Review Article

J Korean Soc Radiol 2024;85(5):848-860
https://doi.org/10.3348/jksr.2024.0066
eISSN 2951-0805

JOURNAL of THE KOREAN SOCIETY of RADIOLOGY

Primer on Generative Artificial Intelligence and Large Language Models in Medical Imaging

의료영상에서 생성형 인공지능과 대형 언어 모델 입문

Kiduk Kim, MD[1], Gil-Sun Hong, MD[2]*, Namkug Kim, PhD[1,2]*

[1]Department of Convergence Medicine, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Korea
[2]Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Korea

Received May 10, 2024
Revised August 22, 2024
Accepted September 19, 2024

***Corresponding author**
Gil-Sun Hong, MD
Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea.

Tel 82-2-3010-4352
Fax 82-2-3010-0090
E-mail hgs2013@gmail.com

Namkug Kim, PhD
Department of Convergence Medicine, Department of Radiology and Research Institute of Radiology University of Ulsan, College of Medicine, Asan Medical Center, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea.

Tel 82-2-3010-6573
Fax 82-2-3010-6196
E-mail namkugkim@gmail.com

The recent advent of large language models (LLMs), such as ChatGPT, has drawn attention to generative artificial intelligence (AI) in a number of fields. Generative AI can produce different types of data including text, images, and voice, depending on the training methods and datasets used. Additionally, recent advancements in multimodal techniques, which can simultaneously process multiple data types like text and images, have expanded the potential of using multimodal generative AI in the medical environment where various types of clinical and imaging information are used together. This review summarizes the concepts and types of LLMs, image generative AI, and multimodal AI, and it examines the status and future possibilities of generative AI in the field of radiology.

Index terms Generative Artificial Intelligence; Large Language Model; Language Vision Model; Image Generative AI; Language Generative AI

서론

최근 OpenAI에 의해 개발된 Chat Generative Pre-trained Transformer (이하 ChatGPT)의 출현으로 생성형 인공지능이 의료에서도 흥미를 끌고 있다. 특히 의료 영상 분야는 적대적 생성 신경망(generative adversarial network; 이하 GAN) (1), 확산 확률 모델(diffusion model) (2), 대형 언어 모델(large language model; 이하 LLM) (3, 4) 등 다양한 종류의 생성형 인공지능을 활용할 수 있는 중요한 영역으로 대두되고 있다.

GAN과 확산 확률 모델과 같은 이미지 생성형 인공지능은 이미지 재건축(image re-

construction)과 이미지 질 향상(image quality improvement과 같은 영역에 주로 활용되고 있으며, 또한 개인 정보의 문제가 없는 현실과 유사한 합성 데이터를 만들어낼 수 있는 능력으로 데이터 프라이버시 영역에서의 대안으로 떠오르고 있다.

ChatGPT나 GPT-4와 같은 LLM과 시각언어모델(vision-language model; 이하 VLM)과 같은 자연어 처리 분야의 생성형 인공지능은 일반적인 용도의 챗봇부터 영상의학과 같은 다양한 도메인 특화 영역에서 추가적인 미세조정 없이도 사용할 수 있는 인공지능 모델로 각광받고 있다.

따라서 본 종설에서는 이러한 이미지와 언어 분야 등 다양한 생성형 인공지능의 영상의학 분야 활용 현황 및 가능성에 대해 종합적으로 알아보고자 한다.

생성형 인공지능(Generative AI)

회귀 모델이나 분류 모델과 같이 정해진 입력값에 따라 결과가 정해지는 결정론적 모델(deterministic model) 들과 달리, 생성형 인공지능은 무작위성을 포함하고 있고 이에 따라 확률론적 모델(probabilistic model)이라는 이름으로도 불린다. 이러한 생성형 인공지능은 이미지, 언어, 혹은 둘 모두 등 그들이 처리하는 데이터의 유형에 따라 분류할 수 있다(5).

이미지 생성형 인공지능(Image Generative AI)

이미지 생성형 인공지능의 경우 주로 변이형 오토인코더(variational auto-encoder; 이하 VAE)(6)와 GAN이 주로 사용되며 최근 확산 확률 모델이 뛰어난 성능으로 주목받고 있다.

VAE (6)는 이미지를 입력값으로 받아 이를 잠재 공간(latent space)에 임베딩하는 encoder와 잠재 공간 속의 벡터값 z (latent vector z)을 입력값으로 받아 이로부터 이미지를 생성해 내는 decoder로 이루어져 있다. 어떤 이미지 x가 주어졌을 때 latent vector z의 분포를 정규분포에 근사한다면, 이 정규분포의 평균(μ)과 표준편차(σ)를 찾는 것이 encoder의 역할이 된다. 여기서 decoder가 주어진 μ와 σ를 이용해서만 이미지를 생성해 낸다면 처음 입력값 x와 동일한 이미지만을 만들어낼 수 있기 때문에 VAE는 latent space에서 평균이 0이고 표준편차가 1인 표준정규분포로부터 noise ε을 sampling 하여 encoder로부터 얻은 σ를 곱하고 μ을 더하여 얻은 latent vector z로부터 이미지를 생성해 내고, 이를 reparameterization trick이라고 한다. VAE는 이 reparameterization trick을 이용해 다양한 이미지를 생성해 낼 수 있다.

GAN (1)은 noise z를 입력값으로 받아 이미지를 생성해 내는 generator와 생성된 이미지와 실제 이미지를 비교하여 주어진 이미지가 생성된 이미지인지 실제 이미지인지를 구별하는 discriminator로 이루어져 있다. GAN의 generator와 discriminator를 설명할 때 흔히 사용되는 비유는 위조범과 탐정의 관계이다. 위조범 generator는 탐정 discriminator를 속이기 위해 사실적인 모조품을 만든다. 탐정은 주어진 그림이 모조품인지 실제 그림인지를 분간한다. 반복적인 경쟁관계를 통해 위조범은 점점 더 사실적인 모조품을 만들어내고, 탐정도 점점 더 사실적인 모조품과 실제 그림을 잘 분간해 내다가 어느 순간 너무 완벽한 모조품이 생기면 탐정은 이게 실제인지 모조

품인지를 분간하지 못하고 확률에 이를 맡기게 되며 학습이 끝난다.

확산 확률 모델(2)은 이미지 생성형 인공지능에 새로운 패러다임을 제시한다. 확산 확률 모델은 주어진 이미지 x_0에 전체 T 시간 중 일정한 시간 t마다 noise를 이미지에 첨가하여 완전한 noise 이미지 x_T가 될 때까지 noise를 이미지에 '확산'시키는 forward process (diffusion process)와 완전한 noise 이미지 x_T로부터 noise가 없는 온전한 이미지 x_0를 만드는 reverse process로 이루어져 있다. 이 과정에서 시간 t-1에서의 이미지의 상태 x_{t-1}는 바로 이전 시간 t에서의 이미지의 상태 x_t에만 의존하는 이산-시간 확률 과정(discrete-time stochastic process)인 Markov chain에 기반하여 noise를 첨가한다. 이를 이용하여 diffusion model은 시간 t에서 noised 된 이미지 x_t에서 시간 t-1만큼 noised 된 이미지 x_{t-1}을 만들어내는 방식(denoising)을 원본 이미지 x_0를 만들 때까지 반복하여 이미지를 생성해 낸다.

언어 생성형 인공지능: 대형 언어 모델(Large Language Model)

기존의 자연어 처리(natural language processing) 방법론은 주로 순환신경망(recurrent neural network) (7)과 장단기 메모리(long short-term memory) (8)를 이용했고 이는 전파 소실(gradient vanishing)이나 병렬처리 불가능, 비효율적인 연산 등의 문제가 있었다. Transformer (9) 구조는 self-attention이라는 메커니즘을 통해 이러한 문제들을 상당 부분 극복해 냈고, 이에 더해 Transformer의 장점인 계산 효율성(efficiency)과 확장성(scalability)으로 인해 엄청난 성능을 보여주는 LLM의 시대가 도래했다(10).

Transformer (9)는 인간이 이해할 수 있는 언어의 의미론적인 단위를 인공지능이 이해할 수 있는 단위인 vector, token으로 변환시켜 주는 embedding 단계, 이렇게 embedding 된 vector를 전달받아 vector들 사이의 관계를 여러 head를 갖는 attention 함수(multi-head self-attention; 이하 MHSA)를 통해 계산한 결과를 N층 쌓은 Encoder 단계, Encoder로부터 받은 데이터를 다시 M층 쌓인 MHSA를 통해 계산한 결과값을 Dense layer와 Softmax layer를 통해 결과를 뽑는 Decoder 단계로 구성된다. Transformer의 핵심이 되는 Attention 메커니즘은 네트워크에 입력된 여러 토큰들을 처리할 때 서로 다른 토큰들(query, key, value)에 주목할(attend) 수 있게 하는 메커니즘이다. 이를 계산할 때는 주로 서로 다른 토큰들의 내적 값의 가중합으로 계산된다. 이 Attention 메커니즘은 주로 encoder-decoder 구조에서 decoder에게 encoder의 input을 주목하게 하는 등 서로 다른 구조 사이에서 서로의 정보를 참조하게 하는 방식으로 이용되었는데, Transformer의 self-attention은 현재 처리 중인 시퀀스 내에서 맥락(context)을 볼 수 있게 하는 방식으로 사용되었다. 이를 통해 트랜스포머는 기존의 순환신경망에서 지적되던 벡터 압축 중 정보 손실과 기울기 소실(vanishing gradient) 문제를 해결할 수 있었다. Fig. 1은 Transformer의 구조를 묘사한다.

많은 수의 상업화된 현대 LLM 모델들의 구체적인 네트워크 구조와 학습방법이 알려져 있지 않지만, 이전에 공개된 GPT 모델(3, 11-13) 들과 구글의 bidirectional encoder representations from transformers (이하 BERT) 모델(4)의 학습 방법과 구조를 살펴보는 것은 현대 LLM의 이해

Fig. 1. Typical structure of a transformer.

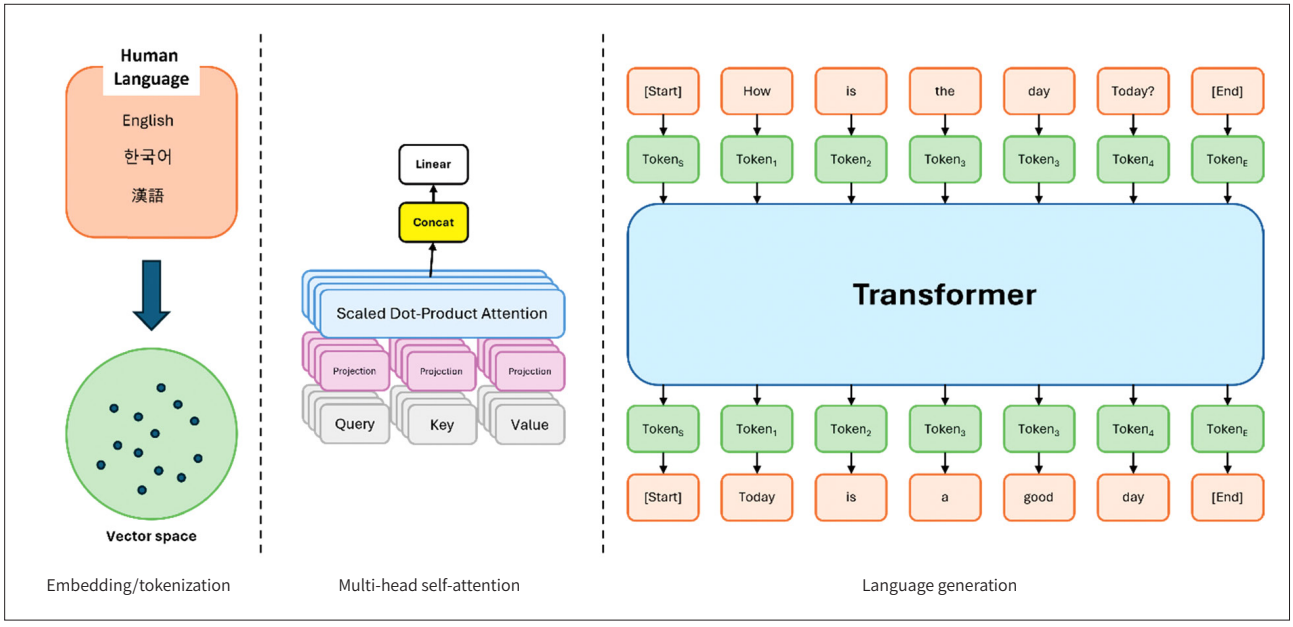
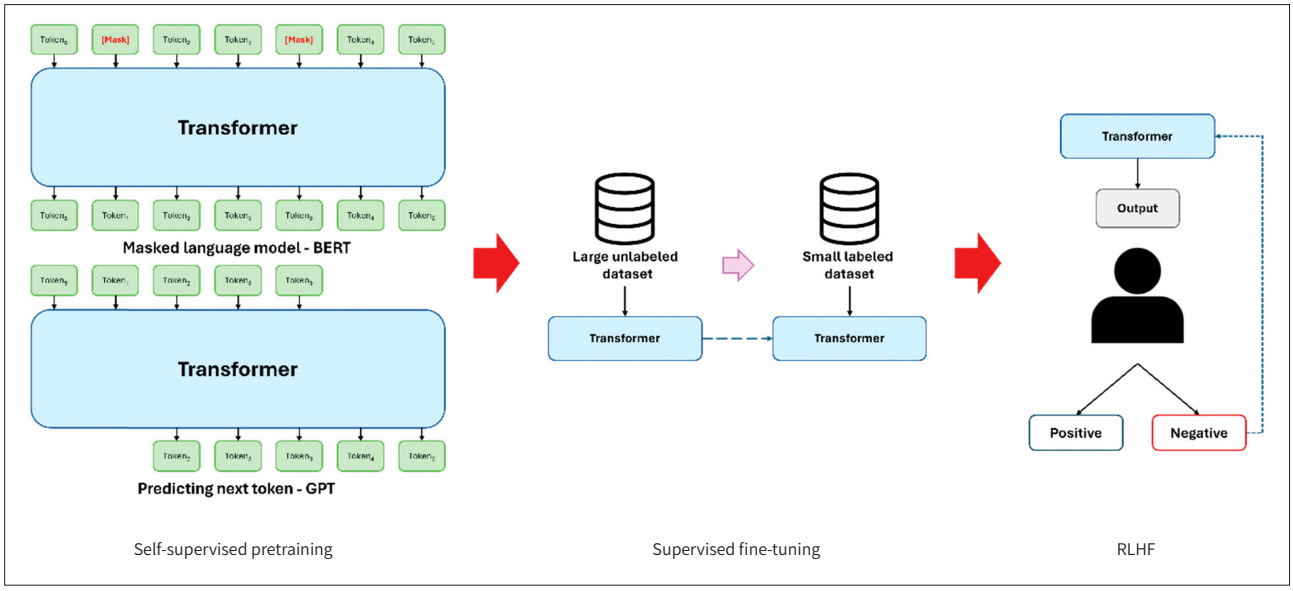


Fig. 2. Typical modern day training pipeline of large language models.



BERT = bidirectional encoder representations from transformers, GPT = generative pre-trained transformer, RLHF = reinforcement learning from human feedback

에 도움을 줄 수 있다. 거대한 모델을 학습시키는 데에는 많은 양의 데이터가 필요하기 때문에, 이 모든 데이터에 일일이 label을 부여하여 지도 학습(supervised learning)의 방식으로 학습시키는 데에는 천문학적인 비용이 발생한다. 이에 따라 GPT와 BERT 모두 모델의 성능을 개선하기 위해서 모델을 잘 학습시키기 위한 임의의 과제를 정해 학습하는 자기지도학습(self-supervised learning; 이하 SSL) 방법론을 사용했다. GPT의 경우 이전의 토큰들의 정보를 이용하여 다음에

올 토큰을 맞추는, 다시 말해 이전까지의 맥락 정보를 이용하여 다음에 올 단어가 무엇인지를 맞추는 SSL 방법을 이용한다(3). BERT의 경우 주어진 문장에서 무작위로 몇 개의 단어를 가리고 가려진 단어를 예측하는 방식의 SSL 방법을 이용하여 문장의 맥락을 배운다. GPT는Transformer의 decoder만 이용하는 네트워크 구조로, BERT는 Transformer의 encoder만 이용하는 네트워크 구조로 이루어져 있다. 그러나 이러한 SSL 만으로 학습된 언어 모델은 충분히 만족스러운 성능을 보여주기 어렵다. 이에 SSL으로 학습된 언어모델을 좋은 질의 고도화된 instruction dataset을 일부 이용하여 지도학습의 방식으로 미세 조정(fine-tuning)하고(11, 12), 이후 인간의 피드백을 통해 보상을 주는 강화학습의 방식(reinforcement learning from human feedback)으로 더 성능을 향상시켜 이용한다(13). Fig. 2는 일반적인 현대 LLM 학습 파이프라인을 묘사한다.

멀티모달 인공지능: 비전-언어 모델(Vision-Language Model)

멀티모달 인공지능(multi-modal AI)의 발달로 이미지만을 처리하고 생성하는 이미지 생성형 인공지능과 언어만을 처리하고 생성하는 언어 생성형 인공지능뿐만 아니라 이미지와 언어 모두를 처리할 수 있는 비전-언어모델(VLM) 역시 가능해졌다(14).

VLM은 시각적 입력(visual input) 혹은 텍스트 입력(textual input) 혹은 둘 모두를 입력(multi-modal input)으로 받아 처리할 수 있고, 처리한 결과를 이미지로 생성(image generation)하거나 텍스트를 생성(textual generation)하거나 혹은 전통적인 분류, 검출, 회귀 등의 이미지 처리 과제를 할 수도 있다. 결과를 처리하는 방식에 따라 크게 생성 과제(generation task)와 인식 과제(perception task)로 나눌 수 있고(5), 생성 과제에는 대표적으로 시각-질의-응답(visual question answering) (15), 시각적 추론(visual reasoning) (16), 시각 캡셔닝(visual captioning) (17), 시각 생성(visual generation) (18) 등이 있고 인식 과제에는 영상 인식(image recognition) (19, 20), 시각 그라운딩(visual grounding) (21), 영상 검색(image retrieval) (22) 등이 있다. 또한 이에 더해 프롬프팅(prompting)을 이용하여 영상을 분할(image segmentation)하는 일반적 분할 모델 (general segmentation model) (23, 24) 역시 비전-언어 모델을 통해 가능해졌다.

의료 영상에서의 이미지 생성형 인공지능의 활용과 한계

최근의 이미지 생성형 인공지능은 GAN과 확산 확률 모델을 주로 사용한다. GAN과 확산 확률 모델 모두 주어진 데이터의 분포를 학습하여 영상을 생성하지만, 이미지를 생성해 내는 방법과 그 특성에서 차이가 다르다. GAN (1)은 generator와 discriminator를 경쟁시키는 방법으로 학습하여 더 빠른 속도로 학습되고 더 적은 수의 데이터로도 학습하는 방법이 있지만, 확산 확률 모델(2)은 이미지에 반복적으로 noise를 주입하고 denoising하는 방법으로 학습함으로써 noise의 크기만큼 섬세한 특징부터 이미지의 전체적인 거친 특징까지 coarse-to-fine grained 특징을 모두 배울 수 있어 영상의 생성 품질이 더 좋은 편이고(25) 모드 붕괴(26, 27)에 빠질 위험이 적다.

GAN이나 확산 확률 모델과 같은 방법론으로 이미지 자체를 생성해 낼 수도 있지만, 활용을 위

해서는 주로 조건부 생성(conditional generation), 제어 가능한 생성(controllable generation)의 방법을 이용하게 된다. 이러한 controllable generation의 방법에는 크게 conditional image synthesis, image editing and manipulation, image-to-image translation이 있다. 엄밀하게 이 셋이 구분되지는 않지만 이러한 방법론을 아는 것이 이미지 생성형 인공지능의 활용을 아는 데에 도움이 될 수 있다.

조건부 이미지 생성(Conditional Image Synthesis)

Conditional image synthesis는 조건부 확률 분포를 이용하여 원하는 클래스 혹은 원하는 상태를 가진 이미지를 생성하는 방법이다. Conditional GAN (28)은 조건부 확률 분포를 직접 이용하여 이미지를 생성한다. 2020년 Nishio 등(29)은 conditional GAN을 이용하여 3D CT에서 생성한 lung nodule image 만으로 학습한 분류기로 실제 CT image를 이용하여 학습한 분류기만큼의 성능을 보여주었다. 확산 확률 모델들은 conditional image synthesis를 위해 분류기-유도 확산 모델(classifier guided diffusion model) (25)이나 분류기 없는 유도 확산 모델(classifier-free guidance diffusion model) (30)을 이용한다. 2024년 Moon 등(31)은 분류기-유도 확산 모델을 통한 data augmentation 기법으로 isocitrate dehydrogenase (이하 IDH)-wild type glioma와 IDH-mutant type glioma를 생성하여 학습한 분류기로 신경영상의학과 의사에 비교할 만한 성능을 보여주었다.

이미지 편집과 조작(Image Editing and Manipulation)

Image editing and manipulation은 주어진 합성 영상 혹은 실제 영상이 주어졌을 때 이를 편집하고 조작하는 기술이다. GAN architecture에서는 주로 잠재 공간을 조작하는 방식으로 image editing and manipulation이 이루어진다. 잠재 공간은 실제 영상을 각 vector에 대응시킨 공간으로, 주어진 실제 영상을 이 잠재 공간에 대응시키는 것을 GAN inversion (32)이라고 한다. 2023년 Lee 등(33)은 GAN inversion을 통해 정상과 척추 측만증 영상을 만들어낼 수 있고, dextroscoliosis에서 levoscoliosis 방향으로 영상을 편집할 수 있음을 시사했다. 확산 확률 모델에서의 image editing and manipulation 역시 잠재 공간의 편집(34)이나 추가적인 인코더를 이용하여(35) 이루어질 수 있지만, 확산 확률 모델의 이미지 편집 기법은 최근 대부분 프롬프트를 이용하는 텍스트-유도 편집 방식을 이용한다(36). 확산 확률 모델을 이용한 의료 영상의 편집에 대한 연구나 활용은 아직 많지 않지만 확산 확률 모델의 성능으로 미루어 보아 무궁한 활용의 여지가 남아있다.

이미지 대 이미지 변환(Image-To-Image Translation)

Image-to-image translation 혹은 스타일 전이(style transfer)라고 불리는 기술은 영상의 구조와 내용물은 보존하면서 이미지의 스타일을 바꾸는 기술이다. Image-to-image translation은 동

일한 위치에 스타일만 다른 동일한 구조물들이 존재하는 완벽한 pair가 존재할 때 할 수 있는지 여부에 따라 지도학습(37) 혹은 비지도학습(38) 방법론으로 시행할 수 있지만, 대부분의 의료영상은 완벽한 pair가 존재하기 어렵기 때문에 비지도학습 방법론을 위주로 시행되고 있다. 2023년 Choi 등(39)은 multi-domain image-to-image translation 기술을 이용해 이미 촬영된 CT 이미지를 소프트 커널과 샤프 커널로 상호 변환할 수 있는 네트워크를 제시했다. 또한 2024년 Kim과 Park (40)은 확산 확률 모델을 이용하여 MRI에서 T1, T1 contrast enhancement, T2, and fluid attenuated inversion recovery 등 다양한 sequence에서 서로 image-to-image translation을 하 수 있는 네트워크를 제시했다.

이미지 생성형 인공지능의 함정과 한계 (Pitfalls and Limitations of Image Generative AI)

이미지 생성형 인공지능이 인상적인 성능을 보여주고 있지만, 작은 문제가 환자에게는 큰 영향으로 다가올 수 있는 민감한 영역인 의료의 특성을 고려할 때 우리는 이미지 생성형 인공지능의 함정과 한계를 잘 인지하고 있어야 한다.

먼저 이미지 생성형 인공지능을 학습할 때 GAN은 모드 붕괴(26, 27)로 충분히 다양한 영상을 생성해 내지 못할 수 있고, 확산 확률 모델은 반복을 통한 생성이 필요해 학습과 생성에 오랜 시간을 필요로 할 수 있다. 또한 이러한 모델들을 학습시키고 활용하는 데에는 상당한 양의 자원과 시간을 필요로 하기 때문에(41, 42), 이런 자원이 제한적인 환경에서는 생성형 인공지능이 활용이 불가능하거나 알려진 수준만큼의 성능을 보이지 못할 수 있다. 마지막으로 영상을 생성하는 과정에서 생성형 인공지능에 의한 인공물이 생길 수 있으며(43), 인공지능의 '블랙박스'적인 특성으로 인해 이 결과를 해석하는 것은 여전히 과제로 남아있다(44).

따라서 영상 생성형 인공지능을 임상에서 활용하기 위해서는 이러한 한계점들을 정확히 인지하고 있어야 한다. 이에 더해 생성형 인공지능의 활용에는 이런 한계와 신뢰성을 잘 알고 있는 전문가들이 꾸준히 결과를 모니터링하며 질을 관리하여 사용해야 한다. 생성형 인공지능의 활용에 관련된 가이드라인 역시 이러한 이해와 질 관리에 도움을 줄 수 있다(5).

의료 영상에서의 LLM과 VLM의 활용과 한계

전통적으로 딥 러닝에서 생성형 인공지능은 주로 영상의 생성형 인공지능이 주목을 받아왔지만, ChatGPT, Gemini, Claude 등의 상업화된 LLM이 엄청난 성능을 보여주며 현재 생성형 인공지능에서는 LLM 등의 언어의 생성형 인공지능이 큰 주목을 받게 되었다. 이에 더해 이전에는 언어, 영상, 신호 등 단일 형태의 데이터만 처리가 가능했지만 최근 멀티모달 인공지능 기술의 발달로 LLM에 영상 처리 기술을 통합한 VLM 기술 역시 이목을 끌고 있다. 또한 LLM을 더 효율적으로 사용하기 위해 프롬프트 엔지니어링(prompt engineering) (45), 효율적 파라미터 파인튜닝(parameter-efficient fine-tuning; 이하 PEFT) (46), 검색 증강 생성(retrieval-augmented gen-

eration; 이하 RAG) (47) 등의 기술이 제시되었고 의료 영상에서도 역시 판독지 요약, 판독지 구조화 등에 LLM을 적용하는 등의 활용 방안이 제시되고 있다.

프롬프트 엔지니어링, PEFT, RAG

LLM의 성능을 개선하고 더 효율적으로 사용하기 위해 최근 프롬프트 엔지니어링, PEFT, RAG 기술 등이 도입되었다. 이 기술들은 추가적인 학습이 없거나 제한된 데이터를 이용한 약간의 학습만으로 원하는 과제를 더 효율적으로 수행할 수 있게 해준다.

프롬프트 엔지니어링(45)은 LLM의 입력으로 주어지는 텍스트의 형식, 내용, 구조 등을 조정하여 원하는 결과를 얻게 해준다. 프롬프트 엔지니어링은 모델의 설정, 출력의 유형과 형식, 모델이 수행할 작업 혹은 지시, 추가적인 문맥, 그리고 응답받고자 하는 입력이나 질문 등으로 구성할 수 있다. 가장 먼저 LLM의 모델의 역할을 설정할 수 있다. LLM에게 특정 역할과 인격을 부여하는 역할극(role-playing)은 LLM에게 원하는 결과를 유도할 수 있다. 모델이 수행할 작업을 간결하게 설명하거나 어떤 식으로 수행할지를 간결하게 설명하는 것 역시 원하는 결과를 얻는 데에 도움이 될 수 있다. 원하는 질문과 유사한 질문과 그에 해당하는 답을 프롬프트에 몇 가지 주고 답을 유도하는 퓨 샷 프롬프팅(few-shot prompting)이나 "Let's think step by step" 등의 LLM에게 중간 추론 단계를 거치도록 하는 생각의 사슬 프롬프팅(chain-of-thought; CoT prompting)도 흔히 사용되는 프롬프트 엔지니어링 기술이다.

일반적으로 언어 모델의 실제 활용을 위해서는 원하는 특정 데이터에 특정 과제를 수행하기 위해 미세 조정을 수행해야 한다. 그러나 LLM의 시대로 오면서 각 언어 모델들이 너무 많은 파라미터 수를 갖게 되어 모든 파라미터를 미세 조정을 하는 것은 시간, 메모리, GPU, 데이터셋 크기 등 현실적인 어려움에 부딪혔다. 이에 시간과 자원 효율적인 미세 조정을 위하여 낮은 intrinsic rank의 미세 조정 만으로 충분한 성능을 얻는 low-rank adaptation (LoRA) 방법론이 소개되었다(46).

검색 증강 생성(47) 기술은 LLM에서 흔히 지적받는 인공지능 모델 업데이트 시점으로부터 뒤처진 정보 문제나 환각(hallucination) 문제를 극복하기 위해 소개된 기술이다. 프롬프트(prompt)와 질문(query)이 RAG 파이프라인에 주어지면, RAG 파이프라인은 외부 지식 소스로부터 관련된 정보를 검색하여 강화된 맥락 정보를 제공하여 프롬프트에 더해 LLM에 전달하고, 이에 대한 답을 제공한다.

의료 영상에서 대형 언어 모델의 활용

의료에서의 LLM은 EHR에 BERT를 임베딩한 시스템 스케일의 활용(48)이나, 여러 LLM을 활용하여 서로에게 피드백을 주는 방식으로 원격 일차 의료 상황에서 의사 수준 이상의 성능을 보여주는 등(49) 다양한 활용이 있다. 의료 영상에서의 LLM 역시 다양한 활용 가능성을 보여주고 있다(50). 2023년 Adams 등(51)은 다양한 언어로 이루어진 Chest radiograph, CT, MRI의 영상 판독지를 GPT-4와 MedBERT를 이용해 구조화된 판독지로 바꿀 수 있는 가능성을 보여주었다. 또

JOURNAL of
THE KOREAN SOCIETY of
RADIOLOGY

2023년 Gertz 등(52)은 GPT-4를 이용하여 환자의 임상 정보에 따라 어떤 body region에 어떤 modality의 영상을 찍어야 하는지, 조영제를 사용해야 하는지 study protocol을 알 수 있음을 보여주었다. 또한 영상 판독지에서 특정한 정보를 추출해 내거나 구조화된 판독지로 바꾸는 등 의학적 활용 외에도, 환자가 이해하기 쉬운 언어로 풀어주는 것 역시 가능하다(53).

의료 영상에서 비전-언어 모델의 활용

의료 영상에서는 판독지도 물론 중요하지만 기본적으로 영상에 많은 정보가 담겨있기 때문에 LLM만으로는 충분하지 않을 수 있다. 따라서 대형 VLM의 출현은 의료 영상에서의 인공지능 활용에 새로운 국면을 맞이하게 할 수 있다. GPT-4V, RadFM (54) 등 다양한 멀티모달 인공지능들이 의료 영상에서의 활용 가능성을 제시했다. 또한 카카오에서는 KARA-CXR (https://karacxr.ai)이라는 chest radiograph을 입력으로 받아 예비 판독문(preliminary report)을 작성하는 상업화된 멀티모달 인공지능을 공개하기도 했다.

대형 언어 모델과 비전-언어 모델의 함정과 한계 (Pitfalls and Limitations of LLM and VLM)

LLM과 VLM은 의료 영상 분야에서 인공지능의 무궁한 활용 가능성을 시사했지만, 생성형 인공지능들이 갖는 여러 문제점 역시 가지고 있다. 사실이 아니거나 알 수 없는 정보를 LLM이 사실인 것처럼 확신에 차 답변하는 환각(hallucination) 문제가 LLM에 가장 흔히 제기되며 가장 클 수 있는 문제이다(5, 50, 55). 인공지능을 학습하는 데이터에 정보가 없거나 편향이 끼어 있을 때 생길 수 있는 문제이다. 웹 검색을 통해 현실의 정보를 포함시킬 수 있는 RAG가 이런 문제의 해결책으로 사용할 수 있을 것으로 보인다. 또한 웹으로 정보를 전송하지 않고 로컬에서 학습시킬 수 있는 LLaMa (56)와 같은 모델들을 제외한 GPT, Gemini, Claude 등의 상업화된 모델들은 환자의 정보가 입력을 통해 OpenAI, Google, Anthropic 등의 회사로 전송될 수 있는 개인정보 문제가 있을 수 있다. 따라서 환자의 개인정보 보호를 위해 환자 정보를 충분히 익명화/가명화한 정보만을 입력하거나 웹으로 전송하지 않는 모델을 활용하는 등의 노력이 필요하다(57).

결론

본 종설에서는 생성형 인공지능에 대해 개략적으로 알아보고 의료 영상에서 다양한 층위의 생성형 인공지능이 어떻게 활용될 수 있는지 사례를 통해 알아보았다. 생성형 인공지능은 다양한 분야에서 뛰어난 성능을 보여주고 있고, 의료 영상 영역에서도 많은 활용 가능성을 보여주고 있다. 하지만 환각, 인공물 생성 등 민감한 의료 영역에서 큰 문제가 될 수 있는 단점들도 존재한다. LLM이 사실이 아닌 정보를 사실인 것처럼 전달할 수도 있고, 이미지 생성형 인공지능이 현실에 존재하지 않는 인공물을 영상에 임의로 생성해 낼 수 있다. 따라서 다양한 생성형 인공지능의 함

정과 한계를 정확히 파악하고 주의를 기울여야 한다. 또한 하나의 문제도 환자에게는 크게 작용할 수 있고 개인정보 등 다양한 민감한 문제가 있을 수 있는 의료에서는 이러한 생성형 인공지능의 환각, 인공물 여부를 감별할 수 있는 전문가의 역할이 더욱 중요하다. 향후 의료 영상 분야에서 전문가들의 업무 부담을 줄이기 위해 생성형 인공지능을 워크플로우에 포함하며 또한 인공지능의 한계를 파악할 수 있는 전문가의 역할이 더욱 중요해질 것으로 기대한다.

Supplementary Materials

English translation of this article is available with the Online-only Data Supplement at https://doi.org/10.3348/jksr.2024.0066.

Author Contributions

Conceptualization, K.K., K.N.; funding acquisition, K.N.; project administration, H.G., K.N.; supervision, H.G., K.N.; writing—original draft, K.K.; and writing—review & editing, all authors.

Conflicts of Interest

Namkug Kim has been an Editorial Board Member of the Journal of the Korean Society of Radiology since 2021; however, he was not involved in the peer reviewer selection, evaluation, or decision process for this article. Otherwise, no other potential conflicts of interest relevant to this article were reported. All remaining authors have declared no conflicts of interest.

ORCID iDs

Kiduk Kim https://orcid.org/0000-0002-9659-897X
Gil-Sun Hong https://orcid.org/0000-0002-0068-9413
Namkug Kim https://orcid.org/0000-0002-3438-2217

Funding

This research was supported by grants from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HI21C1148 and HI22C172300).

REFERENCES

1. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Available at. https://proceedings.neurips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html. Published 2014. Accessed August 10, 2024
2. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Available at. https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html. Published 2020. Accessed August 10, 2024
3. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. Available at. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf. Published 2018. Accessed August 10, 2024
4. 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 August 10, 2024
5. Kim K, Cho K, Jang R, Kyung S, Lee S, Ham S, et al. Updated primer on generative artificial intelligence and large language models in medical imaging for medical professionals. *Korean J Radiol* 2024;25:224-242
6. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.1312.6114. Accessed August 10, 2024
7. Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.1409.2329. Accessed August 10, 2024

8. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;9:1735-1780
9. 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 August 10, 2024
10. 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 August 10, 2024
11. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. Available at. https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf. Published 2019. Accessed August 10, 2024
12. 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 August 10, 2024
13. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Available at. https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html. Published 2022. Accessed August 10, 2024
14. Uppal S, Bhagat S, Hazarika D, Majumder N, Poria S, Zimmermann R, et al. Multimodal research in vision and language: a review of current and emerging trends. *Inform Fusion* 2022;77:149-171
15. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, et al. VQA: visual question answering. Available at. https://openaccess.thecvf.com/content_iccv_2015/html/Antol_VQA_Visual_Question_ICCV_2015_paper.html. Published 2015. Accessed August 10, 2024
16. Zellers R, Bisk Y, Farhadi A, Choi Y. From recognition to cognition: visual commonsense reasoning. Available at. https://openaccess.thecvf.com/content_CVPR_2019/html/Zellers_From_Recognition_to_Cognition_Visual_Commonsense_Reasoning_CVPR_2019_paper.html. Published 2019. Accessed August 10, 2024
17. Hossain MZ, Sohel F, Shiratuddin MF, Laga H. A comprehensive survey of deep learning for image captioning. *ACM Comput Surv* 2019;51:1-36
18. Lee H, Ullah U, Lee JS, Jeong B, Choi HC. A brief survey of text driven image generation and maniulation. Available at. https://doi.org/10.1109/ICCE-Asia53811.2021.9641929. Published 2021. Accessed August 10, 2024
19. 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 August 10, 2024
20. Jia C, Yang Y, Xia Y, Chen YT, Parekh Z, Pham H, et al. Scaling up visual and vision-language representation learning with noisy text supervision. Available at. https://proceedings.mlr.press/v139/jia21b.html. Published 2021. Accessed August 10, 2024
21. Nagaraja VK, Morariu VI, Davis LS. *Modeling context between objects for referring expression understanding.* In Leibe B, Matas J, Sebe N, Welling M, eds. *Computer vision–ECCV 2016*. Cham: Springer 2016:792-807
22. Cao M, Li S, Li J, Nie L, Zhang M. Image-text retrieval: a survey on recent research and development. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2203.14713. Accessed August 10, 2024
23. Mazurowski MA, Dong H, Gu H, Yang J, Konz N, Zhang Y. Segment anything model for medical image analysis: an experimental study. *Med Image Anal* 2023;89:102918
24. 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 August 10, 2024
25. Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis. Available at. https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html. Published 2021. Accessed August 10, 2024
26. Metz L, Poole B, Pfau D, Sohl-Dickstein J. Unrolled generative adversarial networks. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.1611.02163. Accessed August 10, 2024
27. Thanh-Tung H, Tran T. Catastrophic forgetting and mode collapse in GANs. Available at. https://doi.org/10.1109/IJCNN48605.2020.9207181. Published 2020. Accessed August 10, 2024
28. Mirza M, Osindero S. Conditional generative adversarial nets. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.1411.1784. Accessed August 10, 2024

29. Nishio M, Muramatsu C, Noguchi S, Nakai H, Fujimoto K, Sakamoto R, et al. Attribute-guided image generation of three-dimensional computed tomography images of lung nodules using a generative adversarial network. *Comput Biol Med* 2020;126:104032
30. Ho J, Salimans T. Classifier-free diffusion guidance. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2207.12598. Accessed August 10, 2024
31. Moon HH, Jeong J, Park JE, Kim N, Choi C, Kim YH, et al. Generative AI in glioma: ensuring diversity in training image phenotypes to improve diagnostic performance for IDH mutation prediction. *Neuro Oncol* 2024;26:1124-1135
32. Xia W, Zhang Y, Yang Y, Xue JH, Zhou B, Yang MH. GAN inversion: a survey. *IEEE Trans Pattern Anal Mach Intell* 2023;45:3121-3138
33. Lee JS, Shin K, Ryu SM, Jegal SG, Lee W, Yoon MA, et al. Screening of adolescent idiopathic scoliosis using generative adversarial network (GAN) inversion method in chest radiographs. *PLoS One* 2023;18:e0285489
34. Pan Z, Gherardi R, Xie X, Huang S. Effective real image editing with accelerated iterative diffusion inversion. Available at. https://openaccess.thecvf.com/content/ICCV2023/html/Pan_Effective_Real_Image_Editing_with_Accelerated_Iterative_Diffusion_Inversion_ICCV_2023_paper.html. Published 2023. Accessed August 10, 2024
35. Preechakul K, Chatthee N, Wizadwongsa S, Suwajanakorn S. Diffusion autoencoders: toward a meaningful and decodable representation. Available at. https://openaccess.thecvf.com/content/CVPR2022/html/Preechakul_Diffusion_Autoencoders_Toward_a_Meaningful_and_Decodable_Representation_CVPR_2022_paper.html. Published 2022. Accessed August 10, 2024
36. Kim G, Kwon T, Ye JC. DiffusionCLIP: text-guided diffusion models for robust image manipulation. Available at. https://openaccess.thecvf.com/content/CVPR2022/html/Kim_DiffusionCLIP_Text-Guided_Diffusion_Models_for_Robust_Image_Manipulation_CVPR_2022_paper.html. Published 2022. Accessed August 10, 2024
37. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. Available at. https://openaccess.thecvf.com/content_cvpr_2017/html/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.html. Published 2017. Accessed August 10, 2024
38. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. Available at. https://openaccess.thecvf.com/content_iccv_2017/html/Zhu_Unpaired_Image-To-Image_Translation_ICCV_2017_paper.html. Published 2017. Accessed August 10, 2024
39. Choi C, Jeong J, Lee S, Lee SM, Kim N. *CT kernel conversion using multi-domain image-to-image translation with generator-guided contrastive learning.* In Greenspan H, Madabhushi A, Mousavi P, Salcudean S, Duncan J, Syeda-Mahmood T, et al. *Medical image computing and computer assisted intervention–MICCAI 2023.* Cham: Springer 2023:344-354
40. Kim J, Park H. Adaptive latent diffusion model for 3D medical image to image translation: multi-modal magnetic resonance imaging study. Available at. https://openaccess.thecvf.com/content/WACV2024/html/Kim_Adaptive_Latent_Diffusion_Model_for_3D_Medical_Image_to_Image_WACV_2024_paper.html. Accessed August 10, 2024
41. Song J, Meng C, Ermon S. Denoising diffusion implicit models. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2010.02502. Accessed August 10, 2024
42. Wang Z, Zheng H, He P, Chen W, Zhou M. Diffusion-GAN: training GANs with diffusion. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2206.02262. Accessed August 10, 2024
43. Yin Y, Huang L, Liu Y, Huang K. DiffGAR: model-agnostic restoration from generative artifacts using image-to-image diffusion models. Available at. https://doi.org/10.1145/3577530.3577539. Published 2023. Accessed August 10, 2024
44. Hong GS, Jang M, Kyung S, Cho K, Jeong J, Lee GY, et al. Overcoming the challenges in the development and implementation of artificial intelligence in radiology: a comprehensive review of solutions beyond supervised learning. *Korean J Radiol* 2023;24:1061-1080
45. Giray L. Prompt engineering with ChatGPT: a guide for academic writers. *Ann Biomed Eng* 2023;51:2629-2633
46. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2106.09685. Accessed August 10, 2024
47. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowl-

JOURNAL of
THE KOREAN SOCIETY of
RADIOLOGY

edge-intensive NLP tasks. Available at. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html. Published 2020. Accessed August 10, 2024
48. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. *Nature* 2023;619:357-362
49. Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic AI. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2401.05654. Accessed August 10, 2024
50. Kim S, Lee CK, Kim SS. Large language models: a guide for radiologists. *Korean J Radiol* 2024;25:126-133
51. 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
52. 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
53. Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. *Vis Comput Ind Biomed Art* 2023;6:9
54. Wu C, Zhang X, Zhang Y, Wang Y, Xie W. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2308.02463. Accessed August 10, 2024
55. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Comput Surv* 2023;55:1-38
56. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: open and efficient foundation language models. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2302.13971. Accessed August 10, 2024
57. 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

의료영상에서 생성형 인공지능과 대형 언어 모델 입문

김기덕[1] · 홍길선[2*] · 김남국[1,2*]

최근 ChatGPT를 포함한 대형 언어 모델의 출현으로 생성형 인공지능은 다양한 분야에 관심을 끌고 있다. 생성형 인공지능은 학습 방법과 데이터에 따라 텍스트, 이미지, 음성 등 다양한 형태의 데이터를 생성할 수 있다. 이에 더해 최근 텍스트와 이미지 등 여러 종류의 데이터를 동시에 처리할 수 있는 기술의 발달로, 다양한 임상정보와 영상정보를 함께 활용해야 하는 의료 환경에서 이러한 멀티모달 생성형 인공지능의 활용 가능성이 높아지고 있다. 본 종설에서는 대형 언어 모델, 이미지 생성 모델, 멀티모달 인공지능에 대한 개념과 종류 등에 대해 알아보고, 연구 사례를 통해 영상의학 분야에서 생성형 인공지능의 활용과 향후 가능성을 알아보고자 한다.

[1]울산대학교 의과대학 서울아산병원 융합의학과,
[2]울산대학교 의과대학 서울아산병원 영상의학연구소 영상의학과