Review Article

J Korean Soc Radiol 2024;85(5):861-882
https://doi.org/10.3348/jksr.2024.0080
eISSN 2951-0805

JOURNAL of THE KOREAN SOCIETY of RADIOLOGY

Large Language Models: A Comprehensive Guide for Radiologists

대형 언어 모델: 영상의학 전문가를 위한 종합 안내서

Sunkyu Kim, PhD[1,2], Choong-kun Lee, MD[3], Seung-seob Kim, MD[4]*

[1]Department of Computer Science and Engineering, Korea University, Seoul, Korea
[2]AIGEN Sciences, Seoul, Korea
[3]Division of Medical Oncology, Department of Internal Medicine, Yonsei University College of Medicine, Seoul, Korea
[4]Department of Radiology and Research Institute of Radiological Science, Severance Hospital, Yonsei University College of Medicine, Seoul, Korea

Received May 21, 2024
Revised September 18, 2024
Accepted September 21, 2024

***Corresponding author**
Seung-seob Kim, MD
Department of Radiology and Research Institute of Radiological Science, Severance Hospital, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea.

Tel 82-2-2228-7400
Fax 82-2-2227-8337
E-mail k2s0127@yuhs.ac

Large language models (LLMs) have revolutionized the global landscape of technology beyond the field of natural language processing. Owing to their extensive pre-training using vast datasets, contemporary LLMs can handle tasks ranging from general functionalities to domain-specific areas, such as radiology, without the need for additional fine-tuning. Importantly, LLMs are on a trajectory of rapid evolution, addressing challenges such as hallucination, bias in training data, high training costs, performance drift, and privacy issues, along with the inclusion of multimodal inputs. The concept of small, on-premise open source LLMs has garnered growing interest, as fine-tuning to medical domain knowledge, addressing efficiency and privacy issues, and managing performance drift can be effectively and simultaneously achieved. This review provides conceptual knowledge, actionable guidance, and an overview of the current technological landscape and future directions in LLMs for radiologists.

Index terms Natural Language Processing; Large Language Model; Transformer; Radiology; Chatbot; ChatGPT

총론: 대형 언어 모델(Large Language Model)이란?

대형 언어 모델 이전의 자연어 처리(Natural Language Processing) 기술

1950년대 후반에 소개된 Bag-of-Words (이하 BoW, 단어 빈도 모델) 모델은 텍스트 처리를 자동화하려는 초기 시도 중 하나였다(1). BoW는 자주 등장하는 단어가 문서의 주제와 더 큰 연관성과 중요성을 가진다는 가정하에 텍스트 문서를 숫자 벡터로 변환한다.

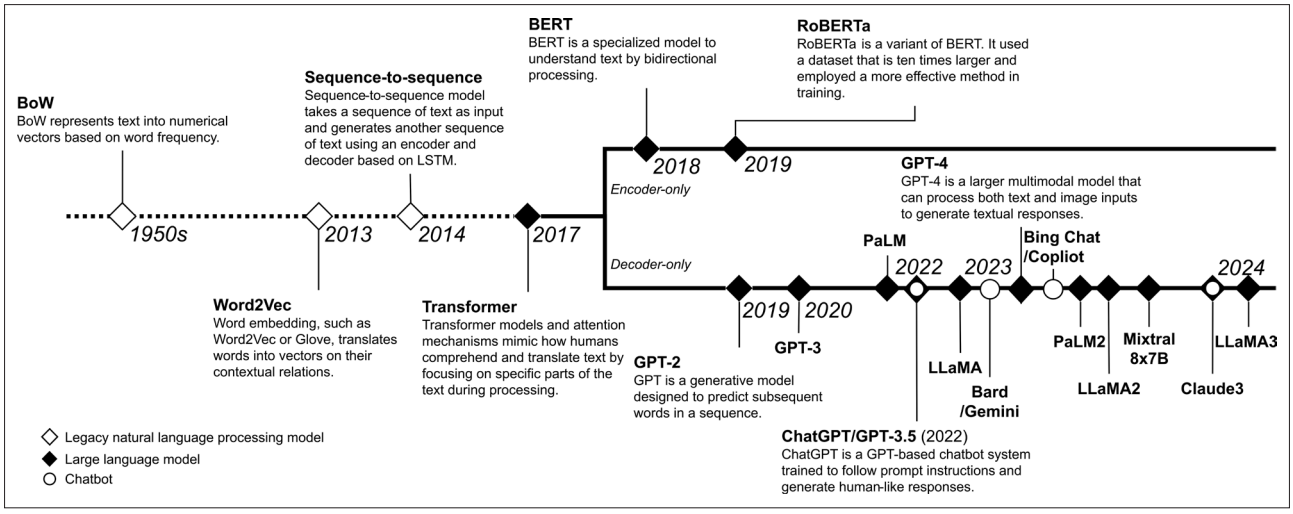
예를 들어, “I love NLP and I love programming”이라는 텍스트를 {“I”:2, “love”:2, “NLP”:1, “and”:1, “programming”:1}와 같은 형태로 표현하였다. 그러나 이 간단한 표현 방식의 주요 한계점은 문장의 맥락을 인식할 수 없었다는 점이다. BoW는 각 단어를 독립적인 요소로만 취급하고, 단어들 사이의 순서나 문맥적 연관성은 고려하지 않는다. 예를 들어, “배가 고프다”와 “배를 타고 여행했다”에서 “배”의 의미는 맥락에 따라 달라지지만, BoW는 이러한 차이를 구분하지 못한다. 결국 언어의 본질을 포착해 내지 못했다는 한계점이 더 발전된 모델들의 등장을 촉진하였다.

J. R. Firth의 유명한 인용구 “you shall know a word by the company it keeps”는 단어 임베딩(word embedding) 기술의 개발로 이어졌다(2). 단어들의 동시 등장 패턴(co-occurrence)을 기반으로 파악한 맥락 관계를 이용하여 단어를 벡터로 변환하였고, 따라서 임베딩은 단어에 대해 더 깊은 의미적 이해를 제공하였다. 예를 들어, “I like you”와 “I love you”라는 두 문장으로부터 “like”와 “love”는 “I”와 “you”라는 단어와 자주 함께 등장한다는 것을 모델이 파악하게 되고, 따라서 두 단어에 유사한 벡터를 부여하게 된다. 이런 식으로 의미적으로 유사한 단어들을 그룹화할 수 있었고, 이는 감정 분석(sentiment analysis)과 같은 작업을 용이하게 하였다. Word2Vec 알고리듬을 사용하여 구현한 임베딩 벡터 공간의 경우, 단어 간의 유추(analogy)도 반영되고 따라서 추상적 의미 간의 재조합도 성립된다는 것이 확인되었다. 예를 들어, “king” 벡터에서 “man” 벡터를 뺀 후 + “woman” 벡터를 다시 더하면 “queen” 벡터와 비슷한 값이 도출됨을 보여주었다. 그럼에도 불구하고, 단어 임베딩 기술은 문장 혹은 문단의 넓은 언어적 뉘앙스를 완전히 포착할 수 없었다. 단어의 임베딩 벡터는 정적으로 생성되기 때문에, 단어의 의미가 여러 맥락에서 다양하게 변화하는 경우는 무시되었다.

순환 신경망(recurrent neural network)은 텍스트 내의 시퀀스, 즉 순서가 있는 데이터를 인식하고 처리하는 데 사용되는 혁신적인 인공 신경망 모델이었다(3). 순환 신경망은 이전 입력의 정보를 일정 기간 동안 유지하면서 입력된 단어의 시퀀스에 따라 정보를 순차적으로 처리할 수 있다. 그러나 길이가 긴 문장을 처리할 때, 시퀀스의 처음 정보가 점차 잊혀지는 ‘장기 의존성 문제(long-term dependency)’가 발생했다. 이 문제를 해결하기 위해 등장한 long short-term memory (이하 LSTM) (4)와 gated recurrent unit (GRU) (5) 모델은 정보를 장기간 기억할 수 있는 아키텍처를 가지고 있어, 전체 문장의 맥락을 보다 효과적으로 파악할 수 있었다.

임베딩 기술과 LSTM 등에 의해 마련된 자연어 처리 기술의 기반 위에서, 인코더-디코더(시퀀스-투-시퀀스) 모델이 번역과 같은 복잡한 자연어 처리 작업을 해결하기 위해 등장하였다(6). 입력 텍스트 시퀀스를 맥락 벡터로 인코딩한 다음 이를 다시 출력 시퀀스로 디코딩함으로써, 이 아키텍처는 인간 대화를 모방하도록, 즉 인간이 질문에 대답을 하는 것처럼 반응이 생성되도록 설계되었다. 인코더와 디코더에는 일반적으로 적층 LSTM이 사용되었으므로, 텍스트를 순차적으로 처리하여 단어 순서와 맥락을 파악할 수 있었다. 예를 들어, “Je t’aime”라는 프랑스어 문장이 주어지면, 인코더-디코더 모델은 영어 번역 “I love you”를 생성할 수 있다. 그러나 LSTM 기반의 인코더-디코더 모델은 순차적으로 텍스트를 처리하기 때문에 문장이 길어질 경우 상당히 오랜 처리 시간이 필요하다는 문제가 있었다. 또한, 문단 수준의 긴 텍스트에 대해서는 인코더에서 임베딩한 정보가 희석되는 등의 문제(information dilution)도 발생하였다. 이에 따라 정보 처리의 효율성을 높이면서 동

Fig. 1. Milestone models leading to modern large language models.



BERT = bidirectional encoder representations from transformers, BoW = Bag-of-Words, GPT = generative pre-trained transformer, LLaMA = large language model meta AI, LSTM = long short-term memory, PaLM = pathway language model, RoBERTa = robustly optimized BERT pretraining approach

시에 장문의 텍스트의 맥락을 유지할 수 있는 새로운 아키텍처의 필요성이 대두되었다(Fig. 1).

대형 언어 모델의 등장

트랜스포머(Transformer) 아키텍처와 대형 언어 모델의 등장

주목(attention) 메커니즘은 모델이 텍스트의 특정 부분에 집중하면서 필요한 정보를 다른 텍스트 부분에서 추출할 수 있도록 하는 기술이다(7). 이 메커니즘을 기반으로 설계된 트랜스포머 모델은 다중 주목 계층(multi-attention layer)을 통해 여러 부분에 동시에 집중할 수 있으며, 이는 텍스트 전체의 맥락을 유지하는 데 있어 LSTM 이상의 효율을 보여주었다. 예를 들어, "It is raining"이라는 영어 구절을 프랑스어 "Il pleut"로 번역할 때, 주목 메커니즘은 "raining"과 "pleut" 사이의 관계를 강조함으로써 전체 영어 구절에 대한 임베딩 벡터 없이 번역에 필요한 부분에만 집중한다. 또한 트랜스포머는 병렬 처리가 가능하므로, 순차적 처리만 가능했던 LSTM과 달리 그래픽 처리 장치(graphics processing unit; GPU)의 이점을 최대한 활용할 수 있었다. 트랜스포머 모델은 이전까지의 자연어 처리 모델과 비교하여 언어 이해 능력이 크게 향상되었고 따라서 더욱 복잡한 문제도 해결 가능하게 되었으나, 동시에 점점 높은 계산량이 요구되기 시작했다.

트랜스포머의 성공을 바탕으로, 양방향 인코더 표현(bidirectional encoder representations from transformers; 이하 BERT)과 생성 사전 훈련 트랜스포머(generative pre-trained transformer; 이하 GPT)와 같은 대형 언어 모델들이 등장하였다(8, 9). 대형 언어 모델은 방대한 양의 텍스트를 학습함으로써 인간과 유사하게 텍스트를 이해하고 생성할 수 있는 고도로 진보된 계산 모델이다. BERT와 GPT는 거대한 규모의 말뭉치에서 사전 훈련을 하고 특정 작업에 대해 미세조정(fine-tuning)을 함으로써, 방대한 데이터셋에서 얻은 지식을 상대적으로 데이터가 제한된 전문

적 응용 프로그램(downstream task)으로 전달(transfer) 할 수 있었다(10). BERT는 33억 단어 수준의 학습 데이터셋, GPT-3는 5000억 토큰(token; 언어 모델이 텍스트 데이터를 처리하고 이해하기 위해 사용하는 기본 단위) 이상의 학습 데이터셋을 이용해 사전 훈련했으며, 이 방대한 지식을 전이 학습(transfer learning) 방법으로 다른 응용문제들을 해결하는 식의 연구들이 활발히 진행되었다. 이렇게 BERT, GPT와 같은 사전 훈련된 대형 언어 모델들, 소위 기반 모델(foundation model)의 등장은 학습 데이터가 적은 분야의 작업을 다루는 방식을 크게 변화시켰다(10). 이러한 패러다임 변화는 자연어 처리 분야에서 매우 중요한 사건이었고, 인간 언어를 더욱 효과적으로 이해하고 생성할 수 있는 모델을 개발하려는 새로운 가능성과 도전을 제공하였다.

BERT 이후로 GPT-2, GPT-3, Pathway Language Model (PaLM), PaLM-2, 그리고 GPT-4에 이르기까지 대형 언어 모델들의 파라미터 개수와 사전 훈련 데이터셋의 크기는 꾸준히 증가하였다. 광범위한 영역의 형식 및 주제, 분야, 언어 등을 모두 아우르는 방대한 데이터셋을 활용한 사전 훈련 덕분에 대형 언어 모델들의 지식 기반은 점점 더 넓어졌고, 그 결과 추가적인 분야별 맞춤 미세조정(domain-specific fine-tuning) 없이도 세부 분야의 전문적인 작업에서 높은 성능을 보이게 되었다(11). 뿐만 아니라 모델이 학습 단계에서 한 번도 명시적으로 배운 적이 없는 작업도 바로(zero-shot learning), 혹은 약간의 예시만으로도(few-shot learning) 수행 가능하게 되었다. 대형 언어 모델의 이러한 성능 도약은 과거의 작은 스케일의 언어 모델에서는 볼 수 없었던 특성이고, 또 모델의 스케일이 증가되는 과정에서 획득되리라고 예측할 수도 없었으며, 스케일이 특정 임계치를 넘어설 때 비선형적인 수준의 비약적인 성능 도약이 이루어진다는 점에서, 소위 창발 능력(emergent ability)의 대표적인 예시로 간주된다(12).

BERT와 GPT의 차이점

BERT는 텍스트를 양방향으로 검토함으로써 단어의 좌우 맥락을 모두 고려하게 되고 따라서 각 단어의 의미를 더욱 정확하게 파악할 수 있다. 실제로 BERT는 명명된 개체 인식(named entity recognition), 감정 분석 및 특정 유형의 질문 응답과 같은, 주로 맥락 의존적인 작업에서 특히 강력한 성능을 보인다. 즉, BERT는 근본적으로 텍스트를 깊이 이해하기 위한 인코더로 설계되었던 것이나, 이는 텍스트 생성을 위한 설계는 아니었다. 따라서 BERT는 사용자의 질의를 높은 정확도로 이해할 수 있었지만, 유창한 응답을 생성하는 능력에는 제한적이었다. 반면, GPT는 주로 텍스트 생성에 초점을 맞추며, 주어진 시퀀스에 이어질 다음 시퀀스를 예측하는 단방향 생성 접근법을 따른다. BERT의 양방향 접근법은 글의 맥락에 대한 더 깊은 이해를 제공할 수 있었지만, 텍스트 생성에는 GPT의 디코더 기반 단방향 설계가 더 적합하였다. 이러한 GPT의 특성은 챗봇 시스템과 같은 대화 생성 프로그램에 있어 큰 강점을 갖는다.

다양한 대형 언어 모델 기반의 챗봇들

인간의 대화를 모방하도록 설계된 프로그램인 챗봇은 GPT와 같은 대형 언어 모델을 사용하게 되면서 큰 발전을 이루었다. 최근의 챗봇들은 일관된 대화 흐름을 유지하고 맥락에 적합한 응답을 생성함으로써 사용자 상호작용을 크게 향상시킨다. OpenAI의 챗봇인 ChatGPT는 현재 약 1조 개

의 매개변수를 가진 GPT-4를 기반으로 하여 개발되어 있으며, 콘텐츠 생성과 같은 창의적인 작업에서 뛰어난 성능을 발휘한다. 인간이 작성한 텍스트를 이해하고 또 인간이 작성한 것과 같은 텍스트를 생성하는 작업에 있어 매우 뛰어난 능력을 보여줌으로써 인간이 기계 지능과 상호작용하는 방식을 혁신적으로 개선하였고, 궁극적으로 인공지능 연구 분야에 있어 기념비적인 도약을 이루어 내었다. 이러한 성능 도약의 핵심 중 하나는 인간 트레이너로부터의 피드백을 바탕으로 모델이 반복적으로 응답을 개선해 나가는 인간 피드백 기반 강화 학습(reinforcement learning from human feedback) 과정이다. 이는 미묘한 인간의 의사소통과 더 비슷하게 일치되도록(align with) 챗봇의 언어 이해력과 출력을 세밀하게 조정해 준다.

현재 다양한 기능과 용도를 가진 챗봇들이 나와 있고 각각 꾸준히 성능이 업데이트되고 있다. GPT의 최신 모델인 GPT-4 turbo는 2023년 12월까지의 데이터로 훈련되어 있고(GPT-4-turbo-2024-04-09 버전 기준), 그 이후의 최신 정보를 필요로 하는 질의에 대해서는 마이크로소프트의 Bing 검색 엔진을 통한 검색 증강 생성(retrieval augmented generation) 기법을 활용하고 있다. 마이크로소프트의 Copilot (구 Bing Chat)의 경우, 마찬가지로 OpenAI의 GPT-4를 기반으로 하지만 온라인 검색 서비스 및 자사 Office 도구와의 연동 목적에 더 최적화되어 있다는 점에서 차이가 있다. 구글의 Gemini (구 Bard)도 구글의 방대한 내부 데이터를 활용하여 지속적으로 성능을 개선하고 있다. 자체 개발한 언어 모델인 language models for dialog applications (이하 LaMDA)와 PaLM-2를 거쳐 현재는 동명의 언어 모델인 Gemini를 기반으로 하고 있다. 구글의 최신 모델인 Gemini 1.5 Pro와 Anthropic의 최신 모델인 Claude3 Opus의 경우 일부 벤치마크(benchmark; 모델의 성능을 평가하고 비교하기 위해 사용되는 표준화된 테스트 지표) 항목에서 GPT-4 turbo 이상의 성능을 보여주기도 하였다. 이러한 벤치마크 성능 순위는 각 업체들의 신규 모델 출시 및 버전 업데이트에 따라 계속 뒤바뀌고 있는 추세이다.

한편 현재 사용 중인 벤치마크 항목들이 과연 제대로 모델의 성능을 평가하고 있는가에 대한 의문이 새롭게 제기되고 있다. 벤치마크 질문의 일부만 수정하는 것만으로도 모델의 성능이 급격히 떨어지는 현상이 관찰되었고, 이는 모델들이 벤치마크의 질문을 이해하고 답하는 것이 아니라 질문과 답을 외워버리는, 소위 데이터 오염(data contamination) 현상 때문인 것으로 추정되고 있다(13, 14). 이러한 데이터 오염 현상을 해결할 수 있는 새로운 방식의 평가법들이 다양하게 제안되고 있기도 하지만 제일 이상적인 방법은 각 연구자가 직접 설계한 작업에서의 실제 성능을 측정하는 것이다. 영상의학 분야와 관련한 실제 작업 성능에 대한 비교, 평가에 대해서는 각론의 [대형 언어 모델들 간의 성능 비교] 챕터에서 보다 상세히 다루었다.

멀티모달(Multimodal) 대형 언어 모델

멀티모달 대형 언어 모델은 텍스트, 이미지, 음성 및 실시간 센서 데이터 등 다양한 형태의 정보를 통합하여 처리할 수 있는 인공지능 모델이다. 이러한 모델은 각각의 데이터 유형에서 중요한 정보를 추출하고 이를 결합하여 보다 정확하고 다양한 형태의 응답을 생성할 수 있다.

구글은 자사의 PaLM 모델을 기반으로 시각적 이미지와 연속적인 센서 데이터를 통합하여 새로운 시각-언어 인공지능 모델인 PaLM-E를 개발하였다(15). 이 모델은 시각 이미지와 텍스트 맥락

을 동시에 이해할 수 있었고, 추가 미세조정 없이도 복잡한 작업을 성공적으로 수행하였다. OpenAI 또한 음성 및 이미지(GPT-4Vision) 인식 기능을 ChatGPT에 통합하여 다양한 모달리티에 대한 인공지능의 접근 방식을 강화하였다(16). 2024년 5월 13일에는 텍스트, 오디오, 이미지를 하나의 모델로 통합하여 한 번에 입력받고 또 출력할 수 있는 GPT-4o (mni)를 출시하였다(17). 기존의 GPT-4의 경우 음성을 입력받으면 이를 텍스트로 변환하는 모델 하나, 변환된 텍스트에 대한 응답 텍스트를 생성하는 모델 하나, 생성된 텍스트를 다시 음성으로 전환하는 모델 하나, 총 세 개의 모델을 연결하는 파이프라인으로 구성되어 있었다. 반면 금번에 출시된 GPT-4o의 경우 하나의 통합 모델이 오디오 자체를 그대로 입력 및 출력으로 이용하기 때문에, 기존의 음성-텍스트 전환 과정에서는 버려졌던 정보들, 예를 들어 말하는 사람의 톤이나 뒤에 깔리는 배경 소음과 같은 정보들도 활용할 수 있게 되었고, 또 웃는다거나 노래를 부른다거나 하는 감정 표현과 같은 정보들까지 모두 반영된 응답을 생성할 수 있게 되었다.

이러한 멀티모달 접근 방식은 특히 의료 분야와 같은 복잡하고 다차원적인 데이터 해석이 필수적인 분야에서 큰 잠재력을 갖는다. 이에 따라 의료 분야에 특화된 멀티모달 대형 언어 모델들[BiomedGPT (18), Med-PaLM M (19)]이 개발되었다. 특히 영상의학 분야의 경우 이미지와 언어를 함께 이해할 수 있는 비전-언어 모델(vision-language model)이 활용도가 높을 것으로 기대되고 있고, 이에 대해서는 각론 [비전-언어 모델] 챕터에서 보다 상세히 다루었다.

프롬프트 엔지니어링(Prompt Engineering)

프롬프트 엔지니어링이란 대형 언어 모델을 통해 사용자가 원하는 정보를 최대한 정확하게 추출해 내기 위해 입력 쿼리, 즉 프롬프트를 상세하게 설계하고 최적화하는 작업을 의미한다. 이는 대형 언어 모델을 효과적으로 사용하기 위한 필수적인 과정으로 간주된다. 몇 가지 프롬프트 엔지니어링 기법들을 정교하게 잘 조합하여 사용하는 것만으로도 out-of-the-box 상태의 GPT-4가 의료에 맞춤 미세조정된 Med-PaLM2보다 의료 관련 벤치마크에서 더 높은 성능을 보였다는 연구결과도 있다(20). 널리 알려져 있는 몇 가지 대표적인 기법들을 소개하였다.

- 명확하고 구체적인 지시를 제공해야 한다. 너무 일반적이거나 불분명한 지시를 하게 되면 모델이 다양한 해석을 할 수 있게 되고 이는 너무 광범위하거나 질문자의 의도와 관련성이 떨어지는 응답으로 이어질 수 있다. 예를 들어, “현재 기술 발전의 최전선에 대해 알고 싶다”보다는 “인공지능과 기계 학습의 최신 동향에 대해 설명해 달라”가 더 구체적이고 명확한 지시의 예이다.
- 대형 언어 모델에 특정한 역할을 부여하는 것은 원하는 결과를 유도하는데 효과적이다. 예를 들어, 모델에게 “너는 영상의학 분야 전문가이다”와 같은 문장을 추가하면 관련 분야에 대한 자세하고 정확한 정보를 제공받는 데 도움이 될 수 있다. 이는 모델이 질문에 대한 답변을 생성할 때 질문자의 맥락이나 요구사항에 부합하도록 돕는다.
- 장문의 맥락 텍스트를 제공할 때에는 삼중 따옴표(""")로 묶는 것이 효과적이다. 이는 모델로 하여금 사용자의 지시와 맥락 텍스트를 구분하는 데에 도움이 된다. 예를 들어, 다음 문장을 한글로 번역해 줘. """*I love dogs*."""와 같이 활용할 수 있다.
- 모델이 능숙하지 않은 작업에 대해서는 여러 예시를 제공하는 few-shot 프롬프팅이 도움이 될

수 있다. 이는 모델에 추가적인 맥락과 지침을 제공하여 작업의 성공률을 향상시키는 데 기여한다. 그러나 모든 경우에서 예시를 제공하는 것이 효과적인 것은 아니며, 때로는 예시를 제공하지 않을 때 더 잘 해내는 경우도 존재한다.

- 중간 추론 단계를 거치는 연쇄 사고(chain-of-thought)를 유도함으로써 보다 정확하고 논리적인 답변을 얻을 수 있다. "단계별로 생각해 보자"와 같은 간단한 프롬프팅을 통해 모델이 추론 과정에서 명확한 구조를 세우게끔 할 수 있다.

대형 언어 모델 사용에 있어 주의해야 할 점

환각과 편향

대형 언어 모델 사용에 있어 가장 널리 알려져 있고 또 치명적인 문제점은 소위 환각(hallucination)이라고 불리는 현상으로, 이는 모델이 겉보기에 그럴듯해 보이지만 실제로는 사실이 아닌 정보를 제공하는 것을 말한다. 예를 들어, Lung-RADS 5와 6 카테고리에 관한 질문을 받았을 때, ChatGPT와 Gemini 둘 다 실제로는 Lung-RADS 5나 6 카테고리 같은 것은 존재하지 않는다고 대답하는 대신 잘못된, 그러나 잘 모르는 사람이 봤을 때에는 매우 사실처럼 느껴지는 대답을 했다는 보고가 있다(21). 이러한 환각 현상이 발생하는 주요한 원인은 사전 훈련 데이터의 양이 부족하거나 거기에 내재되어 있는 오류 때문이다. 따라서 대형 언어 모델들의 사전 훈련 데이터셋의 크기가 점점 증가하면서 환각 현상 또한 많이 개선되었지만 여전히 완전히 해결된 상태는 아님에 주의해야 한다.

또 다른 문제는 편향(bias)이다. 사전 훈련 데이터셋에 편향이 내재되어 있는 경우, 모델 또한 그 편향을 그대로 반영할 수밖에 없다. 한 연구는 ChatGPT, Gemini, Claude와 같은 대형 언어 모델 기반의 챗봇들이 이미 과학적으로 논파가 끝난, 인종에 관해 잘못 퍼져 있는 많은 편견들(예를 들어, 흑인과 백인은 크레아티닌 수치, 폐 기능, 피부 두께 등에 있어 차이가 있다는 주장들)을 그대로 학습하여 대답에 사용하였음을 보였다(22). 또 다른 한 연구는 GPT-4가 여러 의학적 질환들의 실제 유병률 및 지리적 분포를 반영한 응답이 아닌, 그중 일부에 해당할 뿐인 전형적인 인종과 성별의 사례들을 과장하여 반영하는 응답을 했음을 보였다(23). 연구자들은 점점 더 많은 사람들이 대형 언어 모델을 사용하게 됨에 따라 편향된 정보들이 사용자들 사이에서 오히려 더 빠르게 퍼지게 될 상황에 대해서도 우려하였다(23). 이러한 모델의 편향은 검증되지 않은 인터넷 자료나 오래된 서적의 내용 등 낮은 품질의 데이터들이 무분별하게 학습되면서 생긴 것으로 생각되는데, 현재 대형 언어 모델들의 학습 데이터에 대한 정보는 투명하게 공개되어 있지 않기 때문에 직접적으로 어떤 데이터가 어떤 편향을 유발하였는지는 알기가 어렵다. 인간 피드백 기반 강화학습이 이러한 환각과 편향 해소에 어느 정도 도움이 되기는 하였으나 완벽히 해결했다고는 아직 볼 수 없다.

매번 바뀌는 대답

챗봇을 사용하다 보면 같은 내용의 질문을 반복적으로 하였을 때 대답이 매번 조금씩 달라지는 현상을 종종 겪게 된다. 실제로 대형 언어 모델을 이용한 연구 논문들을 보면 동일한 프롬프트를

세 번 정도 반복하여 넣은 결과를 통합적으로 분석하곤 하는데, 이는 응답의 변동성을 확인하고 필요시 보정하기 위한 목적에서였다(24-27). 이러한 현상은 대형 언어 모델의 작동 원리가 본질적으로 결정적(deterministic)이지 않고 확률적(stochastic)이기 때문에 발생한다. 질문에 대한 대답을 생성할 때, 어떤 정해진, 고정된 규칙이 적용되는 것이 아니라 각각의 가능한 토큰들의 확률 분포에 기반한다는 의미이다. 대형 언어 모델의 다양하고 풍부한 반응은 이러한 확률적인 특성 덕분이지만, 한편으로는 그 반응이 실제로 인간의 언어를 이해하여 나온 것이 아니라 오직 수학적인 확률 계산에 의한 기계적인 결과값일 뿐이라는 비판을 불러일으키기도 한다. 이들이 마치 사람과 대화하고 있는 것 같은 느낌을 주지만 이는 결국 학습 데이터셋에 있는 정보들을 앵무새처럼 반복 재생하고 있을 뿐이라는 것이다(28). 이러한 지적 자체는 지극히 사실이지만, 더 중요한 것은 이러한 학습 정보의 반복 재생이 단순한 흉내 내기 수준이 아니라 고도화된 패턴 인식과 적용의 결과라는 점이다. 비록 대형 언어 모델들이 텍스트의 의미를 사람처럼 실제로 "이해"하지 않는 것은 사실이지만, 확률을 활용하여 고도화된 언어적 패턴을 처리, 인식, 적용하는 접근은 인공지능 개발의 영역에 있어 매우 중대한 도약이었다.

모델의 온도(temperature) 파라미터 값을 조정함으로써 이런 확률성을 어느 정도의 수준으로 적용할 것인지를 사용자가 설정할 수 있다(29). 수행하려는 작업의 성격이 가급적 확률성을 배제한 보수적인 답변을 필요로 한다면 온도 파라미터에 0에 가까운 값을 넣으면 되고, 반대로 창의적이고 다양한 반응이 더 도움이 된다면 1에 가까운 숫자를 넣으면 된다. ChatGPT의 경우 맞춤형 지시사항(custom instructions)을 설정하는 기능도 제공한다. 이를 활용하여 본인이 어떤 입장에서, 어떤 목적으로 챗봇을 사용 중인지, 따라서 어떤 스타일의 답변을 원하는지 등을 미리 설정하여 챗봇에게 알려줄 수 있다(30). 이를 통해서도 모델의 확률성을 어느 정도 조절할 수 있으므로, 본인이 수행하려는 작업에 어느 정도의 확률성이 적절할지에 대해 고찰해 보는 것이 필요하다.

한편 모델의 성능 자체가 버전 업데이트에 따라 계속 변화(performance drift)하고 있다는 것도 문제가 될 수 있다. 2023년 3월 버전의 ChatGPT와 6월 버전의 ChatGPT에게 각각 같은 작업을 시켜보았을 때, 어떤 작업들의 성능은 향상된 반면 오히려 성능이 하락한 작업들도 있었음을 보인 연구 결과가 있다(31, 32). 문제는 각 업체들이 어떤 식으로 모델을 수정하고 업데이트하고 있는지가 투명하게 공개되어 있지 않다는 것이다. 따라서 사용 중인 대형 언어 모델의 시간에 따른 성능 변화가 수행 중인 작업에 어떻게 영향을 미치는지를 꾸준히 정기적으로 모니터링할 수 있는 일종의 정도관리용 벤치마크를 확보해 두거나, 아니면 온프레미스(on-premise; 온라인 클라우드 서버가 아닌 사용자의 자체 인트라넷 서버 내부에 구축하는 방식) 모델을 구축하여 업데이트를 직접 관리하는 방식이 권장된다(33).

정보 보안

GPT는 오픈 소스로 공개되어 있지 않은, 상업적으로 이용 가능한 대형 언어 모델이다. GPT에 입력하는 모든 정보는 OpenAI 회사 서버로 전송이 되고 있기 때문에, 실제 환자의 의학 정보를 입력하는 것은 개인정보보호법 위반이 될 수 있다. GPT를 의료나 영상의학 분야에 적용했던 대부분의 연구들이 실제 환자 데이터를 그대로 이용하지 못하고 비식별화 처리를 해야만 했던 것이나,

아니면 공개 데이터셋을 활용하거나, 혹은 아예 가상의 환자 정보, 판독문 등을 만들어서 실험했던 이유가 여기에 있다. 따라서 대형 언어 모델을 병원 시스템과 연계시켜 실제 진료 현장에서 활용하려 한다면, 국외 서버로의 개인 의료정보 전송 및 관련한 국내법 규제 현황 등에 대해 먼저 검토해 보는 것이 필요하다. 다른 대안으로는 온프레미스 모델을 새로 개발하여 병원 내부 인터넷망에 설치하는 방법을 고려할 수 있다. 관련하여 보다 자세한 내용을 아래 "작아진 대형 언어 모델" 단락에 기술하였다.

대형 언어 모델의 발전 방향

검색 증강 생성(Retrieval Augmented Generation)

환각 및 편향 현상을 해결하기 위해 가장 널리 활용되고 있는 방법은 검색 모델을 생성 모델과 연동시킨 검색 증강 생성 기법이다(34). 이 방법은 학습 데이터셋에 포함되어 있지 않던 자료를 대상으로 추가적인 검색을 수행하여 대답을 생성할 수 있기 때문에, 학습 데이터의 부족, 오류로 인해 발생할 수 있었던 환각과 편향 현상을 줄여줄 수 있다. 특히 온라인 데이터도 검색할 수 있기 때문에, 실시간으로 업데이트된 정보가 필요한 작업도 수행 가능하게 해준다. 게다가 생성한 대답의 근거가 되는 자료를 투명하게 제시해 줄 수 있어지므로, 그에 대한 검증을 통해 환각이나 편향 여부를 더욱 쉽고 명확하게 파악할 수 있다. 현재 가장 널리 사용되고 있는 세 개의 챗봇인 ChatGPT와 Copilot, Gemini 모두 온라인 데이터 검색 기능을 갖추고 있다.

작아진 대형 언어 모델(Small Large Language Model)

방대한 양의 사전 훈련 덕분에 추가 미세조정 없이도 zero-shot, few-shot 학습이 가능해진 것은 물론 대형 언어 모델의 큰 장점임에 틀림없지만, 여전히 환각 현상이 남아있다는 사실은 작업의 종류에 따라 더 크게 문제가 될 수 있다. 특히 의료는 환자의 건강을 직접 다루는 분야인 만큼 환각/편향 현상은 자칫 심각한 피해를 초래할 수 있다. 이러한 배경에서 의학 지식들에 대한 미세조정의 필요성이 다시 대두되었다. 마찬가지의 이유로 모델의 성능이 버전 업데이트에 따라 계속 영향을 받는 것도 의료에서는 더욱 큰 문제가 될 수 있다. 이에 더해 개인 의료정보의 경우 개인정보보호법 위반 문제까지 있어, 결국 실제 진료 현장에서 대형 언어 모델을 활용하기 위해서는 온프레미스 모델을 개발하여 의료 기관 내부 전산망에 설치하는 것이 필요하다. 이러한 미세조정 및 온프레미스 모델 개발에 있어 가장 주목받고 있는 컨셉은 작아진 대형 언어 모델이다.

대형 언어 모델들은 한동안 파라미터 개수와 사전 훈련 데이터셋 크기를 증가시켜 나가는 방향으로 발전해 왔으나, 최근에는 작아진 대형 언어 모델이라는, 일견 모순적으로 보일 수 있는 개념의 방향으로 발전해 나가고 있다. 대형 언어 모델의 가장 큰 문제점 중 하나는 개발 및 구동, 유지/보수 비용이 너무 비싸다는 것이다. 따라서 규모가 작은 기업체나 연구소에서는 파라미터의 개수와 학습 데이터셋의 크기는 줄여 훈련 비용은 낮추면서도 성능은 유지될 수 있는, 효율적인 언어 모델을 필요로 하게 되었다. 보다 작은 스케일의 효율적인 모델을 활용하는 것은 지구 환경적 관점에서도 바람직한 방향이겠다. 가장 대표적인 예시는 Meta의 오픈 소스 언어 모델인 large lan-

guage model meta AI (이하 LLaMA)이다. LLaMA-2의 경우, 700억개의 파라미터만으로도 1750억개 파라미터의 GPT-3와 비슷한 성능을 보여주었으며, 2024년 4월 공개된 LLaMA-3는 동일한 700억개의 파라미터로 구글의 가장 최신 모델인 Gemini 1.5 Pro를 능가하는 성능까지 보여주었다. 이보다 더 경량화된 모델들도 계속 개발되고 있다. LLaMA의 경량화 모델인 Alpaca, Koala, Vicuna가 다양한 파라미터의 개수로 세분화되어 공개되어 있고, PaLM-2 또한 Gecko, Otter, Bison, Unicorn의 네 종류의 다양한 파라미터 개수의 경량화 모델들이 개발되어 있다. Mistral AI에서 오픈 소스로 공개한 Mixtral 8x7B는 467억개의 파라미터 개수만으로도 LLaMA-2와 GPT-3.5에 비견할만한 성능을 보였는데, 이는 전문가 혼합(mixture of experts)이라 불리는 방식을 활용한 덕분이다. 전문가 혼합은 모델을 여러 개의 하위 전문가 모델들로 나누고, 그중 수행하려는 작업에 가장 적합한 하위 전문가 모델 하나를 선택하여 계산하게 하는 식으로 모델의 효율성을 제고하는 접근법이다. 현존하는 가장 큰 스케일의 대형 언어 모델인 GPT-4와 Gemini 1.5 Pro 조차 이 전문가 혼합 방식을 차용하고 있는 것으로 알려져 있기도 하다.

의료 영역에서 작아진 대형 언어 모델이 주목받고 있는 가장 큰 이유는, 미세조정을 통해 의료 관련 업무에 특화되도록 모델을 개발하여 환각/편향을 최소화하는 것이 필요한 동시에, 의료가 아닌 다른 분야에 대한 작업 성능은 상대적으로 덜 중요하기 때문에, 경량화 모델이 활용되기에 최적의 상황이기 때문이다. 약 70억개의 파라미터만을 갖는 LLaMA-7B 모델을, 한 실제 온라인 의료 상담 웹사이트로부터 얻은 100000개의 환자-의사 간의 대화들로 미세조정하여 개발한 ChatDoctor의 경우, 의료 상담 작업에 있어 ChatGPT-3.5 이상의 성능을 보여주었다(35). 약 130억개의 파라미터를 갖는 LLaMA의 경량화 모델 Vicuna-13B를 온프레미스로 설치하여, 실제 환자의 흉부 X-ray 영상 판독문을 대상으로 한 작업을 비식별화 과정 없이, 그러면서도 개인 의료정보 유출 없이 성공적으로 수행할 수 있었음을 보인 연구도 있다(36). 이처럼 Vicuna-13B를 온프레미스로 설치하면 사용자가 버전 업데이트를 직접 관리하여 성능 수준을 유지할 수 있고 따라서 높은 재현성(reproducibility)을 담보할 수 있다는 장점도 있다(37).

각론: 영상의학 분야에서의 대형 언어 모델 활용

대형 언어 모델은 전문적인 영상의학 지식을 갖추고 있는가?

ChatGPT는 일상적인 사람 대화를 모사하는 것을 일차적인 목표로 하여 훈련된 범용 목적의 챗봇이지만, 많은 학술적, 전문적인 영역에서도 놀라운 성능을 보여주었다(38). 의학 영역 또한 깊은 전문성이 요구되는 분야이고 특히 영상의학의 경우 다른 곳에서는 잘 사용되지 않는 고유한 용어들까지 있기 때문에, ChatGPT와 같은 범용 대형 언어 모델들이 추가적인 미세조정 없이 바로 활용될 수 있을지에 대해서는 검증이 필요하다.

관련하여 여러 연구들이 진행된 바 있다. 우선 ChatGPT-3.5를 활용하여 폐암(정답률 70.8%), 간암(정답률 75%), 심혈관계 질환(정답률 84%), 유방암(정답률 88%)과 같은 병과 관련하여 일반인들이 궁금해하는 수준의 질문에 대해, 완벽하진 않지만 그래도 대체로 적절한 답변들을 생성해낼 수 있음을 확인한 연구들이 있다(21, 25, 26, 39, 40). Korean Thyroid Imaging Reporting and

Data System (K-TIRDAS)에 관한 퀴즈를 풀게 해보았을 때에는, ChatGPT-3.5의 경우엔 정답률이 73%였으나 ChatGPT-4는 93%의 높은 정답률을 보였다(41). ChatGPT-3.5/4로 하여금 미국 의사 면허 시험(United States Medical Licensing Examination)을 치르게 해 보니 합격 점수, 혹은 거의 그에 상응하는 수준의 점수를 받을 수 있음을 확인한 연구들도 있다(42-44). GPT-4로 하여금 의학 계열 학술지 *New England Journal of Medicine*에서 제공하고 있는 복잡하고 난이도가 높은 의학 퀴즈인 "*Case Challenges*"를 풀게 하되, 실제 퀴즈에 포함되어 있던 영상 사진 대신 그 소견을 정리해 놓은 글과 문제 지문을 함께 제공해 주자 실제 해당 학술지 구독자들의 99.98%보다 높은 성적을 보였음을 보인 연구가 있다(45). 비슷하게 ChatGPT에게 영상의학 계열 학술지인 *Radiology*에서 제공하고 있는 "*Diagnosis Please*"라는 퀴즈를 풀게 하되, 역시 영상 사진 대신 그 소견을 정리해 놓은 글과 해당 환자의 임상 양상을 제시해 주자 ChatGPT-4의 경우 절반 이상의 문제에서 정답을 선택했음을 보여준 연구들도 있다(32, 46). ChatGPT에게 미국과 캐나다, 브라질 영상의학 전문의 자격시험을 치르게 해 보았을 때에도 합격 점수를 넘기거나, 혹은 거의 그에 상응하는 수준의 점수를 받기도 했는데(47-49), 이는 ChatGPT가 갖고 있는 영상의학적 전문 지식이 영어가 아닌 다른 언어를 통해서도 잘 활용될 수 있음을 시사한다.

그러나 텍스트가 아닌 실제 의료 영상 이미지를 해석, 판독하는 능력에 대해서는 아직 부족함이 많다. GPT-4Vision으로 하여금 한 국내 의과대학의 실제 영상의학 과목 시험 문제를 풀게 해보았을 때, 텍스트 기반의 문제들의 경우 85%의 문제에서 정답을 맞혔던 반면, 이미지 기반의 문제들에서는 52%의 낮은 정답률을 보였다(50). 이는 실제 의과대학 3학년 학생들의 평균에도 미치지 못하는 낮은 성적이었다. GPT-4Vision에게 흉부 X-ray 영상을 보여주고 관찰되는 영상 소견을 텍스트로 요약하게끔 했던 연구에서도 F1 점수가 zero-shot의 경우 7.3%–18.2%, few-shot의 경우 11.1%–34.3% 밖에 되지 않았다(51). GPT-4Vision의 의학 영상을 이해하는 능력에는 아직 많은 개선이 필요해 보인다.

한편 이러한 종류의 연구 결과들의 해석에는 주의가 필요하다. 퀴즈나 시험의 경우, 논쟁거리가 전혀 없이 명확히 사실 관계가 확인된 진단 혹은 치료만을 대상으로 하고, 하나의 명확한 정답만이 존재하며, 문제 내에 그 정답을 가리키는 여러 개의 단서들이 일관되게 존재한다는 점에서 실제 임상 상황과는 매우 다르다. 실제 영상의학과 의사가 하는 일은 불명확하고 때로는 부정확하게 기록된 환자 정보를 참고해가며, 하나의 단정적인 진단이 아닌 여러 개의 가능성 있는 감별진단들과 추가적으로 시행해야 할 검사 혹은 치료법 등을 제시하는 일에 더 가깝다(52). 따라서 ChatGPT가 전문적인 영상의학 지식을 어느 정도 갖추고 있고, 제시된 질문을 잘 이해하여 적절한 대답을 해줄 수 있음은 충분히 검증되었으나, 아직까지는 단독으로 실제 임상 현장에 투입되기보다는 영상의학과 의사에 의해 보조적으로 활용되는 정도가 권장된다.

영상의학 영역에서의 대형 언어 모델 활용

영상의학 전문가가 대형 언어 모델을 활용하여 생산성을 향상시킬 수 있는 방법들은 다양하게 제안되고 있다(Table 1). 첫째, 언어적 장벽을 해소하는 데 유용하게 사용될 수 있다. 영상의학은 복잡한 학문 분야로, 경험이 풍부한 전문의조차도 때때로 해석에 어려움을 겪는 영상들을 접하게

Table 1. Potential Clinical and Research Applications of Large Language Models in Radiology

Applications	Inputs	Outputs	Referernce
Reduction of language barriers	Any text written in one's primary language	Text translated into any language	(53)
Generation of radiology reports	Image findings sections within chest radiograph reports	New, short, one-line impression	(57)
	Text-based descriptions of image patterns	List of relevant differential diagnoses	(58)
	TI-RADS features	Distinction between benign and malignant thyroid nodules	(24)
	Radiology reports for patients referred for breast cancer screening or diagnosis	BI-RADS category	(59)
Transformation into structured reporting	Free-text radiology reports for chest radiograph, CT, and MRI	Transformation into structured reporting	(60)
	Free-text CT reports from patients with lung cancer	Extraction of oncologic information	(61)
	Free-text reports on mechanical thrombectomy from patients with acute ischemic stroke	Extraction of procedural details of mechanical thrombectomy	(62)
	Free-text chest radiograph reports from public databases (MIMIC-CXR and NIH datasets)	Presence or absence of 13 predetermined imaging findings	(36)
	Free-text brain MRI reports	Extraction of multiple predefined data elements, including any abnormal findings, and their correlation with headaches	(37)
Error detection in radiology reports	Radiology reports containing deliberately introduced errors	Errors were detected, demonstrating potential for automatic correction	(63)
	CT and MRI reports containing speech recognition errors	Errors were detected and corrected	(64)
Simplification of radiology reports for patients	Free-text radiology reports for chest CT and brain MRI	Radiology reports translated into plain language	(65)
	Free-text radiology reports from public database (MIMIC-III)	Radiology reports simplified using plain language	(66)
	Radiology report impressions from public database (MIMIC-IV)	Radiology report impressions simplified using plain language	(67)
Determination of radiologic study protocol	Medical conditions summarized in ACR appropriateness criteria	Determination of imaging modality and use of contrast agent	(68)
	Radiology request forms	Determination of imaging modality, body region, and contrast phases	(69)
	Clinical presentations regarding breast cancer screening and breast pain	Determination of imaging modality	(27)

ACR = American College of Radiology, BI-RADS = Breast Imaging Reporting and Data System, CXR = chest X-ray, MIMIC = Medical Information Mart for Intensive Care, NIH = National Institutes of Health, TI-RADS = Thyroid Imaging Reporting and Data System

된다. 이러한 영상을 판독할 때에는 문헌 검색이 필수적인데, 문제는 대부분의 유용한 문헌들은 영어로 작성되어 있어 비영어권 의사들에게는 상당한 시간과 노력이 요구된다는 것이다. ChatGPT-4는 다양한 언어 간 번역에서 기존 유료 번역 소프트웨어와 비견될 만한 뛰어난 성능을 제공함으로써 이러한 불필요한 과정을 최소화하는 데 기여한다(53). 영어가 익숙지 않은 연구자의 경우에는 논문 작성에도 도움을 받을 수 있다(54). 생성형 인공지능의 도움을 받아 논문을 작성한 경우

에 대한 정책 및 규정은 학술지별로 조금씩 차이가 있긴 하지만, 전반적으로는 관련 내용을 투명하게 공개하였다는 전제하에 그 자체를 금지하지는 않는 추세이다(55, 56).

다음으로 쉽게 적용해 볼 수 있는 분야는 판독문 작성이다. 일반적으로 영상의학 판독문은 영상 소견을 글로 정리한 부분과 그에 기반한 임시 진단명(impression), 혹은 다수의 가능한 감별진단들의 목록으로 구성된다. ChatGPT-4에게 영상 소견만을 제시해 주었을 때 그에 기반하여 적절한 임시 진단명(57)이나 감별진단의 목록(58)을 생성해 줄 수 있었음을 확인한 연구들이 있다. 종종 복잡하고 어려운 영상을 판독할 때에는 영상 소견으로부터 감별진단이 바로 이어져 나오지 않고 많은 고찰을 요하게 되는데, 이런 경우 ChatGPT를 활용하면 어느 정도 도움을 받을 수 있을 것이다. 대형 언어 모델로 하여금 영상 소견에만 기반하여 최종적인 악성도의 정도를 판별케 함으로써 판독문 작성에 도움을 받고자 했던 시도들도 있었다. 한 연구는 갑상선 결절의 초음파 사진을 영상의학과 의사가 보고 Thyroid Imaging Reporting and Data System (이하 TI-RADS) 소견으로 정리하여 대형 언어 모델에 넣어 주었을 때, ChatGPT-4 및 Gemini가 70%–80% 이상의 높은 정확도로 양성/악성 감별을 해낼 수 있었음을 보이기도 했다(24). 흥미로운 점은, 이 TI-RADS 소견을 주니어 의사가 정리하여 모델에 넣어 주었을 때에는 진단 정확도가 약 74%–82%였고, 시니어 의사가 했을 경우엔 약 83%–86% 정도였는데, 이 과정을 20000개 이상의 갑상선 결절 초음파 사진을 학습시켜 개발하였던 다른 AI 모델로 대체하여 진행하였을 때에는 최종적으로 82%–84% 정도의 정확도를 보였다는 점이다. 즉, 초음파 사진으로부터 TI-RADS 소견을 추출하고, 그것을 바탕으로 악성 여부를 판단하기까지의 전 과정이 인공지능 모델들만으로 이루어졌음에도 불구하고 주니어 영상의학과 의사 이상의 정확도를 보였다는 것은, 급속히 발전하고 있는 현대 인공지능 모델들의 뛰어난 성능을 잘 보여준다. 그러나 배치되는 연구 결과도 있다. 유방암의 감시 혹은 진단 목적으로 내원한 환자들의 실제 영상 판독문만을 보고 영상의학과 전문의와 ChatGPT-3.5/4 및 Gemini로 하여금 각자 Breast Imaging Reporting an Data System (BI-RADS) 카테고리를 결정하게 한 연구가 있다(59). 이 연구에서 대형 언어 모델들의 성능은 영상의학과 전문의에 많이 못 미치는 것으로 보고되었는데, 특히 환자에게 실제로 악영향을 미칠 수 있는 오류가 영상의학 전문의의 경우엔 1.5%였던 반면 ChatGPT-4에서는 10.6%, Gemini에서는 18.1%였다. 주의해야 할 점은 실제 영상의학과 의사가 영상 소견으로부터 임시 진단명을 결정하거나 악성 여부를 판단하는 과정이 오직 영상 소견에만 전적으로 의존하여 이루어지는 것은 아니라는 것이다. 어떠한 임상적 맥락에서 검사가 시행되었는지, 이때 의심되는 질환에는 어떤 것들이 있고 그것들은 각각 어떠한 영상 소견을 보일 수 있는지까지 알고 있어야만 제대로 된 판독이 가능하다. 때론 동일한 영상 소견이라 할지라도 환자의 임상적 맥락에 따라 전혀 다른 의미가 될 수도 있으므로, 대형 언어 모델에 영상 소견만을 제시할 때에는 이러한 한계점이 있음을 인지하고 있어야 한다.

대형 언어 모델은 구조화 판독문(structured reporting) 작성에도 활용될 수 있다. 구조화 판독문을 활용하면 보다 효율적으로 정보를 추출할 수 있고, 추출된 정보의 객관성도 높일 수 있으며, 또 여러 사람 간에 판독문 정보를 쉽고 정확하게 공유할 수 있다. 그러나 가장 큰 단점으로 지적되는 것은 구조화 판독문을 작성하는 과정에 많은 시간과 노력이 필요하다는 것이다. 특히 이미 내러티브 형식으로 작성되어 있던 비구조화 판독문들을 다시 구조화 판독문의 형식에 맡게 고치는

작업에는, 통일되지 않은 다양한 용어와 표현들을 일일이 해석하여 적절한 형식으로 바꿔야 하는 과정도 포함되기에 관련 지식을 깊게 이해하고 있는 전문가만이 할 수 있다. 영상의학 전문가 대신 ChatGPT-4를 이용하면 기 작성되어 있는 비구조화 판독문들을 구조화 판독문으로 적절하게 변환할 수 있음을 보인 연구가 있다(60). 또 다른 연구는 ChatGPT-4를 이용하여 폐암 환자의 판독문으로부터 여러 종양들의 크기 변화 및 전체적인 치료 반응 평가와 같은 종양학적 정보들만을 선택적으로 일괄 추출해 낼 수 있음을 보이기도 했다(61). 기계적 혈전제거술을 받은 급성 뇌졸중 환자의 판독문으로부터 시술 관련 세부 정보들을 항목별로 일괄 추출해내는 작업도 가능함을 보인 연구도 있다(62). ChatGPT뿐 아니라 LLaMA의 경량화 모델인 Vicuna-13B를 이용해서도 흉부 X-ray나 뇌 MRI 판독문으로부터 사전에 정의된 여러 정보들을 추출해낼 수 있었음을 보인 연구들도 있다(36, 37). 이처럼 대형 언어 모델을 활용하면 영상의학 전문가의 도움 없이 편하게 비구조화 판독문을 구조화 판독문으로 전환하거나 특정 정보들을 추출해낼 수 있고, 이는 궁극적으로 영상의학과 의사와 임상 의사 사이의 소통 및 여러 연구자들 간의 데이터 공유에도 도움이 될 것이다.

판독문의 오류, 오타를 찾아내는 데에도 대형 언어 모델이 활용될 수 있다. 200개의 실제 영상 판독문 중 100개에 대해 일부러 단순한 오류를 만들어 놓은 후, GPT-4와 6명의 영상의학과 의사들로 하여금 각자 그 오류들을 찾아내도록 해 본 연구가 있다(63). GPT-4의 오류 발견율은 82.7%였고 이는 6명의 영상의학과 의사들과 비교했을 때 중간 정도에 해당하는 수준의 성능이었다(1등이 94.7%, 6등이 80.0%). 일부러 만들어낸 단순한 오류가 아닌, 실제 판독문에 있던 복잡한 수준의 오류 또한 찾아낼 수 있는지 확인한 연구도 있다. 이 연구는 GPT-4로 하여금 3233개의 실제 CT/MRI 판독문으로부터 오류를 찾아내어 고치라는 작업을 수행하게 하였고, GPT-4는 F1 점수 86.9%–94.3%의 높은 성능을 보였다(64). 이때 GPT-4가 찾아낸 오류의 종류에는 단순 오타, 단어 누락, 혹은 비슷한 발음의 다른 단어가 적혀 있던 경우, 문법 오류와 같이 언어 자체에 대한 지식만으로 찾아낼 수 있는 오류뿐 아니라, 앞뒤가 맞지 않는 내용이 동시에 적혀 있던 경우, 혹은 의미적으로 말이 안 되는 내용이 적혀 있던 경우도 있었다. 이것들을 오류로 인식해 냈다는 것은 GPT-4가 판독문 전체에 걸친 맥락을 파악하는 능력과 적절한 판독문의 내용이 어떠해야 하는지에 대한 이해도 갖추고 있다는 것을 시사한다.

대형 언어 모델은 영상의학과 의사와 환자 간의 소통에도 활용될 수 있다. 영상의학 판독문에는 다른 분야에서는 잘 사용되지 않는 고유한 용어들이 많아 비의료인 환자들이 그 의미를 제대로 이해하는 데에 어려움이 많다. ChatGPT를 비롯한 여러 대형 언어 모델들을 활용하여 영상 판독문을 전문적인 의학 지식이 없는 일반 사람들도 쉽게 이해할 수 있는 언어로 바꿔 간단하게 요약할 수 있음을 보인 연구들이 있다(65-67). 이처럼 쉬운 언어로 간단하게 요약된 판독문을 활용하면, 환자들도 자신들의 건강 상태에 대해 더 잘 이해하게 되고 나아가 치료 방침 결정에도 보다 더 적극적으로 참여할 수 있게 된다.

영상의학 전문가의 업무는 단순히 영상을 판독하는 것에 그치지 않는다. 개별 환자들의 다양한 임상 상황에 맞춰 어떠한 검사를, 어떠한 프로토콜로, 어느 부위를 대상으로 촬영할 것인지, 조영제를 사용할 것인지, 사용한다면 어떤 조영제를 사용할 것인지 등을 결정하는 일 또한 영상의학 전문가만이 할 수 있는 작업들이다. 환자의 과거력과 현재 어떠한 임상적 질문을 해결하고자 하는

지에 대한 정보를 ChatGPT-3.5/4에 제공하였을 때, 적절한 촬영 프로토콜에 관한 세부 사항들을 제시해 주었음을 보인 연구들이 있다(27, 68, 69). 이와 같은 대형 언어 모델을 병원 처방 시스템과 연계시켜 구축하게 된다면, 촬영 프로토콜 관련하여 임상의들을 상담해 주는 영상의학과 의사의 노력을 줄일 수 있고 잘못된 영상 프로토콜 처방으로 환자에게 위해가 가해지는 것 또한 막을 수도 있다.

대형 언어 모델들 간의 성능 비교

OpenAI의 ChatGPT가 등장한 이후 마이크로소프트는 Copilot을, 구글은 Gemini를 개발하였다. ChatGPT의 기반 모델은 GPT-3.5, GPT-3.5 turbo, GPT-4, GPT-4 turbo 버전으로 순차적으로 업데이트되었고, Gemini 또한 LaMDA, PaLM-2, Gemini 1.0, Gemini 1.5로 기반 모델이 바뀌면서 성능이 점점 향상되었다. 메타의 LLaMA도 현재 LLaMA-3까지 공개되어 있는 상태다. 따라서 모델들 간의 성능을 비교한 연구 결과를 해석할 때에는 어떤 모델의 어떤 버전의 단계에서 수행되었는지에 유념해야 한다. 또한 총론에서 언급하였듯, 벤치마크만으로 모델의 성능을 평가하는 데에는 한계가 있으므로 실제 작업 성능을 확인하는 것이 권고되는데, 이때 대상이 된 작업이 어떤 종류의 것이었는지, 특히 프롬프트로 정확히 어떤 내용이 들어갔었는지도 유념해서 보아야 한다. 실제로 각 모델들 별로 프롬프트에 들어가는 내용에 따라 성능 차이가 나는 양상이 모두 다르기 때문이다(49, 66, 67). 그럼에도 불구하고 현재까지의 연구 결과들을 종합해 보면 GPT-4가 대부분의 작업에 있어 가장 높은 성능을 보이고 있다(Table 2). 이러한 GPT-4의 헤게모니가 경쟁 업체들의 지속적인 업데이트로 인해 뒤바뀔 수 있을지(이를테면, Gemini 1.5 Ultra, 혹은 Claude4의 출시), 아니면 GPT-4.5 업데이트, 혹은 GPT-5 출시로 인해 오히려 더 공고해질지는 매우 흥미로운 관전 포인트이다.

영상의학 분야로 미세조정된 대형 언어 모델들

총론의 [작아진 대형 언어 모델] 챕터에서 언급하였듯, 의료 분야의 지식들로 미세조정되어 하부 작업(downstream task)들의 성능을 전체적으로 향상시킬 수 있는 의료 분야 맞춤 기반 모델(medical domain-specific foundation model)의 필요성이 점점 대두되었다(70). 구글의 PaLM을 기반으로 하여 의료 지식에 미세조정된 Med-PaLM (71), Med-PaLM2 (72), 그리고 BLOOM-7B를 기반으로 한 ClinicalGPT (73)와 같은 모델 등이 현재 개발되어 있다. 영상의학 분야로 좀 더 세부 미세조정된 모델들도 있다. Alpaca-7B를 기반 모델로 하여 영상의학 지식에 미세조정된 Radiology-GPT (74), LLaMA-2를 기반 모델로 한 Radiology-LLaMA2 (75) 등이 개발되어 있다. 이러한 모델들은 여러 벤치마크에서 기존의 범용 언어 모델들보다 더 높은 성능을 보였고, 따라서 영상의학 관련 실제 작업에서도 상당한 잠재력을 갖고 있을 것으로 기대된다.

비전-언어 모델(Vision-Language Model)

멀티모달 대형 언어 모델 중 영상의학 분야에서 가장 활용도가 높을 것으로 기대되는 것은 이미지와 언어를 함께 이해하여 서로 정렬시킨(align) 비전-언어 모델이다. 가장 대표적인 비전-언어

JOURNAL of THE KOREAN SOCIETY of RADIOLOGY

Table 2. Performance Comparison of Large Language Models

Tasks (And Metrics for Performance)	GPT-3.5	GPT-4	Gemini	Others	Reference
Response to non-expert questions regarding lung cancer (Percentage of correct answers)	70.8%		51.7%		(21)
Multiple-choice questions regarding K-TIRADS (Percentage of correct answers)	73%	93%	80%	Perplexity (87%)	(41)
USMLE: Self Assessment and Sample Exam (Percentage of correct answers)	49.10%–58.78%	83.76%–86.70%			(44)
"Diagnosis Please" quizzes (Accuracy)	37.3%	57.1%			(32)
Radiology board-style multiple-choice questions (Percentage of correct answers)	69.3%	80.7%			(48)
Brazilian radiology board examinations (Percentage of correct answers)	47.5%–63.3%	65%–81.7%			(49)
Malignancy diagnosis based on TI-RADS (Accuracy)		78%–86%	74%–86%		(24)
BI-RADS category assignment (Percentage of incorrect categorization that would negatively impact clinical management)	14.3%	10.6%	18.1%	Human reader (1.5%)	(59)
Transformation of free-text reports into structured reports (F1 score)		28.11%–92.86%		medBERT.de (24.64%–92.53%)	(60)
Extraction of oncologic information from free-text CT reports (Accuracy)	83.9%–94.2%	97.2%–98.6%			(61)
Extraction of procedural details from free-text reports on mechanical thrombectomy (Percentage of correct data entries)	63.9%–64.2%	90.5%–94.0%			(62)
Detection of speech recognition error in radiology reports (F1 score)	32.2%–59.1%	86.9%–94.3%	20.9%–47.5%	Text-davinci-003 (46.6%–72%) LLaMA-2 (47.7%–72.8%)	(64)
Translation of radiology reports into plain language (Percentage of reports translated with good quality)	55.2%–77.2%	73.6%–96.8%			(65)
Simplification of radiology reports using plain language (Average reading grade level score)	6.7	7.0	9.1	Copilot (9.4)	(66)
Simplification of radiology report impressions using plain language (Average reading grade level score)	7.5	7.5	8.3	Copilot (8.9)	(67)
Determination of imaging modality and use of contrast agent (Percentage of correct answers)	64%–76%	74%–82%		accGPT (82%–86%)	(68)

accGPT = appropriateness criteria context aware GPT, BERT = bidirectional encoder representations from transformers, BI-RADS = Breast Imaging Reporting and Data System, GPT = generative pre-trained transformers, K-TIRADS = Korean TI-RADS, LLaMA = large language model meta AI, LLM = large language model, TI-RADS = Thyroid Imaging Reporting and Data System, USMLE = United States Medical Licensing Examination

모델에는 OpenAI의 contrastive language-image pre-training (이하 CLIP) (76)과 메타에서 개발한 segment anything model (이하 SAM) (77), 마이크로소프트의 large language and vision assistant (이하 LLaVA) (78)가 있다. 그러나 이러한 범용 목적의 비전-언어 모델들의 경우 전문적인 의료 영상에 대해서는 만족스러운 성능을 보여주지 못하였는데, 이는 온라인상에서 쉽게 구할 수 있는 일반적인 이미지들과 전문 의료 영상은 그 성격이 서로 너무 다르기도 하고, 또 공개되어 있는 실제 의료 영상 데이터의 수는 여전히 많이 부족하기 때문이었다. 이에 의료 영상 및 관련 지

식들에 미세조정 된 MedCLIP (79), BiomedCLIP (80), MedSAM (81), LLaVA-Med (82), RadFM (83) 등과 같은 모델들이 개발되었다(84). 이러한 의료 영상 전문 기반 모델들의 성능은 기존의 범용 목적의 모델들보다 더 뛰어난 벤치마크 결과를 보였고, 따라서 영상의학 관련 다양한 하부 작업들에 활용될 수 있을 것으로 기대된다. 실제로 2023년도 국제 영상의학 판독문 요약 컨테스트 (RadSum23) 결과를 보면 기존의 언어 기반 모델들보다 다양한 모달리티를 다루는 멀티모달 인공지능 모델들의 성적이 더 우수하였다(85). 의료 관련 작업에 한 해, 작업의 종류나 범위에 국한되지 않고 일반적으로 적용 가능한 의료 전문 기반 모델, 소위 generalist medial AI (GMAI)라는 컨셉이 제안되기도 하였는데, 이 개념은 필연적으로 멀티모달 입출력을 다룰 수 있어야만 한다(70). 그중 가장 중추가 되는 요소는 비전-언어 모델일 수밖에 없으며, 거기에 더해 청각, 촉각 등의 센서들이 결합되는 형식이 가장 유력한 구조로 기대되는 만큼, 비전-언어 모델의 발전 동향은 비단 영상의학 분야뿐 아니라 모든 의료 분야에 지대한 영향을 미칠 것이다.

결론

대형 언어 모델은 영상의학 전문가의 업무 및 연구 전반에 걸쳐 다양하게 활용될 수 있다. 대형 언어 모델의 작동 원리에 대한 개념적 이해와 초기 단계에서 지적되었던 문제점과 한계점들이 어떻게 해결되어 가고 있는지를 이해하는 것은 이 기술을 안전하게 활용하기 위한 최소한의 필요조건이다. 급속히 발전 중인 대형 언어 모델 기술이 현재 어느 수준까지 도달해 있고 또 어느 방향으로 나아가고 있는지, 어떤 종착점을 두고 인공지능 회사들 간에 경쟁이 이루어지고 있고 현재까지의 승자는 누구인지에 대해 고찰해 보는 과정은, 이 기술에 대한 이해를 더욱 깊게 해줄 것이고 궁극적으로 미래의 변화에 선제적으로 대비하는 데에 도움이 될 것이다.

Author Contributions

Conceptualization, all authors; project administration, K.S.S; resources, L.C.K, K.S.; supervision, K.S.S; visualization, K.S.; writing—original draft, K.S., K.S.S; and writing—review & editing, K.S., K.S.S.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

ORCID iDs

Sunkyu Kim https://orcid.org/0000-0002-0240-6210
Choong-kun Lee https://orcid.org/0000-0001-5151-5096
Seung-seob Kim https://orcid.org/0000-0001-6071-306X

Funding

None

REFERENCES

1. Harris ZS. Distributional structure. *Word* 1954;10:146-162
2. Le Q, Mikolov T. Distributed representations of sentences and documents. International conference on machine learning. Available at. https://proceedings.mlr.press/v32/le14.html?ref=https://githubhelp.com. Published 2014. Accessed May 1, 2024

3. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature* 1986;323:533-536
4. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;9:1735-1780
5. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.1406.1078. Accessed May 1, 2024
6. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Available at. https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html. Published 2014. Accessed May 1, 2024
7. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Available at. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. Published 2017. Accessed May 1, 2024
8. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.1810.04805. Accessed May 1, 2024
9. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Available at. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html. Published 2020. Accessed May 1, 2024
10. Jung KH. Uncover this tech term: foundation model. *Korean J Radiol* 2023;24:1038-1041
11. Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, et al. Scaling laws for neural language models. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2001.08361. Accessed May 1, 2024
12. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2206.07682. Accessed May 1, 2024
13. Golchin S, Surdeanu M. Time travel in LLMs: tracing data contamination in large language models. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2308.08493. Accessed May 1, 2024
14. Zhu K, Wang J, Zhao Q, Xu R, Xie X. Dynamic evaluation of large language models by meta probing agents. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2402.14865. Accessed May 1, 2024
15. Driess D, Xia F, Sajjadi MS, Lynch C, Chowdhery A, Ichter B, et al. PaLM-E: an embodied multimodal language model. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2303.03378. Accessed May 1, 2024
16. OpenAI. ChatGPT can now see, hear, and speak. Available at. https://openai.com/blog/chatgpt-can-now-see-hear-and-speak. Published 2023. Accessed May 1, 2024
17. OpenAI. Hello GPT-4o. Available at. https://openai.com/index/hello-gpt-4o. Published 2024. Accessed May 1, 2024
18. Zhang K, Yu J, Adhikarla E, Zhou R, Yan Z, Liu Y, et al. BiomedGPT: a unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. arXiv [Preprint]. Available at. https://arxiv.org/abs/2305.17100v2. Accessed May 1, 2024
19. Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang PC, et al. Towards generalist biomedical AI. *NEJM AI* 2024;1:AIoa2300138
20. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2311.16452. Accessed May 1, 2024
21. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. *Radiology* 2023;307:e230922
22. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. *NPJ Digit Med* 2023;6:195
23. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health* 2024;6:e12-e22
24. Wu SH, Tong WJ, Li MD, Hu HT, Lu XZ, Huang ZR, et al. Collaborative enhancement of consistency and accuracy in US diagnosis of thyroid nodules using large language models. *Radiology* 2024;310:e232255
25. Cao JJ, Kwon DH, Ghaziani TT, Kwo P, Tse G, Kesselman A, et al. Accuracy of information provided by ChatGPT regarding liver cancer surveillance and diagnosis. *AJR Am J Roentgenol* 2023;221:556-559
26. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology* 2023;307:e230424

27. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an adjunct for radiologic decision-making. medRxiv [Preprint]. Available at. https://doi.org/10.1101/2023.02.02.23285399. Accessed May 1, 2024
28. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? Available at. https://doi.org/10.1145/3442188.3445922. Published 2021. Accessed May 1, 2024
29. OpenAI. Create chat completion. Available at. https://platform.openai.com/docs/api-reference/chat/create. Accessed May 1, 2024
30. OpenAI. Custom instructions for ChatGPT. Available at. https://openai.com/blog/custom-instructions-for-chatgpt. Published 2023. Accessed May 1, 2024
31. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2307.09009. Accessed May 1, 2024
32. Li D, Gupta K, Bhaduri M, Sathiadoss P, Bhatnagar S, Chong J. Comparing GPT-3.5 and GPT-4 accuracy and drift in radiology diagnosis please cases. *Radiology* 2024;310:e232411
33. Spirling A. Why open-source generative AI models are an ethical way forward for science. *Nature* 2023;616:413
34. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Available at. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html. Published 2020. Accessed May 1, 2024
35. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: a medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus* 2023;15:e40895
36. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of using the privacy-preserving large language model Vicuna for labeling radiology reports. *Radiology* 2023;309:e231147
37. Le Guellec B, Lefèvre A, Geay C, Shorten L, Bruge C, Hacein-Bey L, et al. Performance of an open-source large language model in extracting information from free-text radiology reports. *Radiol Artif Intell* 2024;6:e230364
38. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 technical report. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2303.08774. Accessed May 1, 2024
39. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA* 2023;329:842-844
40. Haver HL, Gupta AK, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, et al. Evaluating the use of ChatGPT to accurately simplify patient-centered information about breast cancer prevention and screening. *Radiol Imaging Cancer* 2024;6:e230086
41. Kaba E, Hürsoy N, Solak M, Çeliker FB. Accuracy of large language models in thyroid nodule-related questions based on the Korean thyroid imaging reporting and data system (K-TIRADS). *Korean J Radiol* 2024;25:499-500
42. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2:e0000198
43. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312
44. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2303.13375. Accessed May 1, 2024
45. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. *NEJM AI* 2023;1:AIp2300031
46. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, et al. ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes. *Radiology* 2023;308:e231040
47. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology* 2023;307:e230582
48. Bhayana R, Bleakney RR, Krishna S. GPT-4 in radiology: improvements in advanced reasoning. *Radiology* 2023;307:e230987
49. Almeida LC, Farina EMJM, Kuriki PEA, Abdala N, Kitamura FC. Performance of ChatGPT on the Brazilian radiology and diagnostic imaging and mammography board examinations. *Radiol Artif Intell* 2024;6:e230103

50. Kim H, Kim P, Joo I, Kim JH, Park CM, Yoon SH. ChatGPT vision for radiological interpretation: an investigation using medical school radiology examinations. *Korean J Radiol* 2024;25:403-406
51. Zhou Y, Ong H, Kennedy P, Wu CC, Kazam J, Hentel K, et al. Evaluating GPT-V4 (GPT-4 with vision) on detection of radiologic findings on chest radiographs. *Radiology* 2024;311:e233270
52. Trivedi H, Wawira Gichoya J. The LLM will see you now: performance of ChatGPT on the Brazilian radiology and diagnostic imaging and mammography board examinations. *Radiol Artif Intell* 2024;6:e230568
53. Jiao W, Wang W, Huang JT, Wang X, Shi S, Tu Z. Is ChatGPT a good translator? Yes with GPT-4 as the engine. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2301.08745. Accessed May 1, 2024
54. Hwang SI, Lim JS, Lee RW, Matsui Y, Iguchi T, Hiraki T, et al. Is ChatGPT a "fire of prometheus" for non-native English-speaking researchers in academic writing? *Korean J Radiol* 2023;24:952-959
55. Koga S. The integration of large language models such as ChatGPT in scientific writing: harnessing potential and addressing pitfalls. *Korean J Radiol* 2023;24:924-925
56. Park SH. Use of generative artificial intelligence, including large language models such as ChatGPT, in scientific publications: policies of KJR and prominent authorities. *Korean J Radiol* 2023;24:715-718
57. Sun Z, Ong H, Kennedy P, Tang L, Chen S, Elias J, et al. Evaluating GPT4 on impressions generation in radiology reports. *Radiology* 2023;307:e231259
58. Kottlors J, Bratke G, Rauen P, Kabbasch C, Persigehl T, Schlamann M, et al. Feasibility of differential diagnosis based on imaging patterns using a large language model. *Radiology* 2023;308:e231167
59. Cozzi A, Pinker K, Hidber A, Zhang T, Bonomo L, Lo Gullo R, et al. BI-RADS category assignments by GPT-3.5, GPT-4, and Google Bard: a multilanguage study. *Radiology* 2024;311:e232133
60. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology* 2023;307:e230725
61. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. *Radiology* 2023;308:e231362
62. Lehnen NC, Dorn F, Wiest IC, Zimmermann H, Radbruch A, Kather JN, et al. Data extraction from free-text reports on mechanical thrombectomy in acute ischemic stroke using ChatGPT: a retrospective analysis. *Radiology* 2024;311:e232741
63. Gertz RJ, Dratsch T, Bunck AC, Lennartz S, Iuga AI, Hellmich MG, et al. Potential of GPT-4 for detecting errors in radiology reports: implications for reporting accuracy. *Radiology* 2024;311:e232714
64. Schmidt RA, Seah JCY, Cao K, Lim L, Lim W, Yeung J. Generative large language models for detection of speech recognition errors in radiology reports. *Radiol Artif Intell* 2024;6:e230205
65. Lyu Q, Tan J, Zapadka ME, Ponnatapuram J, Niu C, Myers KJ, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: promising results, limitations, and potential. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2303.09038. Accessed May 1, 2024
66. Doshi R, Amin K, Khosla P, Bajaj S, Chheang S, Forman HP. Utilizing large language models to simplify radiology reports: a comparative analysis of ChatGPT3.5, ChatGPT4.0, Google Bard, and Microsoft Bing. medRxiv [Preprint]. Available at. https://doi.org/10.1101/2023.06.04.23290786. Accessed May 1, 2024
67. Doshi R, Amin KS, Khosla P, Bajaj SS, Chheang S, Forman HP. Quantitative evaluation of large language models to streamline radiology report impressions: a multimodal retrospective analysis. *Radiology* 2024;310:e231593
68. Rau A, Rau S, Zoeller D, Fink A, Tran H, Wilpert C, et al. A context-based chatbot surpasses trained radiologists and generic ChatGPT in following the ACR appropriateness guidelines. *Radiology* 2023;308:e230970
69. Gertz RJ, Bunck AC, Lennartz S, Dratsch T, Iuga AI, Maintz D, et al. GPT-4 for automated determination of radiological study and protocol based on radiology request forms: a feasibility study. *Radiology* 2023;307:e230877
70. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. *Nature* 2023;616:259-265
71. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature* 2023;620:172-180
72. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2305.09617. Accessed May 1, 2024

73. Wang G, Yang G, Du Z, Fan L, Li X. ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2306.09968. Accessed May 1, 2024
74. Liu Z, Zhong A, Li Y, Yang L, Ju C, Wu Z, et al. Radiology-GPT: a large language model for radiology. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2306.08666. Accessed May 1, 2024
75. Liu Z, Li Y, Shu P, Zhong A, Yang L, Ju C, et al. Radiology-Llama2: best-in-class large language model for radiology. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2309.06419. Accessed May 1, 2024
76. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. Available at. https://proceedings.mlr.press/v139/radford21a. Published 2021. Accessed May 1, 2024
77. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment anything. Available at. https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html. Published 2023. Accessed May 1, 2024
78. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. Available at. https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html. Published 2023. Accessed May 1, 2024
79. Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: contrastive learning from unpaired medical images and text. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2210.10163. Accessed May 1, 2024
80. Zhang S, Xu Y, Usuyama N, Xu H, Bagga J, Tinn R, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2303.00915. Accessed May 1, 2024
81. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. *Nat Commun* 2024;15:654
82. Li C, Wong C, Zhang S, Usuyama N, Liu H, Yang J, et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. Available at. https://proceedings.neurips.cc/paper_files/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html. Published 2023. Accessed May 1, 2024
83. Wu C, Zhang X, Zhang Y, Wang Y, Xie W. Towards generalist foundation model for radiology. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2308.02463. Accessed May 1, 2024
84. Azad B, Azad R, Eskandari S, Bozorgpour A, Kazerouni A, Rekik I, et al. Foundational models in medical imaging: a comprehensive survey and future vision. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2310.18689. Accessed May 1, 2024
85. Delbrouck JB, Varma M, Chambon P, Langlotz C. Overview of the RadSum23 shared task on multi-modal and multi-anatomical radiology report summarization. Available at. https://doi.org/10.18653/v1/2023.bionlp-1.45. Published 2023. Accessed May 1, 2024

JOURNAL of
THE KOREAN SOCIETY of
RADIOLOGY

대형 언어 모델: 영상의학 전문가를 위한 종합 안내서

김선규[1,2] · 이충근[3] · 김승섭[4*]

대형 언어 모델은 자연어 처리 분야에 국한되지 않고 기술 산업의 거의 모든 분야에서부터 일상생활에 이르기까지, 전 지구적인 혁신을 가져왔다. 방대한 데이터셋에 대한 광범위한 사전 훈련 덕분에 현대의 대형 언어 모델들은 일반적인 작업뿐 아니라 의료 영상과 같은 전문적인 분야의 작업까지 수행 가능하게 되었다. 업체들은 매우 빠른 속도로 버전 업데이트 및 신규 모델 출시를 발표하고 있고, 그로 인해 초기에 지적되었던 여러 문제점과 한계점들이 하나씩 해결되어 가고 있다. 또한 초기의 스케일링 업 방식의 발전 방향성에서 탈피하여 최근에는 작아진, 온프레미스 오픈 소스 대형 언어 모델의 개념이 주목받고 있고, 이로 인해 전문 의료지식에 대한 미세조정, 훈련 효율성 제고, 개인정보 문제 해결, 성능 변동 관리 등의 이슈들이 해결되어 가고 있다. 본 종설은 대형 언어 모델을 활용하려는 영상의학 전문가에게, 관련 기술에 대한 개념적 지식과 실용적인 지침, 그리고 현재의 기술 지형과 미래 방향성 등을 통합적으로 제공하고자 작성되었다.

[1]고려대학교 정보대학 컴퓨터통신공학과,
[2](주)아이젠사이언스,
[3]연세대학교 의과대학 내과학교실 종양내과,
[4]연세대학교 의과대학 세브란스병원 영상의학과, 방사선의과학연구소