거대언어모델을 활용한 한의학 교육 강화 : 교육용 인공지능 개발을 중심으로

박사윤[1,2]*·김창업[1,3]*

1 : 가천대학교 한의과대학, 2 : 서울대학교병원 의생명연구원, 3 : Stanford University School of Medicine

Enhancing Korean Medicine Education with Large Language Models: Focusing on the Development of Educational Artificial Intelligence

Sa-Yoon Park[1,2]*, Chang-Eop Kim[1,3]*

1 : Department of Physiology, College of Korean Medicine, Gachon University,
2 : Biomedical Research Institute, Seoul National University Hospital,
3 : Department of Neurobiology, Stanford University School of Medicine,

Large language models (LLMs) have introduced groundbreaking innovations in various fields, including healthcare, where they augment medical diagnosis, decision-making, and facilitate patient-doctor communication through their exceptional contextual understanding and inferential abilities. In the realm of Korean medicine (KM), the utilization of LLMs is highly anticipated. However, it demands additional training with domain-specific KM data for seamless integration of KM knowledge. There are two predominant strategies for training domain-specific LLMs in the KM domain. The first approach entails direct manipulation of the LLM's internals by either pretraining a base model on an extensive corpus of KM data or fine-tuning a pretrained model's parameters using KM-related question-answering datasets. The second approach avoids internal model manipulation and leverages techniques like prompt engineering, retrieval augmented generation, and cognitive augmentation. Domain-specific LLMs specialized for KM hold the potential for diverse applications, ranging from personalized medical education plans and content generation to knowledge integration, curriculum development, automated student assessment, virtual patient simulations, and advanced research and scholarly activities. These advancements are poised to significantly impact the field of KM and medical education at large.

keywords : Large language model, Medical education, Korean medicine, Domain-specific LLM

서 론

ChatGPT/GPT-4를 위시한 거대언어모델(Large language model, LLM) 기반의 생성형 AI(generative artificial intelligence) 기술이 거의 모든 과학기술 분야 및 산업에 걸쳐 큰 파장을 일으키고 있다. 비교적 단순한 자연어 처리를 위한 기술 개발로 시작되었던 언어 모델 연구는 최근 모델의 규모가 비약적으로 증대되면서 예상을 크게 뛰어넘는 역량을 보이며 이에 따라 그 활용 범위와 파급효과가 급격하게 증가하고 있다[1]. 최근의 거대언어모델들은 세상에 대한 상당한 수준의 모델(world model)을 바탕으로 복잡한 문맥의 이해 및 인지적 추론능력을 보이고 있으며 인류를 인공일반지능(artificial general intelligence) 시대의 문턱으로 빠르게 이끌고 있다고 여겨진다[2]. 많은 전문가들은 거대언어모델 기반의 생성형 AI 기술로 인해 10년 전 딥러닝 혁명과 구분되는 새로운 혁신 패러다임이 시작되었다는 데 동의하고 있다[3].

의료 분야에서도 언어모델 기반의 생성형 AI는 교육, 진료, 연구 등 다양한 영역에 적용되며 의학의 미래를 변화시키기 시작했다. 특히 기존의 AI와 대비되는 거대언어모델의 두가지 특성으로 인하여 의료 분야에서 더욱 두각을 드러낼 것으로 기대된다. 첫번째는 동적 작업 지정 및 멀티모달(multi-modal) 데이터의 입출력이 가능해짐으로써 사용자는 AI와 유연하게 상호작용하는 등 사용자의 편의성을 증대시키는 방향으로 AI의 이용 방식이 변화할 수 있다는 점이다[4,5]. 두번째는 의료 분야에 대한 사전지식이 부족한 기존의 AI와 달리, 거대언어모델은 공개되어 있는 의료 데이터로 사전학습 되었고 검색을 통한 의료분야 관련 컨텍스트에 접근할 수도 있으므로 훨씬 광범위한 의료 분야 배경지식을 갖추고 있다는 점이다. 별도의 훈련 없이 미국 의사면허시험인 USMLE를 우수한 성적으로 통과한 ChatGPT/GPT-4나[6,7] 프롬프트 학습만으로도 우수한 성능을 보인 MedPaLM 등의 거대언어모델들이 이를 뒷받침한다[5,8].

이러한 특성으로 인해 거대언어모델의 의료 분야에의 적용 가능성은 무궁무진하며 현장 적용 가능성에 대해서 활발하게 검증되고 있다. 예를 들어, 거대언어모델은 방사선 검사 결과, 퇴원기록 작성과 같은 정보의 해석 및 압축과 관련된 반복적인 작업에서 의사를 보조할 수 있다[9,10]. 복잡하고 많은 양의 의료정보를 빠르게 분석하고 정리하여 의료진의 부담을 줄이며, 의료진이 더 집중적으로 환자 치료에 시간을 투자할 수 있게 돕는다. 한편, AI는 질병 진단 및 의사결정의 보조를 위해서도 활용될 수 있다[4,11]. 비정형 임상 기록 데이터를 이용하여 개발된 의료 용어(medical language)를 위한 거대언어모델 NYUTron은 재입원 가능성, 병원 내 사망률, 합병증 및 입원기간 예측 등 광범위한 예측이 가능함을 보였다[12]. 뿐만 아니라, 현재 거대언어모델은 비록 때때로 환자 개개인의 상황에 맞는 정보를 제공하지 못하는 경우가 있지만, 임상 사례를 해석하고 관련 질문에 답할 수 있는 능력을 갖추어[13] 의료진과의 대면 진료가 어려운 상황에서 질병의 초기 진단에 기여할 수 있을 것으로 기대된다. 또한, 거대언어모델의 의료분야 적용을 통하여 의사-환자의 정보 비대칭성 해소에 기여하고 의사의 번아웃을 방지할 수 있다. 이해하기 어려운 의학적 용어나 문장을 일반적인 언어로 번역하거나 해석해줄 수 있으며 간단한 질의응답을 통해 환자와 의료진 사이의 커뮤니케이션을 보조하는 역할도 수행할 수 있다[4]. ChatGPT가 환자 질문에 대한 답변과 의사가 제공한 답변(여가 시간에 소셜 네트워크에서 작성한 답변)을 비교했을 때, 의사

* Corresponding author
Sa-Yoon Park, 214, Yulgok-ro, Jongno-gu, Seoul, 03122, Republic of Korea
·E-mail : psy9228@gmail.com ·Tel : +82-2-6072-5383
Chang-Eop Kim, 1342, Seongnam-daero, Sujeong-gu, Seongnam-si, Gyeonggi-do, 13120, Republic of Korea
·E-mail : eopchang@gachon.ac.kr ·Tel : +82-31-750-5493
·Received : 2023/09/08 ·Revised : 2023/10/19 ·Accepted : 2023/10/24

pISSN 1738-7698 eISSN 2288-2529 http://dx.doi.org/10.15188/kjopp.2023.10.37.5.134
Available online at https://kmpath.jams.or.kr

판정단이 정성적 지표로 평가했을 때 품질과 공감도 측면에서 모델의 답변이 더 선호된 바 있다[14]. USMLE에서 GPT-4와 Med-PaLM 2의 강력한 성능을 고려할 때 교육 측면에서도 거대언어모델은 현재 의사자격시험에서의 성취도가 낮은 학생들에게 유용한 교육 도구가 될 수 있음을 시사한다[6]. 언어 학습 플랫폼 듀오링고(Duolingo) 및 온라인 비영리 교육기관 칸 아카데미(Khan Academy) 등에서 AI-assisted 학습 도구를 개발하고 구현하였으며 이와 유사한 도구를 의학 교육에 적용하여 현재의 의학 교육을 개선시킬 수 있을 것으로 기대된다[7,15]. 이처럼 거대 언어 모델은 의료 분야에서 다양한 역할을 수행할 수 있으며, 이는 의료 서비스의 질 향상, 의료진의 업무 효율성 증대, 그리고 환자의 건강 관리 개선에 기여할 수 있다.

한의학과 같은 동아시아 전통의학은 이러한 변화에 어떻게 대응해야 하며 새로운 기술들을 어떻게 활용해야 하는가? 거대언어모델은 foundation 모델의 학습데이터에 크게 의존적이므로 이러한 거대언어모델의 활용 방식과 효율성에서 현대 서양의학과 큰 차이가 있을 수밖에 없을 것이다. 실제로 장 등[16]의 논문에 따르면 현재 가장 우수한 성능을 보이는 OpenAI의 GPT-4의 경우 한의사국가고시를 통과하는 합격선에 근접하는 성적을 보였지만 많은 한계점 역시 확인되었다. 예를 들어, 국제적으로 표준화된 진단기준에 대한 문제가 주로 출제된 신경정신과학 과목에서는 83.75%의 높은 정답률을 보인 반면, 서양의학 뿐 아니라 중의학과도 차별화되는 이론을 다루는 내과학2 과목은 28.75%의 가장 낮은 정답률을 보였다. 한국의 의료법을 다루는 보건의약관계법규 과목의 경우에도 한의학과 직접적인 연관이 없음에도 불구하고 48%의 낮은 정답률에 머물렀다. 또한 영어로 번역하여 질의 시 한의학 관련 응답의 퀄리티가 전반적으로 더 좋게 나타남을 확인하였다. 종합적으로 보았을 때, 거대언어모델이 복잡한 전통의학에 대해서도 유용하게 활용될 수 있는 가능성을 확인하였으나 주로 영미권에서 생산된 데이터로 모델이 학습되었기 때문에 한의학을 포함한 동아시아 전통의학 및 한글 데이터에 대한 추가 학습이 필요한 상황임을 알 수 있었다. 현재 시점에서 상대적으로 부족한 한의학 지식을 추가 학습시킬 수 있다면 거대언어모델은 한의학 분야에서도 신뢰성 있는 참조 도구로서의 역할을 수행할 수 있을 것으로 기대된다.

이 새로운 기술이 동아시아 전통 의학에 어떤 변화를 가져올 수 있을지 미래를 예측하고 실질적인 노력을 하기 위해서는 기존의 인공지능 개발 및 적용과는 매우 다른 방식으로 전개되고 있는 거대언어모델 기반 인공지능의 특성을 이해해야 한다. 본 고에서는 먼저 거대언어모델의 학습 및 작동 원리, 그리고 한의학에의 적용 방법에 대해 소개하고 이를 바탕으로 향후 한의학 교육에의 활용 방안에 대해 살펴보도록 한다.

본 론

1. 거대언어모델의 학습 및 작동 원리

거대언어모델에 한의학을 추가로 학습시키고 한의학 교육 및 임상에 활용하기 위해서는 먼저 거대언어모델의 학습 및 작동 원리에 대한 이해가 필요하다. 모델이 어떻게 문맥을 이해하고, 그 내용에 따라 대답을 생성하는지에 대해 이해하기 위해 거대언어모델의 학습 및 작동 원리에 대해서 먼저 살펴보고자 한다. 최근 GPT 기반 대화형(assistant) 언어모델들의 학습 파이프라인에 대해서 살펴보면 크게 3단계로 구성되며 서로 순차적으로 이어진다[17,18]: 1) 비지도 사전학습(Unsupervised pretraining) 단계, 2) 지도 미세조정(Supervised finetuning; SFT) 단계, 3) 인간 피드백 기반 강화학습(Reinforcement learning from human feedback; RLHF) 단계(Fig. 1).

1) 비지도 사전학습 단계(Unsupervised pretraining)

사전학습 단계는 전체 학습 시간 중 가장 많은 시간을 차지하며 모델 전체 학습과정에서 이용되는 연산 자원의 대부분이 사전학습 단계에서 필요하다. 기본적인 사전학습 과정은 트랜스포머(transformer) 구조의 모델에 대량의 문서들을 학습 데이터로 넣어주고, 입력된 문장이나 문단에 대해 다음에 올 단어를 예측하게끔 하는 것이다. 이 과정은 자기지도학습(self-supervised learning)이라고 불린다. 이 과정을 통해 도출된 모델은 베이스 모델 혹은 foundation 모델로 아직 대화형(assistant) 모델이 아니다. 즉, 베이스 모델은 질문에 답변하는 것이 아니라 주어진 문장에 대해 이어지는 문장을 생성하는 모델이다. 그러나 이 과정을 통해 모델은 문법, 문맥, 심지어 일부 상식과 같은 언어에 대한 광범위한 이해를 획득하기 때문에 베이스 모델 자체로도 일반적인 표현을 학습하며, 임의의 다운스트림 태스크(downstream task)에 효율적으로 미세조정(fine-tuning)이 가능하다.

사전학습을 위한 학습 데이터는 대규모 텍스트 데이터셋으로 인터넷에서 추출한 책, 기사 웹 페이지, GitHub, 위키피디아, 아카이브 등으로 구성되며, 모든 데이터셋을 혼합한 후 주어진 비율에 따라 샘플링하여 학습한다. 그런데 이 때 실제 모델을 학습시키기 위해서는 데이터셋을 트랜스포머 모델의 입력 형식에 맞추기 위한 토큰화(tokenization)라는 전처리 과정을 거쳐야 한다. 토큰화는 자연어 말뭉치(corpus)를 의미 있는 단위로 분절하여 숫자(integer)로 구성된 시퀀스로 변환하는 작업으로 텍스트에서 토큰으로, 다시 숫자로 변환되는 과정을 거친다. 토큰화를 위한 알고리즘으로는 Byte-pair encoding 방법이 주로 사용된다.

수집된 학습 데이터인 문서들은 트랜스포머에 크기 (B × T)의 행렬 형태로 입력된다. 여기서 B는 batch size를 의미하고, T는 maximum context length(최대 컨텍스트 길이)를 의미한다. 입력 시에 문서를 context length 단위로 묶은 후 행으로 쌓고, 특수 토큰 [endoftext]으로 문서들을 구분하여 새로운 문서가 시작되는 위치를 알려준다.

이때, 연구자가 정한 하이퍼파라미터(hyperparameter)에 따라서 모델의 구성과 학습 결과에 차이가 있게 되는데, 중요한 하이퍼파라미터의 예로는 모델의 파라미터 크기, context length, vocabulary size 등이 있다. 파라미터 크기는 모델의 표현력과 성능, 복잡도, 훈련 및 추론 비용에 직접적인 영향을 미치는데 최초 개발된 ChatGPT의 사전훈련모델인 GPT-3는 175B의 파라미터 크기를 갖고 있으며[17], OpenAI의 폐쇄형 소스 정책에 대항하여 Meta AI가 개발한 LLaMA는 3B-65B의 크기를 갖고 있다[19]. GPT-4의 파라미터 크기는 공식적으로 알려진 바가 없으며, LLaMA2는 7B-70B의 크기를 갖고 있다. 한번에 입력 받을 수 있는 토큰의 양인 context length는 언어모델의 단기기억 용량이라고도 볼 수 있는데, GPT 3.5나 4는 대략 4K-32K, LLaMA와 LLaMA2의 경우 2K-4K의 context length를 갖고 있다[20]. 현존하는 언어모델 중 가장 긴 context length를 갖고 있는 모델은 Anthropic의 Claude로 100K의 context length를 지원한다. 대부분의 거대언어모델들은 10K 이상의 vocabulary size로 훈련되었는데, 예를 들어 GPT-3는 50K, LLaMA는 32K의 vocabulary size를 갖고 있다. 최적의 하이퍼파라미터 조건 뿐 아니라 적절한 데이터셋 구성이 어떠해야 할지에 대해선 활발한 실험과 연구가 진행되고 있다. 예를 들어 LLaMA는 GPT-3보다 훨씬 적은 파라미터 크기를 갖는 대신 (175B vs. 65B), 훈련에 이용된 데이터셋의 토큰 수를 비약적으로 증가시킴으로써 (300B vs. 1.4T) 효과적으로 성능을 확보하는 전략을 택했다.

2) 지도 미세조정 단계(Supervised fine-tuning; SFT)

두번째 단계는 SFT 단계로 사전 훈련된 베이스 모델(pre-trained base model)을 이용하여 질문에 대한 응답이나 번역 등 원하는 특정 작업에 대한 학습을 시행한다. 전체 모델의 파라미터가 업데이트되며, 이 과정을 통해 도큐먼트 완성을 위한 모델이 아닌 질문에 대하여 답변하는 대화형 모델이 도출된다. SFT 단계를 위해서는 대략 수 만개 정도의 상대적으로 작지만 고품질의 데이터셋 수집이 필요하며, 데이터셋은 일반적으로 인간 라벨러(labeler)의 도움으로 작성된 프롬프트와 이상적인 응답 쌍으로 구성된다. 이후 이 데이터셋을 이용하여 학습이 수행되며, 이 때의 학습 알고리즘은 1단계인 사전학습 단계와 같은 next token prediction이다. 사전학습 단계와 비교하면 다량의 저품질 인터넷 도큐먼트 학습에서 소량의 고품질 프롬프트-응답 데이터셋 학습으로 변경된 것이다. 이 때 모델이 학습할 이상적인 응답 데이터 확보를 위해서 인간 라벨러는 라벨링 문서 작성 지침에 따라서 도움이 되고(helpful), 진실하며(truthful), 무해한(harmless) 내용을 작성하도록 요청 받는다.

3) 인간 피드백 기반 강화학습 단계(Reinforcement learning

from human feedback; RLHF)

마지막 단계는 리워드 모델링(Reward modeling) 단계와 강화학습(Reinforcement learning) 단계로 구성된다. 리워드 모델링 단계에서는 동일한 프롬프트에 대해 이전 단계에서 생성된 SFT 모델을 이용하여 얻은 여러 개의 출력값(응답)에 대해 인간 라벨러가 점수를 매긴 데이터셋을 이용한다. 인간 라벨러가 답변을 생성하는 것보다 생성된 답변에 대한 평가를 하기가 훨씬 수월하기 때문에 더 많은 데이터셋을 얻을 수 있다. 이 데이터셋을 이용하여 프롬프트에 대한 응답을 보고 점수를 예측하는 리워드 모델을 학습(지도학습)시킨다. 리워드 모델을 통해 모델 응답의 완성도를 점수화할 수 있다.

강화학습 단계에서는 학습된 리워드 모델을 이용하여 도출된 리워드 점수를 이용하여 SFT 모델을 강화학습한다. 강화학습을 통해 높은 점수를 받았을 때 생성된 단어들(응답)의 샘플링 확률을 더 높아질 것이고, 낮은 점수를 받았을 때 생성된 단어들의 샘플링 확률은 더 낮아질 것이다.

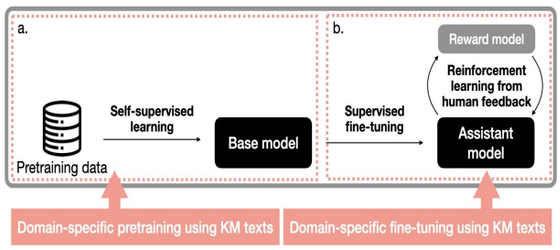

Fig. 1. Making a domain-specific large language model by directly manipulating the model's parameters. a. Pretraining a base model on an extensive corpus of Korean medicine data. b. Fine-tuning an assistant model's parameters using Korean medicine-related question-answering datasets. KM, Korean medicine.

2. 거대언어모델에게 한의학을 학습시키는 방법

거대언어모델에 한의학 지식을 추가하고 원하는 작업을 수행하게끔 하기 위한 방법에 대해 살펴보고자 한다. 한의학 지식을 거대언어모델에 학습시키기 위해서는 크게 모델 내부를 직접 조작하거나, 직접 조작하지 않고 학습시키는 두 가지 접근을 고려해볼 수 있다. 아래에 소개한 방법은 배타적으로 사용할 필요는 없으며 여러 방법의 중복 적용이 가능하다.

1) 모델 내부를 직접 조작

첫번째로 사전훈련 단계에서의 조작 즉 베이스 모델 생성 단계에서부터 한의학 문서로 학습시키는 방법이 있다(Fig. 1a). 이 방법은 가장 근본적인 수준에서 한의학 지식과 체계를 학습시킬 수 있는 방법으로 한의학과 관련된 광범위한 주제에 대해 학습하고, 이를 바탕으로 신규 정보를 이해하거나 기존 정보를 활용하는 능력을 향상시킬 수 있다. 하지만, 현재 상황에서 베이스 모델의 사전 훈련은 연산량과 학습 데이터 양 측면에서 모두 실현 가능성이 낮은 방법이다.

두번째로 고려할 수 있는 방법은 SFT 단계에서의 조작이다(Fig. 1b). SFT 과정을 더 세분화하면 단순히 문장을 완성하는 형태의 사전학습 모델을 대화형 모델로 만들기 위한 instruction-tuning과 특정 분야에 대한 구체적인 도메인 지식을 학습시키는 fine-tuning의 두 단계로 나누어 생각할 수 있다. 한의학 주제의 대화가 가능한 대화형 모델을 만드는 것이 목적이라면, 이미 instruction-tuning이 끝난 모델들을 대상으로 양질의 한의학 텍스트 데이터에 대한 fine-tuning을 진행하는 것이 효율적인 전략일 것이다.

Instruction-tuning 모델의 선택은 크게 오픈소스 모델과 폐쇄형 모델로 나눠진다. 오픈소스(open-source) 진영의 최초 fine-tuned 모델인 알파카(Alpaca)가 개발된 이후, 알파카의 방식을 따라하거나 발전시키면서 다양한 오픈소스 대화형 모델들이 개발되었다. 알파카는 Meta AI의 오픈소스 pre-trained model인 LLaMA에 ChatGPT가 생성한 데이터로 SFT하여 개발된 모델이다. 그리고 최근 2023년 7월 LLaMA2가 연구 및 상업적 용도로 사용할 수 있는 무료버전으로 일반에게 공개되었다[20]. 특히 사전 학습된 베이스 모델(LLAMA2)뿐만 아니라 fine-tuned 대화형 모델(LLaMA2-Chat)까지 두 가지 버전으로 공개됨에 따라 LLaMA2-Chat을 기반으로 양질의 한의학 텍스트를 추가 학습시킴으로써 목적하는 용도에 맞는 거대언어모델을 개발할 수 있다. 이 접근법은 사전 훈련된 모델을 사용하고, 특정 작업에 맞게 가중치를 업데이트하여 한의학에 대한 이해를 향상시킬 수 있지만, 연산량과 학습시킬 데이터를 고려해야 한다. 연산량 측면에서는 LLaMA2-Chat의 경우 모델 파라미터 크기에 따라 각각 7B, 13B, 70B의 세 종류의 모델로 공개되었으므로 7B 혹은 13B 파라미터 크기를 지닌 모델 이용을 고려해볼 수 있다. 모델 학습을 위한 알고리즘은 사전학습 단계와 같은 next token prediction을 이용하며, 학습에 이용되는 데이터는 큐레이션을 거친 라벨이 있는 데이터(labelled data)일 수도 있고, 라벨이 없는 데이터(unlabeled data)를 가지고 자기지도학습을 할 수도 있다. 다만, LLaMA 기반의 베이스 모델의 경우 영어 중심으로 훈련되어서 한국어부터 학습시킨 후 한의학 텍스트를 학습해야 하는 점을 추가로 고려해야 한다.

한편, 폐쇄형 모델을 이용하는 대표 케이스로는 API를 이용하여 OpenAI의 GPT-3.5 Turbo를 훈련시키는 방법을 들 수 있다. 현재(2023.09) GPT-3.5 Turbo 모델에 대한 fine-tuning이 가능하며, 2023년 중으로 GPT-4에 대한 fine-tuning 기능도 제공할 것이라고 밝힌 바 있다. GPT-3.5 Turbo 모델에 대해서 API를 이용하여 fine-tuning하는 방법에 대해서는 OpenAI 홈페이지(https://platform.openai.com/docs/guides/fine-tuning/fine-tuning-examples)에 자세히 나와있으며 사용자가 준비해야 하는 것은 fine-tuning을 위한 데이터를 요구하는 포맷에 맞게 JSON으로 작성하여 파일로 업로드할 수 있도록 해야 한다.

두 방식의 장단점을 간략하게 정리하면 다음과 같다. 오픈소스 모델은 훈련자체에 드는 비용이 폐쇄모델 대비 상대적으로 저렴하며, 데이터 유출 우려가 없다는 점이 장점이다. 반면 훈련을 위한 연산자원을 확보하고 직접 훈련을 진행하는 과정에 약간의 기술적 장벽이 있을 수 있다. 현재기준 전반적으로 오픈소스 모델들의 성능이 폐쇄형 모델의 성능에 못 미치는 상황이다. 한편 폐쇄형 모델의 경우 현재 OpenAI의 기본 모델 훈련 비용이 상대적으로 비싸며, API 호출을 통해 OpenAI 서버 내로 데이터를 보내 서버 내 모델을 훈련시키는 방식이므로 데이터 보안에 대해여 담보할 수 없다. 또한 본인이 훈련시킨 모델을 온전히 보유하고 자유롭게 조작할 수 없으며 OpenAI가 서비스하는 자원과 기능에 의존해야 한다는 점이 단점으로 작용한다. 반면 API 호출을 이용하므로 훈련과정은 보다 수월하며, 성능이 오픈소스 모델에 비해 더욱 우수한 점이 장점이다.

2) 모델 내부를 직접 조작하지 않는 접근방법

모델 내부를 직접 조작하지 않고도 거대언어모델 기반 AI의 성능을 향상시키거나 새로운 도메인 역량을 확보하는 방법은 크게 프롬프트 엔지니어링(Prompt Engineering), RAG(Retrieval augmented generation), Gluing의 세 가지 방법이 있다(Fig. 2).

첫번째, 프롬프트 엔지니어링 혹은 In-context learning은 프롬프트(prompt)를 통해 모델에게 적절한 컨텍스트를 제공함으로써, 모델이 필요한 지식과 상식을 학습시키는 방식이다(Fig. 2a). 프롬프트 엔지니어링이라는 명칭은 행위에 초점을 두고 부르는 것이고, in-context learning은 행동의 변화를 이끌어냈다는 결과에 포커스를 둔 표현이라고 생각할 수 있다. 기본적으로 모델 내부의 웨이트를 직접 조절하지 않고도 AI의 성능 개선이 가능하다는 점은 기존 인공지능 모델에서는 생소한 개념인데 이것이 가능한 이유는 인터넷에 있는 오픈 데이터들로 학습된 LLM의 경우, 인간의 모든 지식이 모델 안에 담기어 숫자로 재구성된 것이기 때문이다. 프롬프팅은 적절한 출력을 구성할 수 있도록 수학적 조건을 설정하는 것이다. 중요한 점은 그 수학적 조건이 도메인 특이적이며, 거대언어모델을 해당 도메인에 적용하기 위해서는 조정이 필요하다는 것이다. 그러나 모델 내부의 웨이트(가중치) 자체는 변하지 않는다. 즉, 모델은 인간의 모든 지식을 습득했고, 당신의 도메인 지식 역시 궁극적으로 인간의 지식에 기반을 두고 있기 때문에 이미 도메인에 대한 잠재적인 지식을 가지고 있어 모델의 웨이트는 바꿀 필요가 없는 것이다.

API 호출을 통해 학습이 완료된 모델에 user와 assistant의 대화 내용을 입력하고, system으로는 모델의 태도를 입력하는 등 적절한 대화의 맥락을 제공함으로써 모델의 파라미터를 업데이트하

지 않고도 모델의 응답을 향상시킬 수 있다. 예를 들어, 삼행시가 무엇인지 모르고 있던 모델에 삼행시의 개념에 대해 예시 없이 설명만으로도(context 제공) 삼행시를 지을 수 있게 되거나(zero-shot learning) 한 개의 예시를 통해 학습이 이루어진다(one-shot learning). 또한 프롬프트 엔지니어링을 통해서 모델이 최대한 깊이 있는 수준의 사고를 하도록 유도할 수 있다. 예를 들어, 'Pretend you have IQ 120' 또는 'Think step by step' 등과 같은 트리거 프롬프트(trigger prompt)를 사용하여 모델이 더 복잡하고 심도 있는 질문에 대해 답변하도록 할 수 있다.

이는 GPT-3에서 본격적으로 나타난 창발적인(emergent) 현상이며[21], 이미 사전훈련 과정에서 학습에 필요한 기반지식과 상식, 모델을 갖추고 있기 때문에 프롬프팅 만으로 학습을 한 것으로 일종의 메타 학습(meta learning)이라고 할 수 있다. 하지만 이 방법의 한계는 결국 모델이 self-attention을 유지할 수 있는 사이즈인 컨텍스트 윈도우(context window)의 길이에 따른 제약이 있다는 점으로, 모델이 받아들일 수 있는 컨텍스트의 양에 제한이 있기 때문에 대량의 복잡한 개념 주입이 필요한 경우에는 적합하지 않다.

두번째로, RAG의 수행 방법은 다음과 같다(Fig. 2b). 도메인 지식 관련 문서를(예를 들면, 동의보감) chunk로 나누고 토큰화를 통해 벡터 임베딩하여 벡터 DB를 사전에 구축해 둔다. 사용자의 query가 들어오면 질의 역시 벡터 임베딩한 후 query 벡터와 DB 벡터 간의 유사도를 계산하여 가장 관련성 높은 문서를 선택한다. API로 모델의 응답 호출 시에 관련 문서 내용을 컨텍스트로 제공하여 모델이 in-context 학습할 수 있도록 한다. 따로 지식 주머니를 달아주는 것으로 생각할 수 있다.

세번째로 Gluing[22] 또는 인지 증대(cognition augmentation) 방법은 일회성 답변 생성이 아니라, 스스로 생성한 답변들을 다시 처리하며 보다 고도화된 사고를 가능하게 하는 방법으로, 메타적 사고를 통해 보다 고차적인 모델로 만들기 위한 방법이다(Fig. 2c). Self consistency with Chain of thought(CoT-SC) 또는 Tree of thoughts(ToT) 등의 방법들이 여기에 속한다. CoT-SC는 같은 query를 여러 번 주고 이에 대한 답변들 중 일관성 있는 답변으로 voting 하는 식으로 진행된다. ToT는 매 단계에서 여러 차례 query를 주고 depth-first 혹은 breath-first search 방식으로 제일 잘한 답변을 선택하는 식으로 진행한다. ToT의 결과 성능이 900% 향상되었다는 연구 결과가 보고된 바 있다[23]. 다만, 이를 위해서는 API를 여러 번 호출해야 하고, 결과들을 접합(gluing)하는 코드가 실행되어야 해서 실용적인 측면에서 API 호출에 대한 시간과 비용이 문제가 될 수 있다.

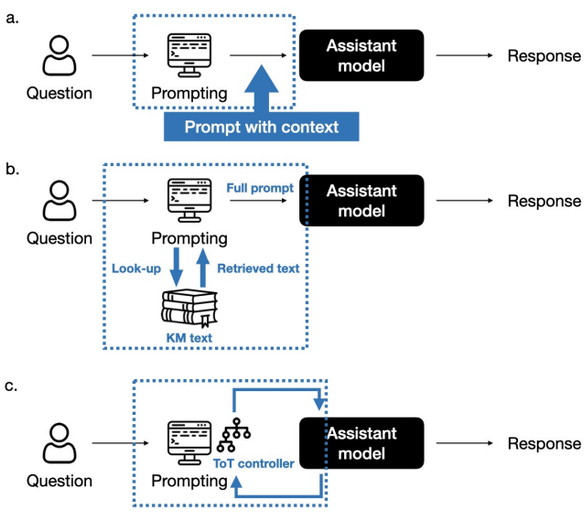

Fig. 2. Approaches to use a domain-specific large language model without internal model manipulation. a. Prompt engineering, b. Retrieval augmented generation, and c. Cognitive augmentation. KM Korean medicine; ToT, Tree of thoughts

3. 거대언어모델의 한의학 교육에의 활용

거대언어모델은 한의학 교육에서 여러 가지 방식으로 활용될 수 있다.

개인 맞춤형 학습 계획 및 학습 자료 제공: 거대언어모델은 상호작용적으로 동작하고 흥미로운 학습 자료를 생성함으로써 기존의 한의학 교수방법론을 혁신할 수 있다. 학생의 개인적인 강점, 약점, 지식 수준, 학습 목표 및 학습 성향에 대한 요구에 따라 맞춤형 학습 계획을 제공할 수 있다[24,25]. 또한, 학생들의 학습 요구와 선호에 따라 플래시카드, 요약 노트, 케이스 스터디 및 연습 문제 등의 맞춤형 학습 자료를 제공할 수 있다[26]. 개인의 학습 방식과 속도를 고려한 맞춤형 학습 계획은 학습의 효율성을 높이고, 학생들이 자신의 이해도를 점검하고 개선할 수 있게 도와줄 것이다.

지식 통합 및 커리큘럼 개발: 한의학은 전통적인 지식과 현대적인 연구가 결합된 영역으로 개인이 방대한 지식체계를 정리하고 통합하기란 대단히 어렵다. 거대언어모델은 다양한 자료들을 통합하고, 사용자가 필요한 정보를 더 쉽고 빠르게 찾을 수 있도록 돕는다. 특히, 한의과학 분야의 최신 연구 추세와 발전에 기반하여 새로운 주제 및 소주제를 교육 과정에 포함시키는 것을 제안할 수 있다. 또한 교육 과정 자료, 강의 노트, 평가 과제 등을 개발하고 수정하는 데 사용될 수 있다. 이를 위해 거대언어모델을 한의학 교과서, 논문 및 그 외 학습 자료로 훈련하여 생성된 콘텐츠가 정확하고 관련성이 있도록 해야 한다.

학생 평가 자동화: 거대언어모델을 이용하여 객관적 평가, 예를 들어 객관식 문제나 빈칸 채우기 문제 뿐만 아니라 에세이나 케이스 스터디와 같은 서술형 문항에 대해서도 채점을 자동화할 수 있다. 에세이나 케이스 스터디와 같은 소논문에 대한 평가 시에 거대언어모델은 글의 구조, 일관성 및 관련성에 대한 자세한 피드백을 즉시 제공할 수 있다는 점에서 유용성이 부각된다.

가상 환자 시뮬레이션 실습: 거대언어모델은 현실적인 가상 환자 시나리오를 제공하여 한의과대학 학생들이 환자 상황을 모의하고 실습하는 데 사용될 수 있다. 예를 들어, 실습의 목표를 입력하면 모델은 해당 증상이나 질환에 대한 환자의 역할을 수행하고, 학생은 진단과 치료 방법을 제시하는 연습을 할 수 있다[27]. 뿐만 아니라, 거대언어모델은 모델의 질문에 대한 학생의 응답을 해석하고, 피드백을 제공함으로써 학생이 한의학의 임상적 의사결정 과정을 연습하고, 환자와의 대화 기술을 습득할 수 있게 한다. 뿐만 아니라, 거대언어모델에 VR 기술을 접목함으로써 실제 임상 현장과 유사한 상황을 시뮬레이션 함으로써 부족한 임상 경험을 보충하는 데 도움 받을 수 있다.

연구 및 학문 활동에서의 활용: 거대언어모델은 한의학 연구에서도 활용될 수 있다. 대량의 데이터와 정보를 훑어서 관련 연구를 식별하고 결과를 종합함으로써 새로운 연구 아이디어를 도출하거나, 기존 연구를 검토하고 요약하는데 사용될 수 있다. 특히 문헌 리뷰와 같은 연구 활동에서 방대한 양의 학술 자료를 빠르게 검토하고, 연구자들이 효율적으로 정보를 수집하고 분석할 수 있게 돕는다[28-30]. 문헌 검토, 데이터 분석, 연구 논문 및 보고서 초안 작성 등의 업무를 수행함으로써 연구에 대한 결과물 도출을 촉진하고 가속화하는 데 도움을 줄 수 있다.

다만, 위에서 살펴본 거대언어모델을 활용한 방법들이 실제로 효과적인 교육 도구로 작동하려면, 거대언어모델의 한의학에 대한 정확하고 심도 있는 이해가 필수적이다. 이를 위해서는 한의학 분야의 전문가들이 모델의 훈련과 평가에 참여하고, 모델이 올바른 정보를 제공하고 있는지를 검증하는 과정 등 적합한 모델 평가 방법에 대한 탐구가 병행되어야 할 것으로 생각된다.

결 론

거대언어모델 기반의 생성형 인공지능 기술은 의료 분야에서도 혁신을 가져오고 있다. 이러한 기술은 복잡한 문맥 이해와 추론 능력을 활용하여 의사의 진단과 의사결정을 보조하며 환자와 의료진 간 의사소통을 보완할 수 있다. 한의학 분야에서도 이러한 기술의 활용이 기대되며, 한의학 지식을 거대언어모델에 통합시키려면 한의학 데이터를 이용한 추가적인 학습이 필요하다. 한의학을 거대언어모델에 훈련시키는 방법에는 크게 두 가지 접근법이 있다. 첫번째는 모델 내부를 직접 조작하는 방법으로, 사전훈련을 통해 베이스 모델을 대량의 한의학 데이터를 기반으로 학습시키거나 한의학 지식에 관한 질의응답 데이터를 이용하여 이미 훈련된 모델의 파라미터를 미세조정하는 방법이다. 두번째로는 모델 내부를 조작하지 않는 방법으로, 프롬프트 엔지니어링, 정보 검색 기반 방법(RAG),

그리고 ToT 등 인지 증대 방법을 활용할 수 있다. 한의학 도메인에 특화된 거대언어모델은 개인 맞춤형 학습 계획 및 학습 자료 생성, 지식 통합과 커리큘럼 개발, 학생 평가 자동화, 가상 환자 시뮬레이션, 그리고 연구 및 학문 활동에 다양한 방식으로 활용할 수 있을 것으로 기대된다. 그러나 이러한 활용을 위해서는 모델의 정확성과 전문성을 확인하고, 전문가들의 참여를 통한 검증 과정이 필수적이다. 거대언어모델을 통한 한의학 교육의 발전은 학습 효율성과 교육 혁신을 도모할 수 있는 새로운 전망을 제시하며, 실제 교육 현장에서의 적용을 위해 활발히 탐구 되어야 하는 중요한 주제이다.

감사의 글

이 논문은 2022년과 2023년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2022R1F1A1068841, RS-2023-00248152).

References

1. Eloundou T, Manning S, Mishkin P, Rock D. Gpts are gpts: An early look at the labor market impact potential of large language models. arXiv preprint arXiv:230310130. 2023.
2. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:230312712. 2023.
3. Miao Q, Zheng W, Lv Y, Huang M, Ding W, Wang F-Y. DAO to HANOI via DeSci: AI paradigm shifts from AlphaGo to ChatGPT. IEEE/CAA Journal of Automatica Sinica. 2023;10(4):877-97.
4. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259-65.
5. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023:1-9.
6. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:230313375. 2023.
7. 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 digital health. 2023;2(2):e0000198.
8. 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 arXiv:230509617, 2023.
9. Arora A, Arora A. The promise of large language models in health care. The Lancet. 2023;401(10377):641.
10. 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(4):e230725.
11. Kraljevic Z, Bean D, Shek A, Bendayan R, Yeung JA, Deng A, et al. Foresight--Deep Generative Modelling of Patient Timelines using Electronic Health Records. arXiv preprint arXiv:221208072. 2022.
12. 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. 2023:1-6.
13. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. Does ChatGPT provide appropriate and equitable medical advice?: A vignette-based, clinical evaluation across care contexts. medRxiv. 2023:2023.02. 25.23286451.
14. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA internal medicine. 2023.
15. Team D. Introducing Duolingo Max, a learning experience powered by GPT-4. Retrieved March. 2023;15:2023.
16. Jang D, Kim C-E. Exploring the Potential of Large Language models in Traditional Korean Medicine: A Foundation Model Approach to Culturally-Adapted Healthcare. arXiv preprint arXiv:230317807. 2023.
17. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901.
18. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems. 2022;35:27730-44.
19. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:230213971. 2023.
20. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288. 2023.
21. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models. arXiv preprint arXiv:220607682. 2022.
22. Karpathy A. State of GPT 2023 [Available from: https://build.microsoft.com/en-US/sessions/db3f4859-cd30-4445-a0cd-553c3304f8e2.
23. Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, et al. Tree of thoughts: Deliberate problem solving with large language models. arXiv preprint arXiv:230510601. 2023.
24. Frommeyer TC, Fursmidt RM, Gilbert MM, Bett ES. The desire of medical students to integrate artificial intelligence into medical education: an opinion article. Frontiers in Digital Health. 2022;4:831123.
25. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nature Medicine. 2023:1-11.
26. Benoit JR. ChatGPT for Clinical Vignette Generation, Revision, and Evaluation. medRxiv. 2023:2023.02. 04.23285478.
27. Kitamura FC. ChatGPT is shaping the future of medical writing but still requires human judgment. Radiological Society of North America; 2023. p. e230171.
28. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. The Lancet Digital Health. 2023;5(4):e179-e81.
29. Biswas S. ChatGPT and the future of medical writing. Radiological Society of North America; 2023. p. e223312.
30. Chen T-J. ChatGPT and other artificial intelligence applications speed up scientific writing. Journal of the Chinese Medical Association. 2023;86(4):351-3.