의료 LLM의 Hallucination 완화를 위한 RAG 동향 연구

소지연‡, 한지원‡, 이연준*
한양대학교 ERICA, 한양대학교 *한양대학교

‡thwldus03@hanyang.ac.kr, ‡jwhero12@hanyang.ac.kr, *yeonjoonlee@hanyang.ac.kr

A Study on the RAG trend for mitigating hallucination in medical LLM

So Ji Youn‡, Han Ji Won‡, Kim Yoo Shin*
‡Hanyang University ERICA, ‡Hanyang University, *Hanyang University

요 약

대규모 언어 모델(LLM)은 다양한 분야에서 활용되며, 특히 의료 분야에서는 생성 정보의 신뢰성이 진료와 치료에 직접적인 영향을 미쳐 큰 주목을 받고 있다. 그러나 그럴듯하지만 사실과 다른 정보를 생성하는 Hallucination 현상은 여전히 큰 문제로 지적된다. 본 논문은 이를 완화하기 위한 방안으로 RAG의 의료 분야 적용 사례를 질의응답, 의료·임상 데이터 처리, 질병 맞춤형 가이드의 세 가지로 나누어 고찰하였다. 반복적 검색, 멀티모달 데이터 활용, 이중 검색 구조 등 다양한 기법을 통해 Hallucination을 효과적으로 줄인 사례들이 소개되었다. 다만 검색 정보의 제한이나 관련성 부족은 여전히 오류 가능성을 내포하고 있으며, 향후에는 보다 정교한 검색 전략과 도메인 지식 통합을 통해 LLM의 응답 신뢰도를 높이는 연구가 요구된다.

Ⅰ. 서 론

최근 LLM(Large Language Model)은 뛰어난 자연어 처리 능력을 바탕으로 번역, 질의응답, chatbot특히 금융, 의료, 법률, 교육과 같이 고도의 전문성이 요구되는 영역에서도 LLM의 도입과 적용이 빠르게 확산되고 있다. 그러나 사용자 조사 결과, Hallucination에 대한 우려가 가장 두드러졌으며, 그림 1에서 확인할 수 있듯이 중국어권과 영어권 사용자 모두 이를 LLM 인터페이스에서 가장 주요한 문제로 인식하고 있음을 알 수 있다[1].

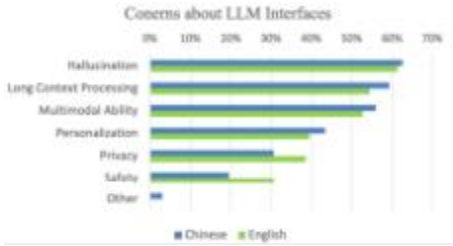

그림 . Wang et al. (2024)에서 중국어권과 영어권 사용자의 LLM 인터페이스에 대한 우려 조사.

Hallucination은 실제로 존재하지 않거나 사실과 다른 정보를 생성하는 문제로, 특히 진료와 치료에 직접적인 영향을 미치는 의료 환경에서는 그 위험성이 더욱 크다. 이를 해결하기 위한 방법 중 하나로 RAG(Retrieval-Augmented Generation)가 주목받고 있다.

본 논문에서는 의료 LLM에서 Hallucination를 완화하려는 최근 RAG 연구 동향을 살펴보고, 향후 개선 방향을 논의하고자 한다.

Ⅱ. 본론

RAG는 LLM이 답변 생성 시 외부에서 검색한 지식을 활용함으로써, 모델의 재학습 비용을 절감하고 최신 정보를 반영할 수 있도록 하는 기법이다. 질의와 관련된 부족한 정보를 보완할 수 있어 LLM의 Hallucination 문제를 완화할 수 있다.

그림 2. LLM의 Hallucination 완화를 위한 RAG 활용의 의료 분야 적용 개요도

Hallucination 현상 감소를 목표로 하는 RAG의 활용은 [1]질의응답, [2]의료·임상 데이터 처리, [3]질병 맞춤형 가이드라는 세 가지 주요 분야로 나눌 수 있다.

A. 질의응답 (QA)

RAG 기반 질의응답 시스템은 chatbot을 활용하여 건강상태를 묻는 개인적인 수준과 의료 관련 시험문제에 답하는 등의 전문적인 수준으로 구분할 수 있다.

개인 건강 정보와 같은 일상적인 질의에 대응하기 위해, COVID-19 관련 문서를 기반으로 RAG를 활용하여 정보를 수집하고 Knowledge base에 추가하는 연구가 나타났다[2]. 이를 통해 도메인 내 핵심 속성과 개체 간 상호작용을 지식 그래프 형태로 구조화하여, 최신 의료 지식에 대한 접근성과 질의 맥락에 따른 정확성을 향상시키고 Hallucination을 완화한다.

보다 전문적인 의학 지식이 요구되는 영역에서도 RAG의 적용 가능성을 검토한 연구들이 활발히 진행되고 있으며, Murali 외[3]는 미국과 인도의 의사면허시험 기반 객관식 질의를 다루기 위해 ReMAG-KR를 제안했다. 의료 질문에서 핵심 키워드를 추출한 후, Dense Retriever와 Cross-Encoder re-ranker를 통해 관련성이 높은 문서를 검색 및 선별하고, 이를 LLM에 결합하여 답변을 얻는다. 광범위한 의학 지식과 데이터를 활용하여 답변을 생성함으로써, Hallucination이 감소하고 추론시간 또한 개선되는 것으로 나타났다. 최근에는 여러 차례의 정보 탐색이 필요한 복잡한 문제를 해결하기 위해, 반복적 질의로 다단계 정보탐색을 수행하는 i-MedRAG가 제안되었다[4]. 초기 질문에 대해 단일 검색만 수행하는 기존

RAG와 달리, LLM이 이전 정보 탐색 결과로 후속 질의를 생성하고, 각 질의에 대해 RAG를 반복 수행함으로써 복잡한 의학 문제를 점진적으로 해결한다. 이는 MedQA 데이터셋에서는 zero-shot 설정 하에 최신 prompt engineering 및 fine-tuning 기법보다도 높은 정확도를 달성하였으며, 기존 RAG 기반 접근법 대비 USMLE 및 MMLU 등의 벤치마크에서 우수한 성능을 보였다.

B. 의료·임상 데이터 처리

수집된 의료 및 임상 데이터를 통합·요약·정제하는 등에 LLM을 활용할 수 있다. 이러한 접근은 복잡하고 비정형적인 임상 정보를 구조화하거나 환자 중심의 요약 또는 보고서 생성을 자동화하는 데 기여하고 있다.

의료 멀티모달[1] LLM(Med-MLLM)의 Hallucination 문제를 완화하기 위해, V-RAG(Visual RAG)를 제안하였다[5]. 유사한 의료 이미지를 검색하여 해당 이미지와 연관된 텍스트 정보를 통합함으로써, 모델의 응답을 시각적 근거에 기반하도록 유도한다.

EHR[2]은 의료진의 임상적 의사결정과 환자 맞춤형 진료에 핵심적으로 활용된다. Saba 외[6]은 EHR의 요약을 위한 RAG 프레임워크를 제안하여, EHR 데이터를 분할·벡터화한 후 의료 전문가가 선정한 핵심 질문에 대한 응답을 추출하는 방식으로 요약을 수행한다. 이를 통해 LLM의 Hallucination 문제를 완화하고 중복 없이 핵심 정보를 담은 요약 생성을 가능하게 한다.

C. 질병 맞춤형 가이드

LLM은 특정 증상이나 질환에 기반한 환자 맞춤형 진료 및 관리 지침 제공하여, 전문가의 임상적 의사결정을 지원하고 정밀의료의 실현 가능성을 높일 수 있다.

한국당뇨병학회(KDA)와 미국당뇨병학회(ADA)의 최신 진료 지침을 기반으로 Hallucination을 최소화하기 위해, RAG의 검색 및 생성 구조를 확장한 이중 검색 기반 RAG 프레임워크가 제안되었다[7]. OpenAI, Upstage 등의 임베딩 모델을 활용한 Dense Retrieval과 언어 특성을 반영한 Sparse Retrieval를 통해 핵심 문서를 효과적으로 선별한다. 이중 검색 결과를 컨텍스트로 통합하여 최신 당뇨병 진료 지침을 주기 때문에, 한국어 및 영어 환경에서 Hallucination을 현저히 줄이고 신뢰할 수 있는 AI 기반 의료 응용 프로그램을 위한 기초를 마련했다.

미국종합암네트워크(NCCN)의 가이드라인에 따라 유방암 환자에게 두 가지 RAG 기반 접근인 Agentic-RAG와 Graph-RAG를 활용해 맞춤형 치료 계획을 제공하기도 하였다[8]. Agentic-RAG는 LLM이 임상 제목을 선택하고 관련 JSON 정보를 반복적으로 탐색하는 방식이며, Graph-RAG는 JSON 데이터를 요약해 치료 관계를 그래프로 구성하고 LLM에 질의하여 권고안을 생성한다. 이를 통해 NCCN 문서와 관련된 치료 권장 사항을 제공했으며, 모두 Hallucination 현상을 나타내지 않았다.

Ⅲ. 결론

본 논문에서는 의료 분야에서 LLM의 Hallucination 문제를 완화하기 위한 RAG의 주요 활용 분야와 기술적 접근을 살펴보았다. 최근 RAG는 단순한 일회성 검색을 넘어서, 반복적 검색, 멀티모달 데이터 활용, 이중 검색 구조 등으로 고도화되며 의료 응용에서의 정확성과 신뢰성을 높이고 있다. 그러나 여전히 검색된 문서의 양적·질적 한계나 과도한 검색으로 인한 비관련 정보의 참조 등으로 인해, 잘못된 응답이 생성될 위험은 존재한다[9]. 향후 의료 도메인에 특화된 검색 전략과 지식 통합 기술이 더욱 발전함에 따라, RAG 기반 의료 LLM의 응답 품질은 지속적으로 향상될 것으로 기대된다.

[1] **멀티모달(Multimodal)**
텍스트, 음성, 이미지, 영상 등 다양한 형태의 데이터를 동시에 처리하거나 이해하는 기술 또는 시스템

[2] **EHR (Electronic Health Records)**
환자의 진료 이력, 검사 결과, 처방 내역 등 다양한 임상 데이터를 포함하는 디지털 기록

ACKNOWLEDGMENT

이 논문은 2025년도 정부 재원(과학기술정보통신부 공학연구팀제 지원사업)으로 과학기술정보통신부와 한국여성과학기술인육성재단의 지원을 받아 수행되었습니다 (WISET 계약 제2025-209호). 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 융합보안핵심인재양성사업의 연구 결과로 수행되었음" (IITP-2024-RS-2024-00423071)

참 고 문 헌

[1] Wang J., Ma W., Sun P., Zhang M., Nie J.-Y., "Understanding User Experience in Large Language Model Interactions," arXiv:2401.08329, 2024.

[2] S. K. S., J. W. K. G., G. M. K. E., M. R. J., Singh A. R. G., Y. E., "A RAG-based Medical Assistant Especially for Infectious Diseases," 2024 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, pp. 1128-1133, 2024.

[3] Murali S., Sowmya S., Supreetha R., "ReMAG-KR: Retrieval and Medically Assisted Generation with Knowledge Reduction for Medical Question Answering," Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics (Vol. 4: Student Research Workshop), pp. 62-67, Bangkok, Thailand, 2024.

[4] Xiong G., Jin Q., Wang X., Zhang M., Lu Z., Zhang A., "Improving Retrieval-Augmented Generation in Medicine with Iterative Follow-up Questions," Pacific Symposium on Biocomputing, vol. 30, pp. 199-214, 2025.

[5] Chu Y.-W., Zhang K., Malon C., Min M. R., "Reducing Hallucinations of Medical Multimodal Large Language Models with Visual Retrieval-Augmented Generation," Proc. of the GenAI4Health Workshop at AAAI 2025.

[6] Saba W., Wendelken S., Shanahan J., "Question-answering based summarization of electronic health records using retrieval augmented generation," arXiv:2401.01469, 2024.

[7] Lee J., Cha H., Hwangbo Y., Cheon W., "Enhancing Large Language Model Reliability: Minimizing Hallucinations with Dual Retrieval-Augmented Generation Based on the Latest Diabetes Guidelines," Journal of Personalized Medicine, vol. 14, no. 12, p. 1131, 2024.

[8] Mohammed A. M., Mansoor I., Blythe S., Trujillo D., "Developing an Artificial Intelligence Tool for Personalized Breast Cancer Treatment Plans based on the NCCN Guidelines," arXiv:2502.15698, 2025.

[9] Xia P., Zhu K., Li H., Zhu H., Li Y., Li G., Zhang L., Yao H., "RULE: Reliable Multimodal RAG for Factuality in Medical Vision Language Models," Proc. of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 1081-1093, Miami, Florida, USA, 2024