대한내과학회지: 제100권 제5호 2025 eISSN 2289-0769
What's hot? https://doi.org/10.3904/kjm.2025.100.5.219

의료 현장 맞춤형 거대 언어 모델 활용 전략과 데이터 보안 대응 방안 탐색

울산대학교 의과대학 서울아산병원 정보의학과

전태준

Tailored Strategies for Applying Large Language Models in Clinical Settings and Addressing Data Security Challenges

Tae Joon Jun

Department of Information Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

Since the advent of ChatGPT in 2022, large language models (LLMs) have rapidly evolved, and their clinical applications are currently being explored. This paper introduces three practical strategies for applying LLMs in healthcare settings: text-to-text, any-to-text, and retrieval-augmented generation. Each strategy is described using real-world examples and analyzed for potential data security risks. Although LLMs offer promising efficiency and performance benefits, they also pose new challenges regarding privacy and information leakage, particularly when trained using sensitive patient data. We propose tailored learning and governance approaches to mitigate such risks, emphasizing the necessity of de-identification techniques and robust guardrails for ensuring safe and effective deployment in clinical settings. **(Korean J Med 2025;100:219-224)**

Key Words: Large language models; Computer security; Generative artificial intelligence

서 론

2022년 OpenAI (San Francisco, CA, USA)의 ChatGPT 등장 이후 거대 언어 모델(large language model, LLM) 기술이 빠르게 발전하고 있다. 2025년 현재 ChatGPT뿐만 아니라 Claude (Anthropic PBC, San Francisco, CA, USA) Gemini (Google, Mountain View, CA, USA), Grok3 (xAI, Palo Alto, CA, USA) 등 다양한 상용 LLM이 서비스 중이다. 그러나 이들 상용 모델은 모델의 가중치를 공개하지 않아 연구나 상업적 활용에 제약이 있다. 이러한 한계를 극복하기 위하여 연구 및 상업적 활용이 가능한 오픈소스 LLM 개발 또한 활발히 이루어지고 있으며 대표적인 사례로 LLaMA (Meta Platforms, Menlo Park, CA, USA), Qwen (Alibaba Cloud, Singapore), Mistral (Mistral AI, Paris, France) 등이 있다.

의료 데이터 기반 LLM은 이러한 오픈소스 모델을 바탕으로 의료 분야의 특화된 데이터를 활용하여 미세 조정(fine-tuning) 과정을 거쳐 학습된다. 주로 활용되는 의료 데이터는 PubMed의 논문 초록(abstract)과 미국의사면허시험(United

Received: 2025. 5. 27 **Revised:** 2025. 7. 24 **Accepted:** 2025. 7. 28

Correspondence to:
Tae Joon Jun, Ph.D.
Department of Information Medicine, Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea
Tel: +82-2-3010-8642, Fax: +82-2-3010-8642, E-mail: taejoon@amc.seoul.kr

Copyright©2025 The Korean Association of Internal Medicine
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted noncommercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

States Medical Licensing Examination, USMLE)의 문제들이며 ChatGPT와 같은 상용 LLM이 생성한 합성 데이터도 널리 활용된다. 현재 가중치가 공개되어 있는 대표적인 의료 특화 LLM으로는 Med42 [1], Clinical Camel [2], MediTron [3], MedPaLM [4] 등이 있다. 그러나 이러한 모델들은 주로 USMLE 점수 또는 소수의 임상의 평가를 기반으로 성능 평가가 이루어지고 있어 실제 의료 현장의 다양한 업무에 직접 적용하기에는 여전히 여러 가지 한계점을 안고 있다.

따라서 실제 의료 현장에서 LLM을 효과적으로 활용하려면 병원 내부의 데이터를 이용하여 특정 목적에 최적화된 형태로 모델을 학습하는 전략이 필요하다. 본 논문에서는 이러한 목적 특화 전략을 크게 세 가지로 구분하였다. 첫째 text-to-text 학습 전략, 둘째, any-to-text 학습 전략, 셋째, 검색 증강 생성(retrieval-augmented generation, RAG) 활용 전략이다. 이어지는 본문에서는 각 전략의 구체적인 방법과 실제 적용 사례를 소개하고 이 과정에서 발생할 수 있는 의료 데이터 보안 이슈와 이에 대한 대응 방안을 함께 논의하고자 한다.

본 론

Text-to-text 학습 전략

Text-to-text 학습 전략은 텍스트 형태의 데이터를 입력받아 LLM이 텍스트 형태의 결과물을 출력하는 방식이다. 예를 들어 병원에서 흔히 사용하는 입퇴원기록지, 외래경과기록지 등 다양한 임상노트를 LLM에 입력으로 제공하면 모델은 이를 바탕으로 요약된 내용을 생성하거나 새로운 임상기록을 작성해 준다. 이러한 방식은 방대한 분량의 의료 기록에서 중요한 임상 정보만을 간략하게 정리하여 의료진이 빠르게 정보를 파악하고 활용할 수 있도록 도와준다. 또한 새로운 기록지 작성 시 의료진이 초안을 작성하는 데 들어가는 시간을 크게 절약할 수 있어 업무의 효율성도 높일 수 있다[5].

학습 과정은 크게 두 단계로 나누어진다. 첫 번째 단계에서는 병원의 임상노트를 비롯한 다양한 의료 데이터를 활용하여 LLM의 의료 분야 이해도를 높이는 학습을 수행한다. 이 단계에서 주로 활용되는 기반 모델은 Llama, Qwen, Mistral 등 오픈소스 LLM이다. 임상노트는 학습 전 미리 지정한 크기(chunk)로 나누어 처리되며 학습 방법으로는 continued pre-training이 사용된다. Continued pre-training이란 이미 기본적인 학습을 마친 LLM에 의료와 같이 특정 분야의 전문 지식을 추가로 주입하는 과정이라고 이해할 수 있다. 이 학습 방식은 모델 전체의 가중치(parameter)를 대상으로 추가 학습을 수행하므로 막대한 계산 자원과 긴 학습 시간이 필요하다. 또한 추가 학습하는 분야의 지식이 기존에 학습된 지식과 크게 다를 경우 기존 지식을 일부 잃어버리는 파국적 망각(catastrophic forgetting) 현상이 발생할 수도 있다. 따라서 최종적으로 수행할 작업의 복잡성이 높지 않다면 첫 번째 단계인 continued pre-training 단계를 생략하고 바로 다음 단계인 supervised fine-tuning (SFT) 단계로 넘어가는 전략도 고려해 볼 수 있다.

SFT는 LLM이 특정 목적의 업무를 더 잘 수행하도록 세부적으로 조정하는 학습 과정이다. SFT를 수행하기 위해서는 목표로 하는 업무에 맞추어 데이터셋을 구축해야 한다. 예를 들어 퇴원기록지를 생성하는 모델을 구축한다고 할 때 입력 데이터(question)는 환자의 기본 정보, 진단명, 치료 과정 등 퇴원기록지 생성을 위한 필수 정보와 이를 지시하는 프롬프트(prompt)를 포함하고 있어야 한다. 그리고 해당 입력에 대한 이상적인 출력(answer)으로 완성된 퇴원기록지를 준비한다.

이렇게 question과 answer가 짝지어진 데이터셋을 구축한 후 이를 이용하여 SFT 학습을 수행한다. 일반적으로 SFT는 LLM 전체 파라미터가 아닌 일부 파라미터만을 대상으로 학습한다. 이는 주로 LoRA [6], QLoRA [7]와 같은 기법을 사용하여 이루어지는데 이 방법들은 원본 LLM의 대부분의 파라미터를 고정(freeze)하고 adapter라 불리는 소수의 추가적인 파라미터만을 학습하여 원본 LLM에 결합하는 방식이다. 이 방법을 통해 학습에 필요한 계산 자원을 절약하면서도 목적에 특화된 성능 향상을 효과적으로 달성할 수 있다.

Text-to-text 학습 전략은 기존의 다양한 의료 인공지능 모델 개발 방식과 유사하기 때문에 비교적 학습의 난이도가 낮고 그 과정 또한 직관적이다. 따라서 필자는 현재 의료 현장에서 가장 손쉽게 도입할 수 있는 전략이라고 생각한다. 그러나 Text-to-text 학습 전략을 사용할 때 반드시 고려해야 할 중요

한 보안 이슈가 있다. 그것은 바로 학습 데이터가 LLM을 통해 외부로 유출될 수 있다는 점이다. Text-to-text 전략은 병원의 텍스트 정보를 직접 LLM에 학습시키는 방식이기 때문에 만약 학습 데이터에 환자의 개인 정보나 병원의 민감한 정보가 포함될 경우 LLM이 학습된 내용을 답변하는 과정에서 외부로 노출할 가능성이 존재한다[8]. 특히 기존의 전통적인 의료 정보 보안은 환자 개인 정보에 대한 접근 권한을 역할에 따라 차등적으로 부여할 수 있지만 LLM 내에 이미 학습된 데이터는 접근 권한을 별도로 차등화하기 어렵다는 기술적 한계가 있다. 따라서 환자의 개인 정보가 꼭 필요하지 않은 task라면 학습 데이터에서 미리 개인 정보를 삭제하거나 가명화 처리하는 과정이 필수적이다. 한편 만약 환자 개인 정보를 포함하여 학습을 진행해야 하는 task라면 LLM이 생성하는 답변 및 질의에 대해 엄격한 가드레일(guardrail)을 설정하는 방법을 고려할 수 있다. 가드레일이란 모델이 특정 정보를 언급하거나 외부로 노출하지 않도록 사전에 정한 규칙과 제한을 의미한다.

Any-to-text 학습 전략

Any-to-text 학습 전략은 다양한 형태의 데이터를 입력받아 텍스트 형태로 결과를 생성하는 방식이다. 이 전략에서 입력 데이터는 영상(image), 생체 신호(signal), 텍스트(text) 등 매우 다양한 형태의 의료 데이터가 될 수 있으며 단일 형태의 데이터뿐 아니라 두 가지 이상의 형태를 결합한 멀티모달(multimodal) 데이터도 가능하다.

대표적인 예로 흉부 방사선 촬영(chest posteroanterior [PA]) 이미지를 입력받아 이상 소견을 자동으로 판독하고 그 결과를 텍스트로 기술하는 업무를 들 수 있다[9]. 병리조직 슬라이드 이미지의 경우에도 조직의 상태를 분석한 후 조직학적 진단 결과를 텍스트로 제시할 수 있다[10]. 또한 심전도(electrocardiogram, ECG)와 같은 생체신호 데이터를 입력받아 심장 리듬 이상을 탐지하고 이를 텍스트로 설명해주는 업무 역시 any-to-text 학습 전략의 대표적인 활용 사례이다[11].

이러한 업무들은 의료진이 영상 및 신호 데이터를 분석할 때 보통 이미지나 신호 데이터만으로 판단하지 않고 임상적 정보(환자의 병력이나 검사 목적 등)가 담긴 텍스트를 함께 참고한다. 따라서 any-to-text 학습 전략은 이미지나 신호 데이터와 임상적 정보 텍스트를 함께 활용하여 더 정확하고 풍부한 의료적 판단을 내리는 데 도움을 준다. 의료진이 다루는 다양한 형태의 자료를 동시에 고려하여 보다 통합적이고 정확한 판독 결과를 제공하는 데 효과적이다.

Any-to-text 학습 전략의 학습 단계는 앞서 설명한 text-to-text 학습 전략과 기본적으로 유사한 과정으로 구성된다. 하지만 any-to-text 전략은 입력 데이터의 형태(modality)가 다양하므로 데이터의 modality에 따라 선택하는 모델 및 전처리 방법에서 차이가 발생한다. 예를 들어 입력이 이미지와 텍스트를 함께 포함하는 멀티모달 데이터라면 텍스트만을 다루는 기존 모델과는 달리 이미지 데이터를 추가로 인식하고 처리할 수 있는 LLM을 활용해야 한다.

대표적으로 이미지와 텍스트 데이터를 함께 입력으로 받아 텍스트 형태의 결과를 생성하는 경우 Llama-4-Maverick (Meta Platforms)이나 Qwen-VL (Alibaba Cloud)과 같이 이미지 입력이 가능한 멀티모달 전용 모델을 활용할 수 있다. 또한 기존의 텍스트 기반 LLM 모델 중에서도 이미지 데이터 입력을 추가적으로 지원하도록 개발된 Gemma-3 (Google), Mistral 등의 모델도 활용 가능하다. 이러한 멀티모달 전용 또는 지원 모델들은 이미지 데이터를 효과적으로 처리하기 위하여 이미지 인코딩(image encoding) 모델을 LLM과 결합한 형태를 갖추고 있다. 이미지 인코딩 모델로는 주로 vision transformer [12]와 같은 시각 인지 모델을 사용하며 이를 통해 이미지 데이터에서 시각적 특징(visual feature)을 추출하여 텍스트 정보와 함께 LLM에 제공한다. LLM은 이미지와 텍스트에서 얻은 정보를 동시에 고려하여 최종 텍스트 출력 결과를 생성하게 된다.

Any-to-text 전략은 기존의 의료 인공지능 서비스들이 영상(image)이나 생체 신호(signal) 데이터를 입력으로 받아 다양한 진단 및 예측을 수행하는 형태에서 한 단계 더 나아가 판독 결과를 자연어 형태의 텍스트로 확장하려는 시도이다. 예를 들어 chest PA 이미지에서 폐 결절을 자동으로 탐지하거나 자기공명영상(magnetic resonance imaging) 및 컴퓨터단층촬영(computed tomography)에서 병변을 자동으로 분할하는 인공지능 소프트웨어에 판독 소견을 텍스트 형태로 생성하

는 기능을 추가하여 더 고도화된 의료 서비스로 발전시킬 수 있다. 또한 ECG 신호에서 심방세동(atrial fibrillation)과 같은 비정상적인 파형을 탐지하는 기존의 신호 처리 인공지능 소프트웨어에도 텍스트 기반의 상세 판독 결과를 제공하는 기능이 추가될 수 있다.

그러나 앞서 소개한 text-to-text 전략에 비해 any-to-text 전략은 기술적으로 구현 난이도가 상당히 높다. 이는 영상이나 신호 데이터를 이해하고 해석하여 텍스트 형태로 결과를 생성해야 하기 때문이다. 또한 텍스트 데이터와 달리 영상-텍스트, 신호-텍스트와 같이 멀티모달 형태로 구성된 공개 데이터셋이 절대적으로 부족하기 때문에 데이터 확보와 학습 데이터 구축에 큰 어려움이 있다. 이러한 이유로 any-to-text 전략이 의료계에 빠르게 도입되기는 쉽지 않으며 실제 현장에서 안정적으로 활용하기 위해서는 앞으로 더 많은 연구와 데이터 구축이 선행되어야 할 것으로 사료된다.

검색 증강 생성 활용 전략

RAG 기술은 앞서 설명한 text-to-text 또는 any-to-text 학습 전략과는 개념적으로 차이가 있다. Text-to-text와 any-to-text 전략은 기본적으로 병원 내의 임상 데이터(텍스트, 영상, 신호 등)를 이용하여 LLM을 추가로 학습시키고 이렇게 학습된 모델이 특정 업무에 대한 답변이나 진단 결과를 직접 생성하는 방식을 사용한다. 즉 모델 자체가 의료 지식을 저장하고 있는 형태라고 볼 수 있다.

반면 RAG 기술은 LLM 자체를 의료 지식을 저장하는 저장소로 사용하는 것이 아니라 오히려 사용자가 제공하는 외부 데이터베이스(database)나 자료를 참조하여 답변을 생성하는 방식이다. 여기서 LLM은 사용자가 질문하였을 때 필요한 정보를 외부 데이터에서 직접 검색(retrieval)하여 가져온 후 이를 바탕으로 자연스러운 형태로 재구성(generation)하여 텍스트 결과를 출력하는 일종의 생성 엔진 역할을 한다.

따라서 RAG 방식을 사용할 때는 모델 학습 단계에서 방대한 양의 의료 데이터를 모두 학습시키기보다는 별도의 의료 문

Table 1. A comparative analysis of strategies for utilizing customized large language models in clinical settings

Classification	Text-to-text	Multimodal-to-text	RAG
Concept	Generates text from a text input	Generates text from various data inputs, such as images and biosignals	Searches an external database and generates answers based on the results
Data handling	The model is trained on and stores medical data directly in text format	The model is trained on and stores various forms of medical data (multimodal)	The model itself is not used as a knowledge repository; it references an external DB
Use cases	Summarizing and drafting clinical records	Medical image interpretation Biosignal analysis and explanation	Answering questions based on the latest clinical guidelines Identifying candidates for clinical trials Searching academic papers and Q&A
Technical difficulty	Relatively low and straightforward	High technical implementation difficulty and challenges in data acquisition	Lower training burden on the model itself, but requires technology for DB construction and integration
Security issues	Risk of sensitive information from training data being leaked through the model Difficulty in implementing differentiated access controls for the learned information	Includes all security issues of the text-to-text model Increased complexity due to diverse data types (multimodal)	Risk of information theft through direct attacks on the vector DB Possibility of inferring sensitive information from questions and answers
Mitigation strategies	Data de-identification and pseudonymization Implementing guardrails on the model	Apply the mitigation strategies from the text-to-text approach	DB access control Strengthening traditional database security measures

RAG, retrieval-augmented generation; DB, database.

서 데이터베이스나 병원 내부 자료를 구성해 두고 실제 답변 과정에서 이 데이터베이스를 검색하여 필요한 정보를 즉시 가져온다. 예를 들어 특정 환자에 대한 질병 진단 및 치료 계획에 대한 질문이 입력되면 RAG 방식의 모델은 먼저 병원에서 관리하는 최신 진료지침이나 관련 논문, 기존 환자 기록 등을 참조하여 가장 적합한 정보를 찾아내고 이를 이용하여 자연스러운 텍스트 형태로 답변을 생성하게 된다[13].

RAG 기술을 활용하여 병원 내 환자 데이터를 기반으로 LLM에 질문하고 답변을 받는 과정을 구체적으로 예시를 들어 설명하면 다음과 같다. 우선 병원에서 활용하고자 하는 환자 데이터를 데이터베이스 형태로 구축한다. 이때 구축하는 데이터는 환자의 전자의무기록(electronic medical record, EMR)과 같이 표로 정리된 구조화된 데이터일 수도 있고 의료 영상이나 임상노트와 같은 비정형 데이터도 함께 포함될 수 있다.

이처럼 다양한 형태의 데이터는 임베딩(embedding)이라는 과정을 거쳐 수치로 이루어진 벡터 형태로 변환된 후 벡터 데이터베이스(vector database)에 저장된다. 임베딩이란 데이터를 의미적으로 표현하는 수치 벡터로 변환하는 과정을 의미하며 유사한 의미를 가진 데이터들은 벡터 공간 내에서 가까운 위치에 배치되고 반대로 관련성이 낮거나 상반된 의미의 데이터들은 벡터 공간 내에서 멀리 떨어진 위치에 배치된다.

사용자가 참조하고자 하는 데이터를 벡터 데이터베이스에 저장한 후 실제 활용 단계에서 사용자는 LLM을 통해 원하는 질문을 입력한다. 예를 들어 사용자가 “오늘 내원한 A 환자의 과거 투약 기록을 모두 보여줘”라고 질문하였다고 가정해 보자.

이때 RAG 기술은 사용자의 질문을 먼저 벡터 형태로 변환하고 이 질문 벡터를 벡터 데이터베이스에 저장된 환자 데이터 벡터들과 비교하여 의미적으로 가장 유사한 데이터를 검색한다. 이는 인터넷 검색 시 입력한 검색어와 가장 관련성 높은 결과를 찾아 제공하는 검색 엔진의 작동 방식과 유사하다. 그 다음 LLM은 사용자의 질문 벡터와 벡터 데이터베이스에서 찾은 가장 유사한 환자 데이터 벡터를 참고하여 사용자가 이해하기 쉬운 텍스트 형태의 답변을 생성하여 제시한다.

RAG 기술의 가장 큰 장점은 LLM 자체를 환자 데이터로 직접 추가 학습하지 않고도 별도의 의료 데이터를 기반으로 즉시 질의응답이나 데이터 요약 등 다양한 서비스를 제공할 수 있다는 점이다. 따라서 LLM 자체의 성능이 향상될수록 자연스럽게 제공되는 답변의 품질 역시 높아지게 된다. 이러한 이점 때문에 현재 많은 의료 분야에서 RAG 기술을 활용한 다양한 연구가 활발히 진행되고 있다. 대표적인 예로 임상시험 대상자 탐색[14], 의학 논문 검색 및 질의응답 시스템 개발 등을 들 수 있다[15].

RAG 기술은 의료 데이터를 LLM 자체에 직접 학습시키지 않기 때문에 모델 자체가 적대적 공격(adversarial attack)을 받아 환자의 개인 정보가 유출될 가능성은 비교적 낮은 편이다. 다만 환자의 데이터가 직접 저장된 벡터 데이터베이스를 대상으로 한 공격으로 개인 정보가 탈취될 위험성은 여전히 존재하며 이러한 이슈는 오히려 기존의 병원 데이터베이스를 노리는 전통적인 보안 문제와 유사하다[16]. 특히 벡터 데이터베이스는 기존의 관계형 데이터베이스나 NoSQL 데이터베이스에 비해 역사가 짧아 아직 발견되지 않은 데이터베이스 자체의 보안 취약점이 더 많을 수 있다는 점에 주의해야 한다.

또한 비록 LLM 자체에 직접적인 개인 정보가 포함되지 않는다 하더라도 사용자의 질문과 모델의 답변을 통해 간접적으로 민감한 개인 정보를 추출하거나 유추하는 공격이 발생할 수 있다. 이러한 경우 접근 권한을 철저히 제어하는 방식으로 일부 대응이 가능하지만 RAG 기반의 서비스가 아직 의료계에 널리 도입되지 않은 만큼 미처 예상하지 못한 다양한 보안 공격 가능성도 존재할 수 있다는 점을 충분히 염두에 두어야 한다.

결 론

본고에서는 실제 의료 현장에서 LLM을 효과적으로 활용하기 위한 세 가지 대표적인 전략인 text-to-text 학습 전략, any-to-text 학습 전략, RAG 활용 전략에 대해 소개하고 각 전략에서 발생할 수 있는 의료 데이터의 보안 이슈 및 그 대응 방안을 논의하였다(Table 1). 이러한 전략들은 의료 기관이 목표로 하는 업무의 성격과 난이도에 따라 각각 독립적으로 적용하거나 여러 전략을 동시에 활용할 수 있으며 기존에 개발된 의료 인공지능 소프트웨어 및 서비스와 융합하여 보다 강력하고 효과적인 솔루션을 제공할 수 있다.

다만 LLM 기술은 현재도 빠르게 발전하고 있는 만큼 아직

충분히 고려되지 않은 보안 취약점이나 데이터 유출 가능성이 항상 존재한다는 점을 염두에 두어야 한다. 특히 한번 LLM에 학습된 데이터는 형태와 방법에 따라 외부로 유출될 가능성을 완전히 배제하기 어렵다. 이러한 데이터 유출 위험을 줄이기 위한 근본적인 대안으로 외부 클라우드 서비스가 아닌 병원 내부망에 독립된 서버를 구축하여 LLM을 운영하는 온-프레미스(on-premise) 방식을 고려할 수 있다. 현재 모델에 가드레일(guardrail)을 설정하는 등의 다양한 유출 방지 기술들이 연구되고 있지만 무엇보다 근본적인 보안 방안은 처음부터 학습 데이터에 환자 개인 정보와 같은 민감한 정보를 포함하지 않는 것이다. 이를 위해서는 임상 노트와 같은 비정형 텍스트에서 환자의 개인 정보를 효과적으로 제거하거나 가명화(anonymization) 및 비식별화(de-identification)를 수행할 수 있는 기술 개발과 이에 대한 추가 연구가 적극적으로 이루어져야 할 것이다.

중심 단어: 거대 언어 모델; 컴퓨터 보안; 생성형 인공지능

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

FUNDING

None.

AUTHOR CONTRIBUTIONS

TJJ contributed all aspects of the article.

ACKNOWLEDGEMENTS

None.

REFERENCES

1. Christophe C, Kanithi P, Munjal P, et al. Med42 - evaluating fine-tuning strategies for medical LLMs: full-parameter vs. parameter-efficient approaches. In: AAAI 2024 Spring Symposium on Clinical Foundation Models; 2024 Mar 25-27; Stanford, CA. Washington, DC: Association for the Advancement of Artificial Intelligence, 2024.
2. Toma A, Lawler PR, Ba J, Krishnan RG, Rubin BB, Wang B. Clinical camel: an open-source expert-level medical language model with dialogue-based knowledge encoding. arXiv 2023;2305.12031.
3. Chen Z, Cano AH, Romanou A, et al. MEDITRON-70B: scaling medical pretraining for large language models. arXiv 2023;2311.16079.
4. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nat Med 2025;31:943-950.
5. Williams CYK, Subramanian CR, Ali SS, et al. Physician- and large language model-generated hospital discharge summaries. JAMA Intern Med 2025;185:818-825.
6. Hu EJ, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. arXiv 2021;2106.09685.
7. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLORA: efficient finetuning of quantized LLMs. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, eds. 37th Conference on Neural Information Processing Systems; 2023 Dec 10-16; New Orleans, LA. Red Hook (NY): Curran Associates, 2023:10088-10115.
8. Kim M, Kim Y, Kang HJ, et al. Fine-tuning LLMs with medical data: can safety be ensured NEJM AI 2025;2:1-8.
9. Sloan P, Clatworthy P, Simpson E, Mirmehdi M. Automated radiology report generation: a review of recent advances. IEEE Rev Biomed Eng 2024;18:368-387.
10. Lu MY, Chen B, Williamson DFK, et al. A visual-language foundation model for computational pathology. Nat Med 2024;30:863-874.
11. Tang J, Xia T, Lu Y, Mascolo C, Saeed A. electrocardiogram report generation and question answering via retrieval-augmented self-supervised modeling. In: Ashok A, Kissos I, Rasul K, et al. eds. NeurIPS 2024 Workshop on Time Series in the Age of Large Models; 2024 Dec 15; Vancouver. San Diego (CA): NeurIPS, 2024:1-7.
12. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. arXiv 2020;2010.11929.
13. Kresevic S, Giuffrè M, Ajcevic M, Accardo A, Crocè LS, Shung DL. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. NPJ Digit Med 2024;7:102.
14. Ozan U, Shin J, Mailly CJ, et al. Retrieval-augmented generation-enabled GPT-4 for clinical trial screening. NEJM AI 2024;1:1-11.
15. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc 2025;32:605-615.
16. Nazary F, Deldjoo Y, di Noia T. Poison-RAG: adversarial data poisoning attacks on retrieval-augmented generation in recommender systems. In: Hauff C, MacDonald C, Jannach D, eds. 47th European Conference on Information Retrieval (ECIR 2025); 2025 Apr 6-10; Lucca, Italy. Cham: Springer, 2025:239-251.