거 기반 응답을 생성할 수 있음을 보여주며, 이는 검색 성능이나 임상적 유용성의 정량적 입증과는 구분된다.

첫째, RAG는 모델 재학습 없이 외부 한의학 자료를 답변 생성에 활용하는 구조를 제공하였다. 범용 LLM의 한의학 전문지식에는 한계가 보고되어 있으며[4,6], 질의 시점에 관련 문헌을 제공하는 방식은 이를 보완하기 위한 접근이 될 수 있다. 새 진료지침이나 문헌이 발간되면 지식베이스에 추가하는 것만으로 지식베이스에 반영할 수 있다는 점도 운영상 장점이다[8-9].

둘째, 근거를 추적할 수 있는 기능은 한의학 문헌 활용에서 중요하다. 고문헌의 내용, 온톨로지의 관계 정보, 용어사전의 설명, 개별 연구 결과, 임상진료지침의 권고는 자료의 목적과 근거 수준이 다르다. KMGPT는 자료원과 원문 발췌물을 함께 제시하여 사용자가 답변의 근거를 검토할 수 있도록 하였다. 그러나 출처가 존재하거나 연결된 원문이 일치한다는 사실만으로 답변의 해석이 타당하다고 보장할 수는 없다. 또한 근거가 없는 질의에 답변 불가를 명시하도록 한 설계는 환각의 위험을 줄이는 데 기여한다[7]

셋째, 선행 연구와 비교하면 다음과 같은 차별점이 있다. 의료 분야의 RAG 연구는 주로 영문 의학 문헌과 서양의학 진료지침을 대상으로 하였다[10-11]. 국내 한의학 분야의 LLM 연구는 CPX 연습 등 교육 목적의 활용이나 범용 LLM의 한의학 지식 평가에 집중되어 있었고[5-6], 중의학 분야에서는 특화 모델의 파인튜닝이나 지식그래프 기반 RAG가 시도되었다[12-14]. KMGPT는 고문헌·논문·진료지침 같은 텍스트 자료와 한의 온톨로지 같은 그래프 자료를 텍스트로 변환하여 하나의 지식베이스로 통합하고, 동의어 등 용어의 처리를 파이프라인에 반영하였다는 점에서 구별된다. 또한 프라이빗 환경에 구축하여 기관 내부 데이터와 사용자 질의의 보안을 확보하였다.

넷째, 내부 서버와 사용자 인증을 활용한 구성은 기관 내부 자료를 다루는 서비스의 운영 기반이 될 수 있다. 다만 내부 서버에 배치했다는 사실만으로 보안성이 검증되는 것은 아니며, 실제 운영에서는 접근권한, 로그 관리, 외부 통신 및 자료 이용 범위를 함께 관리해야 한

다. 본 연구는 보안성 시험이나 개인정보 처리 적합성을 평가한 연구는 아니다.

활용 측면에서 KMGPT는 연구자의 문헌 탐색, 한의대 학생의 개념 학습, 임상가의 진료지침 조회를 지원할 수 있다. 교육용 CPX 시스템에 문헌 검색 기능을 연계하거나[5], 임상 분야에서는 진료지침의 권고사항을 빠르게 조회하는 의사결정 보조 도구로 활용될 수 있다. 향후 수록 고전 문헌을 확대하여 근거 자료의 범위를 확대할 수 있다. 이러한 활용은 후속 연구에서 사용자 요구와 실제 업무 효과를 검증할 예정이다.

그러나 본 연구에는 다음과 같은 한계가 있다. 첫째, 검색 정확도와 답변의 사실성·근거 충실도를 정량적으로 평가하지 않았다. 향후 한의학 질의응답 데이터셋을 구축하고 Recall@K, MRR 등의 검색 지표와 답변의 근거 지지 여부, 인용 정확성 및 누락을 평가할 필요가 있다. 근거 부족 시 응답 제한에 대해서도 적절한 답변 거부와 불필요한 거부를 구분하여 평가해야 한다. 둘째, 선정된 대표 사례와 소수 전문가의 검토는 질의 전체의 성능을 대표하지 않는다. 사례 선택과 검토자의 연구 참여 여부에 따른 편향 가능성이 있으며, 전문가 간 판단 일치도와 독립적인 외부 검토가 필요하다. 셋째, 수록 자료와 전처리의 범위에 따라 검색 가능한 정보가 제한된다. PDF 추출 오류, 청크 분할에 따른 문맥 손실, 용어 매핑의 오류, 온톨로지의 텍스트 변환 과정에서 관계 구조가 충분히 보존되지 않을 가능성을 검토해야 한다. 최신 문헌과 지침 개정판의 갱신, 상충하는 근거의 처리도 후속 과제이다. 넷째, 연구자·학생·임상가를 대상으로 한 사용성 평가와 실제 활용 효과 검증은 이루어지지 않았다. 문헌 탐색 시간, 근거 확인의 용이성, 사용자 만족도 및 실제 업무에서 발생하는 오류를 평가하여 적용 범위를 구체화할 필요가 있다.

V. 결론

본 연구에서 개발한 KMGPT는 한의학 문헌 기반 지식베이스와 RAG를 결합하여, 문헌을 근거로 답하고 그 출처를 제시하는 한의 생성형 AI 시스템이다. 대표 질의 사례를 통해 범용 LLM의 근거 부재와 한의학 지식 부족을 보완할 수 있음을 보여주었다. KMGPT는 한의학 연구의 문헌 탐색, 한의대 학생의 학습, 임상 의사결정 보조에 활용될 수 있으며, 한의학 분야에서 신뢰할 수 있는 생성형 AI를 구축하기 위한 기반을 마련하였다. 향후 평가 데이터셋을 활용한 정량 평가, 한의 지식베이스 확대, 실사용자 대상 실증 연구가 이루어진다면 KMGPT는 한의학 지식 서비스 플랫폼으로 발전할 수 있을 것이다.

감사의 글

본 연구는 한국한의학연구원의 “RAG 기반 한의 프라이빗 생성형AI 서비스 개발 및 임상실증(KSN2512021)”의 지원을 받아 수행되었다.

참고문헌

1. Bommasani R, et al. On the Opportunities and Risks of Foundation Models.

arXiv:2108.07258. 2021. https://doi.org/10.48550/arXiv.2108.07258

2. OpenAI. GPT-4 Technical Report. arXiv:2303.08774. 2023. https://doi.org/10.48550/arXiv.2303.08774
3. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-180. https://doi.org/10.1038/s41586-023-06291-2
4. Park SY, Kim CE. 거대언어모델을 활용한 한의학 교육 강화: 교육용 인공지능 개발을 중심으로. J Physiol & Pathol Korean Med. 2023;37(5):134-138. http://doi.org/10.15188/kjopp.2023.10.37.5.134
5. Kim JD, Lee HY, Kim JH, Kim CE. Development of an LLM-based CPX Practicing Chatbot for Korean Medicine Education: Implementation of Automated Scoring and Feedback Generation Framework. J Korean Med. 2024;45(4):215-230. https://doi.org/10.13048/jkm.24067
6. Jang D, Yun TR, Lee CY, Kwon YK, Kim CE. GPT-4 can pass the Korean National Licensing Examination for Korean Medicine Doctors. PLOS Digit Health. 2023;2(12):e0000416. https://doi.org/10.1371/journal.pdig.0000416
7. Ji Z, Lee N, Frieske R, et al. Survey of Hallucination in Natural Language Generation. ACM Comput Surv. 2023;55(12):1-38. https://doi.org/10.1145/3571730
8. Lewis P, Perez E, Piktus A, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Adv Neural Inf Process Syst. 2020;33:9459-9474. https://doi.org/10.48550/arXiv.2005.11401
9. Gao Y, Xiong Y, Gao X, et al. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997. 2023. https://doi.org/10.48550/arXiv.2312.10997
10. Zakka C, Shad R, Chaurasia A, et al. Almanac—Retrieval-Augmented Language Models for Clinical Medicine. NEJM AI. 2024;1(2). https://doi.org/10.1056/AIoa2300068
11. Xiong G, Jin Q, Lu Z, Zhang A. Benchmarking Retrieval-Augmented Generation for Medicine. Findings of ACL 2024. 2024:6233-6251. https://doi.org/10.18653/v1/2024.findings-acl.372
12. Jia Y, Ji X, Wang X, et al. Qibo: A Large Language Model for Traditional Chinese Medicine. Expert Syst Appl. 2025;284:127672. https://doi.org/10.1016/j.eswa.2025.127672
13. Yue W, Wang X, Zhu W, et al. TCMBench: A Comprehensive Benchmark for

Evaluating Large Language Models in Traditional Chinese Medicine. arXiv:2406.01126. 2024.

https://doi.org/10.48550/arXiv.2406.01126

14. He J, Guo Y, Lam LK, et al. OpenTCM: A GraphRAG-Empowered LLM-based System for Traditional Chinese Medicine Knowledge Retrieval and Diagnosis. arXiv:2504.20118. 2025.

https://doi.org/10.48550/arXiv.2504.20118

15. 한국한의학연구원, 온톨로지 기반 한의학 지능형 정보체계 연구, 2013년 연차보고서, 2013.

16. 한국한의학연구원. 오아시스, Available from: https://oasis.kiom.re.kr (accessed 2026-09-01).

17. 대한한의학회. 대한한의학회 표준한의학용어집. Available from: https://cis.kiom.re.kr/terminology (accessed 2026-09-01).

18. 한국한의약진흥원. 한의표준임상진료지침. Available from: https://nikom.or.kr/nckm/module/practiceGuide/index.do?menu_idx=14&gubun=INT&code_gubun=mds (accessed 2026-09-01).

19. Qwen Team. Qwen3-Embedding-8B. Available from: https://huggingface.co/Qwen/Qwen3-Embedding-8B (accessed 2026-09-01).

20. Milvus. Milvus Documentation. Available from: https://milvus.io/docs/overview.md (accessed 2026-09-01).

21. Robertson S, Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond. Found Trends Inf Retr. 2009;4(1-2):1-174. https://doi.org/10.1561/1500000019

22. Karpukhin V, Oguz B, Min S, et al. Dense Passage Retrieval for Open-Domain Question Answering. Proc EMNLP 2020. 2020:6769-6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

23. Cormack GV, Clarke CLA, Buettcher S. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. Proc SIGIR 2009. 2009:758-759. https://doi.org/10.1145/1571941.1572114

24. Beijing Academy of Artificial Intelligence. BAAI/bge-reranker-v2-m3. Available from: https://huggingface.co/BAAI/bge-reranker-v2-m3 (accessed 2026-09-01).

25. Qwen Team. Qwen3.8-27B. Available from: https://huggingface.co/Qwen/Qwen3.8-27B (accessed 2026-09-01).