

거대언어모델의 발전과 기초의학교육에 미치는 영향에 대한 고찰

이현수¹

¹부산대학교 의과대학 의학과 생리학교실

Reflections on the Development of Large Language Models and Their Impact on Basic Medical Education

Hyunsu Lee¹

¹Department of Physiology, School of Medicine, Pusan National University

Abstract : Rapid advances in artificial intelligence, particularly in natural language processing, are transforming medical education. Large language models (LLMs), which are capable of processing vast amounts of text and generating coherent, contextualized text, offer an unprecedented opportunity to enhance teaching and learning in medicine. This paper reviews the educational potential of LLMs and explores the potential challenges associated with their integration and use. LLMs can facilitate personalized learning experiences and support research by broadening access to a wide range of medical knowledge and streamlining literature reviews. They can also provide a dynamic and immersive educational experience through interactive learning environments that can mimic clinical settings. However, the integration of LLMs in medical education is not without its concerns. It is important to consider the accuracy and reliability of the information provided, the ethical implications, and the over-reliance on technology, which may affect the development of critical thinking skills. Therefore, the integration of LLMs into medical curricula must be approached with careful consideration, ensuring a harmonious balance between cutting-edge technology and traditional learning methodologies. To achieve this, this paper presents guidelines that emphasize the importance of adopting LLMs based on evidence-based reviews, fostering multidisciplinary collaboration, ensuring transparency, and effectively managing cognitive load. These recommendations aim to facilitate a discourse on the effective integration of LLMs into medical education, ultimately helping to prepare healthcare professionals for the challenges and opportunities of the digital age.

Keywords : Large language models, Medical education technology, Artificial intelligence in healthcare, Ethical implications of AI, Cognitive load

이 과제는 부산대학교 기본연구지원사업(2년)에 의하여 연구되었음.

저자(들)는 '의학논문 출판윤리 가이드라인'을 준수합니다.

저자(들)는 이 연구와 관련하여 이해관계가 없음을 밝힙니다.

Received: June 7, 2024; **Revised:** June 17, 2024;

Accepted: June 18, 2024

Correspondence to: 이현수 (부산대학교 의과대학 의학과 생리학교실)

E-mail: hyunsu.lee@pusan.ac.kr

서론

인공지능이라는 단어 “artificial intelligence”는 1956년 다트머스(Dartmouth) 여름학교에서 존 매카시(John McCarthy), 마빈 민스키(Marvin Minsky) 등의 연구자가 제안한

것으로 알려져 있다[1]. 인공지능은 사람처럼 행동하거나 이성적으로 사고하는 기계라고 흔히 정의한다[2]. 이와 비슷한 의미를 지닌 로봇(robot)은 “인공지능”이라는 단어보다 앞선 시기인 1921년 연극 “R.U.R. Rossum’s Universal Robots”을 통해 처음 사람들에게 선보였다[3]. 체코 출신 극작가 카렐 차페크(Karel Capek)가 노동 또는 노예와 관련이 있는 로봇이라는 체코어를 지능이 있는 인공노동자를 일컫는 단어로 자신의 극에 사용한 것이다. 우리가 현재 사용하고 있는 인공지능이나 로봇이라는 단어는 비록 100년 남짓 되었지만, 인간의 노동을 대신해 줄 수 있는 자동인형이라는 개념 자체는 고대 시대로 거슬러 올라간다[3].

근대 이전의 사람들이 기대했던 대로 근현대 과학기술의 발전과 산업혁명은 기계화를 통해 많은 영역에서 육체노동의 부담을 줄여 주었다. 육체노동뿐만 아니라 지적 능력까지 자동화하는 것에 이르렀고, 이에 컴퓨터과학과 인공지능의 발전이 기여하고 있다. 인간의 지적 능력은 인간이 다른 동물과 스스로를 구분 짓는 중요한 능력 중 하나다. 이를 함양시키는 교육 영역도 자동화의 예외는 아니다. 인터넷을 비롯한 정보 통신 기술의 발전, 빅데이터를 처리할 수 있는 하드웨어 및 소프트웨어 기술의 발전, 그리고 거대언어모델(대형언어모델 혹은 대규모 언어모델 등의 용어가 현재 혼용되어 사용되고 있다. 이하 거대언어모델로 통일한다), 생성모델 등 최신의 인공지능 발전은 인공지능이 교육 영역에서도 교육자의 역할을 보조하는 것을 넘어 축소, 대체할 가능성도 있다는 전망을 하게 한다[4].

인공지능 기술은 개인화된 학습, 자동화를 통한 효율성 증진, 교육에 활용되는 매체의 다양화, 접근성 및 포용성의 향상 등 교육의 다양한 측면에 영향을 미치고 있다. 특히 COVID-19로 인한 팬데믹 기간을 거치며 대부분의 교육 환경에 컴퓨터 혹은 스마트폰이 매개하는 소통과 e-러닝 플랫폼이 도입되어 왔다[5,6]. 피교육자와 교육자를 연결하고, 피교육자의 학습 활동을 추적하며 학습을 지원하는 기술은 교육의 접근성과 유연성을 향상시켰다. 교육 환경에 정보통신 기술의 도입과 진화는 인공지능 기술의 발전이 반영되어 학습 경험을 더욱 향상시키는 방향으로 전환되고 있다.

하지만 데이터를 기반으로 있는 최근 인공지능 기술의 특성을 고려했을 때, 교육 환경에 인공지능을 도입하는 것에 우려되는 측면이 없는 것은 아니다. 피교육자 개인의 데이터 추적에 따른 보안 문제, 개인정보의 윤리적 사용 문제 등이 흔히 제일 먼저 제기된다. 학습 자료 추천, 피교육자 평가 등이 데이터에 기반한 인공지능에 의해 이루어지는 만큼 그것의 정확성, 신뢰성, 투명성, 교육적 효과 등이 보장되어야 한다. 또한 컴퓨터 혹은 스마트폰은 디지털기기를 활용해야 하는 만큼 디지털 격차로 인해 야기될 수 있는 불평등 문제도

있다. 디지털 교육, 인공지능 기술 도입의 가속화로 신중한 도입의 기회를 잃고 있다는 우려가 제기된다.

현대 의학과 의료는 현대 과학기술의 최전선에 있으며, 한해 의·생명 분야에서 출판되는 논문의 양은 기하급수적으로 증가하고 있다[7]. 예비 의료인을 양성하는 의학교육은 빠르게 증가하는 의료와 의학 분야의 지식을 다룰 수 있는 능력을 함양하도록 해야 한다. 여기에 더해 예비 의료인들이 미래의 의학과 의료환경에 미래를 대비할 수 있도록 해야 한다. 교육과정에서 가르쳐야 하는 내용과 그 내용을 다루는 방식으로 의학교육을 나누어 볼 때, 내용과 방식 모두 미래 지향적이어야 할 필요가 있다. 미래의 의료환경에서 가장 영향력이 클 것으로 예상되는 기술은 인공지능의 발전이다[8]. 따라서 의학교육은 의료인공지능의 발전과 영향에 대해서 다뤄야 하며, 가르치는 방식에도 인공지능이 도입되어 예비 의료인들에게 인공지능의 장단점, 한계에 대해 직접 체험하고 생각할 수 있게 할 필요가 있다.

기초의학교육은 건강한 인체의 구조와 기능을 가르치는 해부학, 조직학, 생리학 등이 근간을 이룬다. 수직통합, 수평통합 등의 교육 과정 개편과 의료 현장에서 필요한 실천적 지식과 기술 습득에 대한 강조로 인해 기초의학에 할애되는 시간은 상대적으로 감소하고 있는 경향이 있다[9]. 그에 반해 기초 의학을 구성하고 있는 학문 영역은 지속해서 발전하고 복잡해지면서 지식의 양은 증가하고 있다. 상반된 두 추세로 인해 예비 의료인들은 정확하고 상세한 기초의학 지식을 짧은 시간 안에 대량으로 습득해야 하는 상황에 직면해 있다. 이러한 교육 현장에 인공지능 기술의 도입은 폭발하는 의학 지식의 양과 학업 부담으로부터 구원자가 될 수도 있다. 하지만 그 자체의 한계와 문제점에 대한 고민이 필수적으로 수반되어야 한다.

본 종설에서는 인공지능 기술의 의학교육 도입 및 통합에 대해 논한다. 인공지능 기술 중에서는 최근 주목받고 있는 거대언어모델 중심으로 살펴본다. 자연어 처리 기술의 발전 흐름을 살펴보고, 거대언어모델의 원리에 대해 간략하게 개관한다. 거대언어모델이 가지고 있는 잠재적 장점과 유의해야 할 점을 살펴보고, 학습이라는 관점에서 읽기와 바람직한 어려움 개념을 통해 거대언어모델을 분석하고 논의한다. 이를 바탕으로 향후 언어모델이 의학교육 과정에 접목되어 발전해 나가야 할 방향에 대해 고찰하고자 한다.

거대언어모델의 역사적 진화

인공지능 기술의 초창기에는 규칙 기반 시스템으로 인공지능을 만들려 했다. 자연어 처리 영역도 예외는 아니었다.

1960년대 중반에 개발되어 선보인 ELIZA와 같은 시스템은 사용자가 입력한 문장의 키워드를 중심으로 미리 정의된 패턴에 따라 응답을 함으로써 대화를 진행해 나갔다[10]. 프로그래밍된 규칙에 의해서만 작동된다는 점에서 한계가 분명했다. 새로운 패턴에 대응을 못했고, 대화가 동어반복적으로 진행되었다. 입력할 수 있는 문장의 길이와 출력되는 문장의 길이 역시 한 두 문장 정도 짧아 언어모델이라고 불리기에 부족한 점이 많았다. 그럼에도 불구하고 자연어를 컴퓨터가 처리할 수 있다는 가능성을 보여주었다는 점에서 의미가 있는 결과였다.

1980년대 후반 이후 통계 기반 확률적 인공지능 모델의 발전은 자연어 처리를 거대 텍스트 데이터의 관점에서 접근하도록 변화시켰다. 인터넷의 발전으로 인해 텍스트 데이터의 양이 대폭 증가하였고, 은닉 마르코프 모델 같은 확률 모델이 등장하면서 텍스트 내 단어의 흐름을 확률적 네트워크로 파악하고 처리할 수 있게 되었다. 이런 모델들은 단어 사이의 연결에서 각 단어나 글자가 등장할 확률이 서로 독립적이지 않다는 점을 이용한다. 즉, 먼저 등장한 문장, 단어, 글자의 연쇄가 다음에 등장할 단어나 글자의 확률을 결정한다. 이러한 확률적 관계를 학습함으로써, 텍스트 데이터의 양이 늘어날수록 모델의 정확도는 더욱 향상되었다. 이와 같은 통계적 접근법은 현재 사용되고 있는 거대언어모델에도 적용되고 있다.

2000년대 이후 심층신경망, 합성곱신경망, 순환신경망 등의 발전과 역전파알고리즘을 비롯한 최적화 학습 기법의 도입으로 크고 복잡한 데이터를 효율적으로 학습시킬 수 있게 되었다. 자연어 처리 영역에서 획기적인 발전은 트랜스포머 모델의 등장이다[11]. 주의(attention)라는 개념을 인코더(encoder, 입력 데이터를 처리 가능한 새로운 형태의 표현으로 변환하는 장치 혹은 함수)와 디코더(decoder, 인코더에서 생성된 데이터를 받아 원래 형식으로 복원하거나, 새로운 형태의 출력 데이터로 변환하는 장치 혹은 함수)에 도입해 텍스트 사이의 관계와 맥락을 인공지능이 스스로 학습할 수 있게 했고, 학습된 주의벡터를 기반으로 일관된 맥락을 가진 긴 텍스트를 생성하는 모델을 만드는 데 크게 기여했다. BERT (Bidirectional Encoder Representations from Transformers) 및 GPT (Generative Pre-trained Transformer) 등이 잘 알려진 트랜스포머 기반 거대언어모델의 시초라고 볼 수 있다[12,13].

거대언어모델의 의학교육에서의 역할

거대언어모델이 교육 영역, 더 좁게는 의학교육 영역에 미

Table 1. Opportunities and challenges of large language model on medical education

Opportunities	Challenges
Personalized learning	Accuracy and reliability
Access to vast medical knowledge	Ethical considerations
Support for research	Over-reliance on technology
Interactive learning environments	Integration into curriculums

칠 수 있는 영향은 다양하다. 학습자와 교수자 입장에서 더 효율적인 교육이 이루어질 기회가 거대언어모델을 통해 열린다고 할 수 있다(Table 1).

거대언어모델로 인해 교육 영역에 발생하고 있는 기회를 순서대로 설명하면 다음과 같다. 첫 번째, 개인화된 학습이 가능해진다. 거대언어모델에 대화 형식으로 질문과 응답을 이어 나가면 언어모델은 입력된 문장(프롬프트)이 요구하는 수준에 맞게 출력을 내어 준다. 예를 들어, 초등학교 수준, 대학생 수준, 박사 수준으로 채장의 기능과 구조에 대해 설명 하라고 입력할 경우, 단순히 탄수화물, 단백질, 지방을 분해 한다는 수준의 쉬운 설명에서부터 알파-아밀라아제(α -Amylase)의 자세한 기능에 대해 설명하는 수준까지 사용자가 원하는 수준에 맞춰 출력이 나온다(Table 2).

이러한 기능은 곧 거대언어모델의 두 번째 기회인 급속도로 증가하고 있는 의학지식을 쉽고 빠르게 습득하도록 도울 수 있다는 점으로 연결된다. 현재 많은 거대언어모델들이 플러그인 기능을 이용해 문헌 검색 기능을 제공한다. 이 경우 검색된 문헌과 그에 대한 요약물 기반으로 거대언어모델이 출력을 생성하기 때문에, 흔히 문제로 제기되는 환각현상(hallucinations)을 방지할 수 있고, 출력된 정보의 신뢰성을 검증하기 용이하다. 특정 주제에 대해 연구를 하려고 할 때, 문헌 검색 플러그인을 활용하면 관련된 논문을 검색해 주는 것을 넘어, 요약 및 정리까지 거대언어모델이 처리하여 출력을 제공하므로 사전 조사 및 논문 작성을 거대언어모델이 지원할 수 있다. 이것이 세 번째 기회에 해당한다.

현재 서비스되고 있는 거대언어모델은 대부분 채팅형으로 제공된다. 따라서 자연스럽게 상호작용을 기반한 몰입형 교육환경이 형성되며, 이것이 네 번째 기회다. 텍스트를 키보드로 입력하고 텍스트 형태로 출력하는 것을 뛰어넘어, 음성 입력과 출력을 제공하고 있는 모델들은 실제 대화하는 환경을 재현할 수 있다(2024년 5월 13일 OpenAI에서는 ChatGPT의 새로운 버전인 ChatGPT4o를 발표했다. 발표 영상에서는 ChatGPT가 영상 입력과 음성 입력을 동시에 처리하는 것을 보여주며, 반응 지연시간도 상당히 줄여 대화가 자연스럽게 이어지는 것을 볼 수 있다). 비록 기초의학교육 영역보다는

Table 2. Example of pancreatic function and enzyme production explained at various academic levels

Prompt: What is the role of the pancreas in the digestive system and what enzymes do it produce to aid in digestion? Explain at three levels: elementary school level, undergraduate level, and postdoctoral level.

Elementary school level	Undergraduate level	Postdoctoral level
(Preceding omitted) It makes three main types of helpers: One that helps break down sugars (like in bread or fruit), Another that helps break down proteins (like in meat or beans), and One more that helps break down fats (like in butter or cheese). (Later omitted)	(Preceding omitted) The key enzymes include: Amylase: This enzyme helps in the breakdown of carbohydrates into simple sugars. For instance, it converts starches from foods like pasta and bread into glucose that our body can use for energy. Proteases (such as trypsin and chymotrypsin): These are responsible for breaking down proteins into amino acids, which are the building blocks our body uses to make new proteins for muscle growth, repair, and other functions. Lipases: These enzymes tackle fats, breaking them down into fatty acids and glycerol, which our body can then absorb and use for energy and other needs. (Later omitted)	(Preceding omitted) These enzymatic secretions include: Pancreatic α -amylase: Catalyzes the hydrolysis of α -1,4-glycosidic linkages in polysaccharides to yield disaccharides, oligosaccharides, and limit dextrins, facilitating carbohydrate digestion. Serine proteases (e.g., trypsinogen activated to trypsin, and chymotrypsinogen to chymotrypsin): These zymogens, once activated in the intestinal lumen, play key roles in proteolysis, breaking down polypeptides into absorbable amino acid units. Their activation cascade is tightly regulated, beginning with the duodenal enzyme enterokinase converting trypsinogen to trypsin, which then activates other proenzymes. Pancreatic lipase: This is the primary enzyme responsible for the hydrolysis of dietary triglycerides into monoglycerides and free fatty acids, a critical step for lipid assimilation. (Later omitted)

입상교육에 가까운 내용이지만, 거대언어모델에서 특정 질문을 가진 환자 역할을 하라고 프롬프트를 줄 수 있기 때문에, 진단 과정을 시뮬레이션할 수도 있다(Table 3).

거대언어모델 도입에 대한 우려

1. 의학교육 도입의 잠재적 단점

앞서 거대언어모델이 플러그인 기능 등을 활용해 이전에 비해 정확한 정보를 제공하는 것에 대해 언급하였다. 하지만 여전히 부정확한 정보를 제공할 위험은 존재한다[14]. 거대언어모델이 학습한 자료와 정보가 모두 정확하다고 보기 힘들기 때문이다. 특히 거대언어모델의 특성상 어떤 프롬프트를 사용하는가에 따라 출력이 달라지므로, 사용자에게 따라 정보의 정확도가 달라질 위험이 있다. 또한 배경지식의 수준이 낮은 사용자일수록 프롬프트를 정확하게 입력하지 못할 가능성이 높고, 거대언어모델이 출력한 정보의 진위를 판단하기 어려워할 가능성이 높다는 점도 교육 현장에서 사용될 때 고려되어야 할 점이다.

거대언어모델이 출력하는 정보의 정확성 기준이 무엇이나는 문제도 제기될 수 있다. 텍스트 데이터를 기반으로 학습하는 거대언어모델의 특성상 인터넷, 책 등에서 방대한 양의 텍스트를 학습 데이터로 사용하게 되는데, 양이 많은 만큼 텍스트의 질이 담보되지 못한다. 또한 인터넷 정보의 90% 이상이 영어로 되어 있기 때문에 영어권이 아닌 문화권의 관점, 교육 콘텐츠 등의 양이 절대적으로 부족하여 편향되거나 부정확한 결과가 나올 가능성이 높다. 예를 들어 ChatGPT에게 고구려 멸망에 관해 설명해보라고 했을 경우, 7세기에 멸망할 때 후백제와의 갈등이 있었다고 하는 등 상당한 오류를 보인다(2024년 4월 3일, ChatGPT 3.5를 사용한 결과에 대해 기술한 것이다. 유료로 사용할 수 있는 ChatGPT 4.0은 668년 나당연합군에 의해 멸망했음을 정확하게 언급한다. 무료인 네이버 클로바와 구글 Gemini 역시 668년 나당연합군에 의한 멸망했음을 정확하게 언급한다. 하지만, 구글 Gemini는 668년에 연개소문이 사망했다고 언급하는 오류를 보였다. 네이버 클로바의 경우 연개소문 사망 시기에 대해 언급하지 않고 연개소문 사망 후 아들들의 권력다툼이 국력을 약화시킨 원인이라고 ChatGPT 4.0이나 구글 Gemini에 비해 더 자세

Table 3. Example of simulation of clinical interaction: Role-playing patient scenarios

Prompt	Response
Act like a patient with symptoms of urinary stones. However, don't mention urinary stones directly. I'm a urologist.	<p>Hello Doctor, for the past few weeks, I've been experiencing some really uncomfortable symptoms, and I'm not sure what's causing them. I've had sharp, cramping pains in my lower back and side, sometimes it even moves down towards my groin area. The pain comes and goes but when it hits, it's almost unbearable.</p> <p>Additionally, I've noticed some changes when I go to the bathroom. My urine isn't clear like it usually is; it's been quite cloudy, and a couple of times, I've even noticed a pinkish color, which I assume might be blood. It's also been painful to urinate, and I feel like I have to go a lot more frequently than normal, but each time, it's just a little amount.</p> <p>I've been drinking water thinking it might help, but it doesn't seem to make much of a difference with the pain or the frequent urge to urinate. I haven't had any recent injuries or done anything different in my routine that I think could have led to this. Can you help me figure out what's going on?</p>

하게 지적하고 있다).

거대언어모델이 학습한 데이터의 질을 사용자가 파악하기 어렵다는 점은 교육 현장에서 특히 유의해야 할 점 중 하나다. 인터넷 자료 중 많은 부분이 신빙성이 높지 않기 때문이다. 신뢰성 문제는 기초의학교육보다 임상의학교육 분야에서 더 치명적이다. 의학 분야는 빠르게 지식이 증가하고 개선되고 있어, 최신 진료 지침의 변경 여부를 모르는 상태에서 거대언어모델이 제공하는 정보가 과거의 지침을 기준으로 작성되었을 경우, 사용자가 잘못된 의사결정을 내릴 위험을 배제할 수 없다. 조혈모세포 이식과 관련한 연구를 보면 복잡한 임상 사례에 대한 의사결정일수록 거대언어모델은 수련 받는 전공의보다 잘못된 판단을 내릴 가능성이 높다[15]. 최신 정보를 제공한다 하더라도, 논쟁이 많은 문제에 대해 거대언어모델이 논쟁의 지점을 충분히 다룬다는 보장이 없다. 만성 통증 관리를 위한 대마 사용, 영양제 사용의 효과, 보완대체의학의 효능 등 전문가들 사이에서도 의견이 일치하지 않는 주제들이 많다. 사용자가 충분히 주의를 기울이지 않은 상태에서 프롬프트를 입력한다면 편향된 결과를 거대언어모델이 제공하게 될 가능성이 높다.

2. 읽기 관점에서 본 거대언어모델

기술과 경제 수준이 향상됨에 따라 많은 교실과 강의실의 칠판과 화이트보드가 스크린이나 전자칠판으로 바뀌었다. 전달 매체가 변화하면서 각종 인터넷과 스마트기기를 활용한 정보통신기술도 교실과 강의실에 도입되고 있다. 이렇게 디지털화된 매체 기반 교육과 그것을 뒷받침하는 기술을 묶어 거칠게 에듀테크라고 부를 수 있을 것이다. 인공지능을 비롯한 새로운 기술들을 서둘러 교육 현장에 도입하는 배경에는

기술이 주는 효율성과 새로운 기술이 더 나은 학습 효과를 낼 것이라는 기대가 은연중에 깔린 듯하다. 학습 효과에 대한 엄밀한 검증이 뒷받침되지 않더라도, 새로운 기술들을 교육 현장에 도입하여 가르치지 않으면 미래 사회에 학생들이 적응할 수 있는 기회를 놓칠지도 모른다는 우려도 적지 않게 작용한다. 거대언어모델의 교육 현장 도입 역시 알고리즘 기반의 디지털 교육 적용의 연장선상으로 이해할 수 있다.

전달 매체가 전자기기로 변화하는 것은 무거운 종이책이 가벼운 스마트기기에 들어간다는 것 이상의 의미가 있다. 읽기라는 행위는 학습 과정에 중요한 비중을 차지하고 있다. 읽음으로써 명시적 기억(사실)을 획득할 때, 읽는 과정이 인코딩되는 절차적 기억(맥락과 과정) 역시 중요하다. 기억은 부호화-저장-인출의 세 단계로 간략하게 나뉘볼 수 있는데, 맥락은 모든 단계에 관여한다[16,17]. 종이책의 글자, 문장, 도식 등의 위치는 물리적으로 고정되어 있다. 우리는 종이책으로 읽은 사실에 대해 종이책 중간 어디쯤 펼쳐진 면의 가운데라고 맥락과 함께 기억을 떠올린다[18]. 그러나 디지털기기의 콘텐츠의 맥락은 고정되어 있지 않아, 디지털기기를 활용한 읽기에는 맥락이 개입할 여지가 상대적으로 적다. 끝없이 이어지는 스크롤 위에 글자와 그림들은 부유하며, 우리 기억의 맥락도 마찬가지로 부유하게 된다[19,20].

전체적인 맥락을 이해하는 깊은 사고가 필요한 읽기에서는 종이책이 분명 강점이 있지만, 그렇다고 해서 디지털기기를 통한 읽기나 온라인 읽기를 배척해야만 하는 것은 아니다. 디지털 읽기는 빠르게 검색하고 내용을 훑어볼 수 있게 하기에 여러 사실들과 주장을 동시에 검토하고 대조하기 용이하다. 종이책이 깊이 읽기를 통해 비판적 읽기를 유도할 수 있다면, 디지털기기는 동시적 읽기(syntopical reading)를 통해

비판적 읽기를 유도한다[21].

거대언어모델의 경우, 사용자의 능력에 따라 다른 읽기 환경을 제공한다. 사용자가 모호하고 광범위한 프롬프트를 입력할 경우 맥락이 제거되고, 맞지도 틀리지도 않는 일반적인 출력을 내놓는 경향이 있다. 사용자가 구체적이고 엄밀한 사항을 요구해야만 거대언어모델은 자세하고 깊이 있는 내용을 출력으로 내어놓는데, 이때 거대언어모델의 장점을 살리기 위해서는 사용자가 이 출력에 대해 집요하게 비판적인 질문과 의견을 제기해야 한다. 거대언어모델과의 대화 과정이 일종의 맥락으로 작용하게 되며, 거대언어모델이 접근 가능한 정보의 질과 양에 따라 높은 수준의 동시적 읽기 환경을 사용자에게 제공할 수 있다.

거대언어모델이 줄 수 있는 교육 성과를 평가할 때는 피교육자가 거대언어모델의 출력을 읽어낸다는 점에서 다른 도구(종이책, 온라인 읽기)와 비교할 필요가 있다. 종이책은 깊이 읽기에 효율적이지만, 서로 다른 자료를 비교하는 것은 어렵다. 그러나 온라인 읽기나 거대언어모델은 긴 맥락을 바탕으로 한 깊이 읽기를 유도하지는 않지만, 서로 상충되는 많은 자료를 동시에 읽고 검토하기 유리하다. 어떤 목적으로 학습하는가에 따라 적절한 도구를 활용하는 것이 좋을 것이다. 기초의학에서 요구하는 필수 학습 성과라면 깊이 읽기를 통해 기본적인 이해를 하도록 유도하는 것이 좋을 것이며, 심화에 해당하는 최신 연구 결과 혹은 논쟁의 여지가 있는 영역은 온라인 읽기나 거대언어모델을 활용해 많은 자료를 비판적으로 검토하는 것이 도움 될 것이다.

3. 바람직한 어려움 관점에서 본 거대언어모델을 이용한 학습

바람직한 어려움(Desirable Difficulty)이란 피교육자에게 적절하게 어려운 방식으로 학습 내용을 전달함으로써 학습 성과를 더 좋게 한다는 개념을 의미한다[22]. 연구에 따르면, 강의 평가가 좋은 강의를 들은 졸업생의 성과가 반드시 강의 평가 점수와 비례하여 향상되는 것은 아니다[23-25]. 강의 평가가 좋게 나오는 경우는 종종 교수자가 학생들의 편의를 위해 쉽게 강의하는 방식을 도입했을 때이다. 하지만 이런 강의를 들은 학생들이 장기적으로 해당 분야에 대한 이해도가 오히려 낮다는 결과를 보인다. 반면에, 강의가 어렵다는 이유로 강의 평가가 낮게 나온 강의를 들은 학생들과 비교할 경우, 쉬운 강의를 들은 학생들보다 어려운 강의를 들은 학생들이 장기적으로 해당 분야에 대한 이해도가 더 높다. 학생 입장에서 때로는 모순적이고, 쉽사리 이해되지 않는 내용을 적극적으로 소화하려는 과정을 통해 학습한 내용을 장기적으로 가져갈 수 있다는 것이다. 쉬운 강의와 어려운 강의 효과

를 메타인지 측면에서 설명할 수도 있다[26]. 쉬운 강의를 들은 경우, 피교육자는 해당 분야를 잘 알고 있다고 착각하게 된다. 그러나 그 착각으로 인해 실제 테스트에서 더 낮은 성과를 보인다.

손쉽게 정리된 내용을 프롬프트 입력으로 얻을 수 있는 거대언어모델 사용은 강의 평가가 좋은 쉬운 강의와 유사하게 작용할 가능성이 있다. 바람직한 어려움이 거대언어모델 사용으로 제거된다는 것이다. 바람직한 어려움을 통해 피교육자는 학습하는 내용에 대해 능동적으로 개입하고, 몰입하여 내용을 습득 내지는 체득할 수 있지만, 이러한 과정을 디지털 교육과 거대언어모델이 앗아갈 가능성이 있다.

바람직한 어려움은 장기적 성과를 위해 교육 현장에서 권장되어야 하지만, 학생에게 어렵게 가르치는 것이 꼭 좋은 성과를 내는 것은 아니다. 바람직한 어려움이란 학생이 받아들일 수 있고, 학생 스스로 개입할 여지가 있을 때 좋은 효과를 낸다. 짧은 시간에 많은 내용을 전달해야 하는 의학교육의 특성상 피교육자는 이미 인지 과부하 문제를 겪을 가능성이 높다[27]. 전달해야 하는 내용에 더해 외적인 부분에서 피교육자가 인지 자원을 써야 한다면 문제는 악화된다. 예시로 비슷한 내용을 서로 다른 교수님이 가르칠 경우 학생들은 중요하지 않은 세부적인 내용의 불일치로 혼란을 겪을 수 있다. 지나치게 많은 참고 자료, 특히 거대언어모델이 생성하는 내용도 인지 과부하의 원인이 될 수 있다. 이런 문제는 “바람직하지 않은 어려움(Undesirable difficulty)”이라고 부른다면, 거대언어모델의 활용이 바람직하지 않은 어려움을 최소화하고 바람직한 어려움이 적절히 노출되도록 유도하는 것이 중요하다.

바람직하지 않은 어려움을 교육 시스템 차원, 커리큘럼에서 제거하는 것은 교육자와 교육행평가의 몫이 될 것이다. 피교육자 입장에서 바람직하지 않은 어려움을 제거할 수 있는 좋은 방법은 연습이다. 고리타분한 이야기지만 해부학 등 기초의학 강의를 시작하기 전 방학을 활용해 “골학 오리엔테이션” 등으로 기본적인 의학용어를 미리 익히는 의대생이 많은 것을 보면 학생들도 연습의 중요성을 잘 알고 있다. 연습에 거대언어모델을 활용하도록 하는 것은 학생들에게 맞춤형 학습 자료를 접할 수 있는 기회를 제공하고, 강의 시간에 겪을 인지 과부하를 줄일 수 있다는 점에서 권장할 만하다.

거대언어모델의 접목과 방향 제시

정보통신기술 발달로 인해 인터넷이 보급되고, 위키피디아 등이 등장하는 등 정보에 대한 접근성이 좋아져 왔다. 높아진 정보접근성이 교육에 미치는 영향에 대한 우려와 논쟁은 반

복되어 왔다. 현재 거대언어모델에 대한 논쟁도 유사한 양상을 띤다. 위키피디아가 널리 사용되기 시작되던 시점에도 교육 현장에서 위키피디아에 대한 신뢰성 문제를 제기하였고, 학생들이 위키피디아에 지나치게 의존하는 것이 교육의 효과를 떨어뜨린다는 우려가 있었다[28]. 거대언어모델을 통해 학습하는 것에도 비슷한 우려가 제기된다. 위키피디아의 정보를 복사 붙여넣기하는 것을 넘어, 거대언어모델은 새로운 내용을 그럴듯하게 써준다는 점에서 작성자가 지식을 습득하여 과제를 작성한 것인지 아니면 단순히 거대언어모델의 출력을 그대로 사용한 것인지 평가자는 알 수 없다. 피교육자가 지나치게 거대언어모델에 의존하면 독립적이고 비판적인 사고 능력을 함양할 수 있는 기회를 잃게 될 우려가 있다. 따라서 거대언어모델을 비롯한 인공지능 기술들을 교육과정에 통합할 때, 피교육자의 학습 성과를 해치지 않는 방식에 대해 면밀히 검토하는 것이 필요하다.

임상 현장에 거대언어모델 도입에 대한 논의를 기반으로 교육 현장에 거대언어모델을 도입할 때 유의해야 할 점을 나열하면 다음과 같다[29]. 첫째, 근거를 기반해 교육 현장에 도입 여부와 방식을 결정해야 한다. 둘째, 학생의 참여가 증가하는 것만 따지는 것이 아니라, 교육 성과지표가 개선되는지 조사해야 한다. 기초의학교육에서 학생들이 처음 접하는 의학용어와 개념들이 거대언어모델을 통해 얼마나 효과적으로 전달되는지 평가해야 한다. 의학용어의 정확성, 학습 내용의 이해도, 적용 능력 등 교육 성과 지표를 중심으로 도입 근거가 될만한 추가 연구가 필요하다. 셋째, 학제 간 협업을 기반으로 거대언어모델의 개발과 도입이 이루어져야 한다. 거대언어모델의 안전하고 효율적인 도입을 위해서는 기초의 학교육자뿐만 아니라 임상 의사, 인지과학자, 인공지능 개발자 등 다양한 분야의 전문가들의 협력이 필요하다. 넷째, 거대언어모델 사용 지침과 사용 여부에 대해 투명하고 신뢰할 수 있는 방식으로 교육자와 학생들 간의 소통이 이루어져야 한다. 다섯째, 거대언어모델이 생성하는 교육 콘텐츠에 차별적 요소가 없는지 면밀히 검토하고 제거하거나 최소화하도록 해야 한다. 의료 현장에서는 장애인, 성소수자, 선천성 질환, 유전성 질환, 난치성 질환 등 다양한 소수자 그룹을 만나게 된다. 이러한 소수자 그룹에 대한 사회적 낙인을 방지하고, 포괄적이며 공정한 교육이 이루어질 수 있도록 주의를 기울여야 한다. 이에 더해 거대언어모델 도입 시 고려해야 할 윤리적인 요소로 학생들이 의료인으로서 평생 가지게 될 학습과 의술에 대한 태도도 고려해야 한다. 여섯째, 거대언어모델 도입으로 인해 인지 부하가 증가하는 것이 아니라, 감소하는 방향으로 교육과정이 설계되어야 한다.

거대언어모델과 같은 인공지능 기술을 기초의학에 도입하

는 것은 잠재적인 이점도 있지만, 우려되는 점이 없지 않다. 이러한 기술이 제공하는 편리함과 새로운 학습 기회는 분명 큰 장점이지만, 학생들의 독립적이고 비판적인 사고 능력을 저해할 위험성도 존재한다. 거대언어모델이 지닌 편향성, 환각현상, 모델 학습에 쓰인 훈련 데이터에 대한 낮은 신뢰도, 프롬프트에 의존적인 답변 등의 단점을 고려할 때, 사용자가 비판적이고 회의적인 태도로 거대언어모델을 활용할 수 있는 능력을 함양시켜 주는 것이 필요하다. 따라서, 교육과정에 거대언어모델을 통합할 때는 신중하고 체계적인 접근이 필요하다. 다양한 전문가들의 협력을 통해 거대언어모델의 안전성과 효율성을 확보하고, 교육자와 학생 간의 투명한 소통을 통해 신뢰를 구축해야 한다. 또한, 학생들이 이러한 기술을 올바르게 활용하도록 지도하고, 차별적 요소를 최소화하는 등 윤리적 고려를 포함한 종합적인 전략이 필요하다. 이렇게 한다면, 거대언어모델이 교육의 보조 도구로서 유용하게 사용되면서도, 학생들의 학습 성과와 비판적 사고 능력이 훼손되지 않을 것이다.

REFERENCES

1. Buchanan BG. A (very) brief history of artificial intelligence. *AI Magazine*. 2005;26:53-53.
2. Russell SJ, Norvig P, Davis E. *Artificial Intelligence: A Modern Approach*. Hoboken: Prentice Hall; 2010.
3. Nocks L. *The Robot: The Life Story of a Technology*. Baltimore: Johns Hopkins University Press; 2008.
4. Lee H. *The Rise of ChatGPT: Exploring its Potential in Medical Education*. *Anat Sci Educ*. 2023.
5. Li C, Lalani F. The COVID-19 pandemic has changed education forever. This is how. *World Economic Forum*. 2020. Available from: <https://www.weforum.org/agenda/2020/04/coronavirus-education-global-covid19-online-digital-learning/2024-05-20>.
6. Pokhrel S, Chhetri R. A Literature Review on Impact of COVID-19 Pandemic on Teaching and Learning. *High Educ Future*. 2021;8:133-41.
7. Kamtchum-Tatuene J, Zafack JG. Keeping Up With the Medical Literature: Why, How, and When? *Stroke*. 2021;52:e746-8.
8. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25:44-56.
9. Kim NH. Reduction in Basic Medical Education and Practical Training for Aspiring Physician-Scientists. 2022. Available from: <http://www.monews.co.kr/news/articleView.html?idxno=313172>.

10. Shah H, Warwick K, Vallverdú J, Wu D. Can machines talk? Comparison of Eliza with modern dialogue systems. *Comput Hum Behav.* 2016;58:278-95.
11. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *NIPS.* 2017;10: S0140525X16001837.
12. Radford A, Narasimhan K. Improving Language Understanding by Generative Pre-Training. 2018. Available from: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
13. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Stroudsburg: North American Chapter of the Association for Computational Linguistics;* 2019:2234.
14. Duong DV, Solomon B. Analysis of large-language model versus human performance for genetics questions. *Eur J Hum Genet.* 2024;32:466-8.
15. Civettini I, Zappaterra A, Ramazzotti D, Granelli BM, Rindone G, Aroldi A, et al. Evaluating the Performance of Large Language Models in Hematopoietic Stem Cell Transplantation Decision Making. *Blood.* 2023;142:3726-3726.
16. Heald JB, Wolpert DM, Lengyel M. The Computational and Neural Bases of Context-Dependent Learning. *Annu Rev Neurosci.* 2023;46:233-58.
17. Heald JB, Lengyel M, Wolpert DM. Contextual inference in learning and memory. *Trends Cogn Sci.* 2023;27:43-64.
18. Fischer MH. Perceiving Spatial Attributes of Print. Reading as a Perceptual Process. Amsterdam: Elsevier; 2000. p. 89-117.
19. Reich SM, Yau JC, Xu Y, Muskat T, Uvalle J, Cannata D. Digital or Print? A Comparison of Preschoolers' Comprehension, Vocabulary, and Engagement From a Print Book and an e-Book. *AERA Open.* 2019;5:233285841987838.
20. Haddock G, Foad C, Saul V, Brown W, Thompson R. The medium can influence the message: Print-based versus digital reading influences how people process different types of written information. *Br J Psychol.* 2020;111:443-59.
21. Chen JY, Lee WH. The Breadth and Depth of E-reading and Paper-reading. Annual Meeting of the Cognitive Science Society. 2015.
22. Bjork RA, Bjork EL. Desirable difficulties in theory and practice. *J Appl Res Mem Cogn.* 2020;9:475.
23. Ruslim NM, Ee NL, Saharun N, Baharuddin N, Bakar NAA, Karim MKA. The Correlation between Teaching Evaluation and Lecturers' Performances. *Asian Social Science and Humanities Research Journal (ASHREJ).* 2020;2:32-7.
24. Carrell SE, West JE. Does professor quality matter? Evidence from random assignment of students to professors. *J. Pol Econ.* 2010;118:409-32.
25. Braga M, Paccagnella M, Pellizzari M. Evaluating students' evaluations of professors. *Econ Educ Rev.* 2014;41:71-88.
26. Persky AM, Lee E, Schlesselman LS. Perception of learning versus performance as outcome measures of educational research. *Am J Pharm Educ.* 2020;84:ajpe7782.
27. Qiao YQ, Shen J, Liang X, Ding S, Chen FY, Shao L, et al. Using cognitive theory to facilitate medical education. *BMC Med Educ.* 2014;14:79.
28. Kulesza J, DeHondt G, Nezelek G. More technology, less learning? *Information Systems Education Journal.* 2011;9:4.
29. Stade EC, Stirman SW, Ungar LH, Boland CL, Schwartz HA, Yaden DB, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *Npj Ment Health Res.* 2024;3:12.

간추림 : 인공지능, 특히 자연어 처리 분야의 급속한 발전으로 인해 의학교육도 변화하고 있다. 거대언어모델은 방대한 양의 텍스트를 처리하고 일관된 문맥을 가진 텍스트를 생성할 수 있는 기능을 통해 의학 분야의 교육과 학습을 향상시킬 수 있는 유례 없는 기회를 제공한다. 이 논문에서는 거대언어모델의 교육적 잠재력을 검토하고, 통합과 사용에 관련된 잠재적 문제점을 살펴본다. 거대언어모델은 개인화된 학습 경험을 제공하고, 광범위한 의학 지식에 대한 접근성을 넓히며, 문헌 검토를 간소화하여 연구를 지원할 수 있다. 또한 임상 환경을 모사할 수 있는 대화형 학습 환경을 통해 역동적이고 몰입감 있는 교육 경험을 제공할 수 있다. 그럼에도 불구하고 의학교육에 거대언어모델을 통합하는 데 우려되는 점이 없는 것은 아니다. 제공되는 정보의 정확성과 신뢰성, 윤리적 영향, 비판적 사고력을 저해할 수 있는 기술에 대한 과도한 의존성에 대한 신중한 고려가 필요하다. 따라서 의학교육 과정에 거대언어모델을 통합할 때 신중하게 접근할 필요가 있으며, 기술과 전통적인 학습 방법론 간의 균형이 필요하다. 이를 위해 본 논문은 근거 기반 거대언어모델 도입, 다학제 협업, 투명성, 인지 부하 개념 등을 강조하는 지침을 제시한다. 본 논문이 거대언어모델을 의학교육에 효과적으로 통합하기 위한 논의를 촉발하여 디지털 시대의 의료 전문가를 양성하는 데 도움이 될 것으로 기대한다.

찾아보기 낱말 : 거대언어모델, 기초의학교육, 의료 분야의 인공지능, 인지 부하, 인공지능의 윤리