재활치료과학 제 13 권 제 1 호
Therapeutic Science for Rehabilitation
Vol. 13. No. 1. 2024.
https://doi.org/10.22683/tsnr.2024.13.1.065

ChatGPT는 한국작업치료사면허시험에 합격할 수 있을까?

홍준화*, 김나연*, 민혜민*, 양하민*, 이시현*, 최서진*, 박진혁**

*순천향대학교 작업치료학과 학부생

**순천향대학교 의료과학대학 작업치료학과 교수

국문초록

목적 : 본 연구는 대규모 언어 모델에 기반한 인공지능인 ChatGPT가 한국작업치료사면허시험에 통과할 수 있는지 알아보고자 하였다.

연구방법 : 한국보건의료인국가시험원에서 제공하는 2018년부터 2022년도까지의 한국작업치료사면허시험 문항 중 공개되지 않은 작업치료실기 문항을 제외하고 작업치료학기초, 의료관계법규, 작업치료학 문항을 활용하였다. 시험문항과 함께 가장 적절한 정답을 제시하도록 프롬프트를 영어로 구성하였고 이를 입력한 후 ChatGPT가 제시하는 답을 채점하였다. 2명의 연구자가 독립적으로 전체 과정을 진행하였으며, 2명의 연구자 채점한 정확도를 평균으로 5개년도의 시험에 대한 합격 여부를 확인하였고 연구자 간 ChatGPT 답에 대한 일치도를 확인하였다.

결과 : ChatGPT는 2020년에서만 합격하였고 나머지 4개년도 시험은 탈락권 점수를 보였다. 구체적으로 의료관계법규 문항의 정확도는 25~57% 범위를 보였고 다른 문항의 정확도는 모두 60% 이상을 기록하였다. 또한 의료관계법규 문항을 제외한 연구자 간 ChatGPT는 높은 일치도를 보였으며, 이는 정확도와 유의미한 상관관계를 보였다.

결론 : 언어나 문화권에 영향을 받는 문항의 경우 아직 ChatGPT를 적용하는 데 제한이 있음을 확인하였다. 추후 프롬프트의 최적화 작업과 함께 지속적인 데이터의 학습에 따라 작업치료학을 전공하는 학생들의 학습 도구로서 활용될 수 있는지에 대한 지속적인 연구가 필요하다.

주제어 : 대규모 언어 모델, 인공지능, 작업치료면허시험

교신저자 : 박진혁(roophy@naver.com)

|| 접수일: 2023.09.25 || 심사일: 2023.10.12
|| 게재승인일: 2023.11.14

Ⅰ. 서 론

최근 수십 년 동안 기계학습 중 심층학습의 발달은 다양한 산업 분야에서 인공지능을 혁신적으로 바꾸어 왔다(Wang & Siau, 2019). 특히 이미지, 텍스트 등의 자료를 정확하게 분류할 수 있는 인공지능의 능력은 인간 수준의 언어 번역을 가능하게 했다(Wang & Siau, 2019). 이런 흐름에 따라 최근에는 자연어 입력에 기반한 대규모 언어 모델(large language model)에 기반한 인공지능을 활용한 연구가 활발히 진행 중이다.

특정 영역 자료에 제한된 기존의 인공지능과는 달리 대규모 언어 모델에 기반한 인공지능은 영역에 구애받지 않는 자료를 분석할 수 있는 장점이 있다(Brown et al., 2020). 대규모 언어 모델의 이러한 특성은 특정 영역에 특화된 학습 자료를 구축해야 하는 업무를 줄일 수 있어 다양한 분야에 활용되고 있다. 특히 의료 분야에서 개인 맞춤형 건강관리, 질병 예측, 의료영상 분석을 포함한 다양한 업무에 대규모 언어 모델 기반 인공지능의 활용 가능성의 탐색이 적극적으로 이루어지고 있다(Sallam, 2023).

따라서 의료 분야에서 대규모 언어 모델 기반 인공지능을 대표하는 ChatGPT (Chatbot Generative Pre-trained Transformer)를 활용하는 데 많은 관심이 높아지고 있다. ChatGPT는 인터넷 접근을 통해 얻을 수 있는 텍스트 자료를 기반으로 학습을 한 대규모 언어 모델 기반 인공지능을 기반으로 하고 있으며, 자연어 처리를 위해 미세조정되었다(Castelvecchi, 2022). 선행연구에서는 ChatGPT의 의학 지식을 기반으로 환자와 상호작용하는 능력을 평가하기 위해 의학면허시험에서 ChatGPT의 능력을 평가하기도 하였다. 실제 ChatGPT는 미국 의학면허시험에서 의학 관련 자연어 질문에 대해 합격 수준의 정확도(60%)를 달성하였다(Kung et al., 2023). 이는 ChatGPT가 전문적인 훈련을 받은 인간의 도움 없이도 미국 의학면허시험에 합격 수준의 능력을 보였다는 것을 시사하며, ChatGPT의 활용 가능성을 확인하였다는 데 의의가 있다. 다른 선행연구에서는 ChatGPT가 일본 의학면허시험 및 대만 약사면허시험도 통과할 수 있다는 것을 보여주고 있다(Kasai et al., 2023; Wang et al., 2023). ChatGPT는 각 시험 문제에 대한 정답과 더불어 정답에 대한 이론적 배경까지 함께 제시하기 때문에 사용자는 ChatGPT 외에 추가적인 문헌 고찰 및 검색을 하지 않고서도 학습을 할 수 있다. 즉 ChatGPT는 다양한 국가에서 의학을 전공하는 학생들을 위한 의학 지원 또는 자기 학습 도구로 활용될 수 있다는 것이 여러 연구를 통해 입증되고 있다. 특히 영어로 번역을 한 이후 프롬프트로 입력하는 방식을 통해 영어권 국가가 아닌 곳에서도 ChatGPT를 의학 관련 학습 도구로 활용할 수 있다(Nisar & Aslam, 2023).

지금까지 의사, 간호사, 약사 면허시험에 대한 ChatGPT의 활용 가능성이 주로 연구되어 다른 분야에서의 ChatGPT의 적용 가능성은 아직 불명확하다(Taira et al., 2023; Wang et al., 2023). 재활시스템 내에서 의사와 함께 작업치료사 역시 중요한 역할을 수행하며, 각각의 전문성과 역할은 서로 다르다. 구체적으로 의사는 주로 질병의 진단과 치료에 중점을 둔다. 하지만, 작업치료사는 클라이언트의 작업 참여를 촉진하는 데 크게 기여한다. 구체적으로 작업치료사는 클라이언트의 신체적, 인지적, 심리적, 감각 지각적 요소와 같은 다양한 측면을 다루어 그들의 작업 참여를 촉진한다(American Occupational Therapy Association, 2020). 따라서 ChatGPT를 작업치료사 교육을 위한 학습 도구로서의 역할을 수행할 수 있는지에 대한 연구도 필요하다.

따라서 본 연구에서는 한국작업치료사면허시험의 문항을 사용하여 ChatGPT의 능력을 평가하고자 한다. 한국작업치료사면허시험은 작업치료사에게 필수적인 모든 지식 영역을 아우르고 있다(Korea Health Personnel Licensing Examination Institute, 2023b). 한국작업치료사면허시험의 문항의 난이도와 복잡성은 전문가 패널을 통해 표준화되고 규제되고 있으며, 지난 5년 동안 매우 안정된 원점수와 심리측정 속성을 가지고 있어 인공지능 평가에 적합하다(Korea Health Personnel Licensing Examination Institute, 2023a). 또한 한국작

업치료사면허시험 문항은 모두 텍스트 기반의 객관식 문항으로 이루어져 있어 ChatGPT의 능력 평가에 적절하다.

Ⅱ. 연구 방법

1. 대규모 언어 모델 기반 인공지능

ChatGPT는 OpenAI에서 개발한 고급 언어 모델로 자연어 대화 속에서 맥락에 적절한 응답을 생성하기 위해 학습 데이터를 활용한다. ChatGPT는 다른 대화 인공지능이나 챗봇과 달리 인터넷 검색이나 외부 데이터베이스에 접근하지 않고 자체 구축한 데이터를 기반으로 하는 대규모 언어 모델이다. 따라서 ChatGPT가 생성하는 모든 응답은 자체 신경망 속에서 단어 간의 추상적인 관계를 기반으로 생성된다(Kung et al., 2023). 본 연구에서는 ChatGPT-3.5 (ChatGPT-3.5) 버전을 사용하였다. 최신 버전인 ChatGPT-4 (ChatGPT-4)는 ChatGPT-3.5에 비해 더 많은 학습 데이터를 사용하여 학습이 이루어졌으며 응답도 더 빠르고 텍스트뿐만 아니라 이미지 분석도 가능한 장점이 있다. 반면, ChatGPT-4의 사용을 위해선 유료 결제가 필요하여 접근의 제한이 있고 본 연구의 특성상 텍스트 기반의 객관식 문항만을 사용한다는 점을 고려하여 조금 더 대중적인 ChatGPT-3.5를 사용하였다.

2. 입력자료

본 연구에서는 한국보건의료인국가시험원에서 제공하는 2018년부터 2022년까지 공개된 5개년도의 한국작업치료사면허시험 문항만을 이용하였다. 한국작업치료사면허시험은 한국의 3년제 또는 4년제 작업치료학과를 졸업할 예정이거나 졸업한 학생들이 응시하는 시험이다. 시험은 3교시로 구성되어 있고 1교시에는 두 개의 하위 단원으로 구성되어 있다. 1교시의 첫 번째 소단원에는 해부학, 생리학, 공중보건학 등 작업치료에 대한 기초 지식을 70문항으로 평가하고 두 번째 소단원에는 의료관계법규에 대한 지식을 20문항으로 평가한다. 2교시에는 신경계 및 근골격계 작업치료, 아동작업치료, 정신작업치료 등 작업치료 전문 지식을 100문항으로 평가한다. 3교시에는 그림과 가상의 임상 사례가 제시되는 50개의 문항으로 실기능력을 평가한다. 합격 기준은 각 교시별 평균 60% 이상이며, 전체 정답률에 관계없이 각 교시별 정답률이 40% 미만일 경우 불합격 처리된다. 이 시험에 합격한 학생은 보건복지부로부터 작업치료사 면허를 취득하고 작업치료사로 등록된다 (Korea Health Personnel Licensing Examination Institute, 2023a).

본 연구에서는 저작권 제한으로 한국보건의료인국가시험원에 공개되지 않은 3교시 문항을 이용할 수 없어 1, 2교시 문항만 이용하였다. 접근의 제한과 더불어 3교시 문항 중 주로 가상 사례를 묘사하기 위한 그림 또는 표가 포함되어 있어 ChatGPT-3.5에 활용하기 어려운 측면도 있었다. 반면, 1, 2교시 문항은 모두 정답이 한 개인 객관식 문항으로 구성되어 있고 정답에 대한 이유 또한 제시할 필요가 없어 ChatGPT-3.5에 입력하는 데 적절하다.

ChatGPT-3.5 프롬프트 입력 전 입력 데이터가 대표성을 갖출 수 있도록 1, 2교시 문항에서 각 20문항씩 무작위로 표본을 추출하였고 인터넷에 검색하여 5개년도 면허시험 문항이 색인되지 않았는지 확인하였다. 그 이후 총 950문항(연 190개 문항)을 프롬프트 입력에 사용하였다.

3. 프롬프트 입력 방법

프롬프트는 독립된 연구자 2명이 각각 입력하였다. 본 연구에서 프롬프트 입력은 면허시험 문항 그대로 진행하였다. 한 개의 정답을 제시하게 하도록 각 문항과 더불어 “다음 중 가장 적절한 답은 무엇입니까”라는 프롬프트를 일관되게 추가하였다. 이 프롬프트를 제3

의 연구자가 영어로 번역하였고, 영어번역본은 영어를 모국어로 하는 원어민에게 검토 받은 이후 프롬프트로 확정되었다(Figure 1). ChatGPT-3.5의 기억 유지 편향을 줄이기 위해 각 문항을 입력할 때마다 ChatGPT-3.5의 새로운 채팅 세션을 시작하였다. ChatGPT-3.5가 각 문항에 대해 한 개의 정답을 제시하지 못하는 경우, 질문 프롬프트를 그대로 재입력하였다. 해당 문항은 최대 세 번까지 다시 입력되었다.

4. ChatGPT-3.5 능력 평가 및 사용자 간 일치도

두 명의 연구자는 독립적으로 정확도를 계산하여 ChatGPT-3.5의 능력을 평가하였다. 정확도는 ChatGPT-3.5가 각 문항에 대해 정답 번호를 제시하거나 정답 번호에 해당하는 텍스트를 제시하는 경우 정답으로 간주하여 계산하였다. 두 명의 연구자는 각 문항의 정답을 엑셀 파일로 코딩하였고 제3의 연구자에게 전달하였다. 제3의 연구자는 엑셀 파일을 검토한 이후 ChatGPT-3.5의 정확도 및 2명의 연구자 간 일치도를 평가하였다(Figure 2).

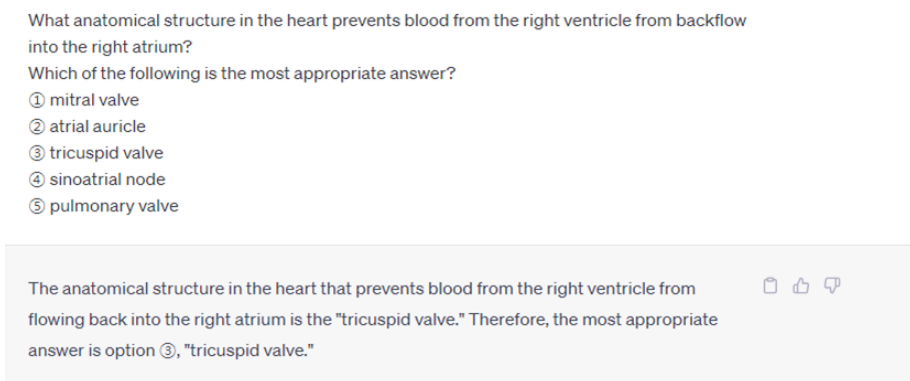

Figure 1. The Example of a Inputted Prompt and ChatGPT's Response

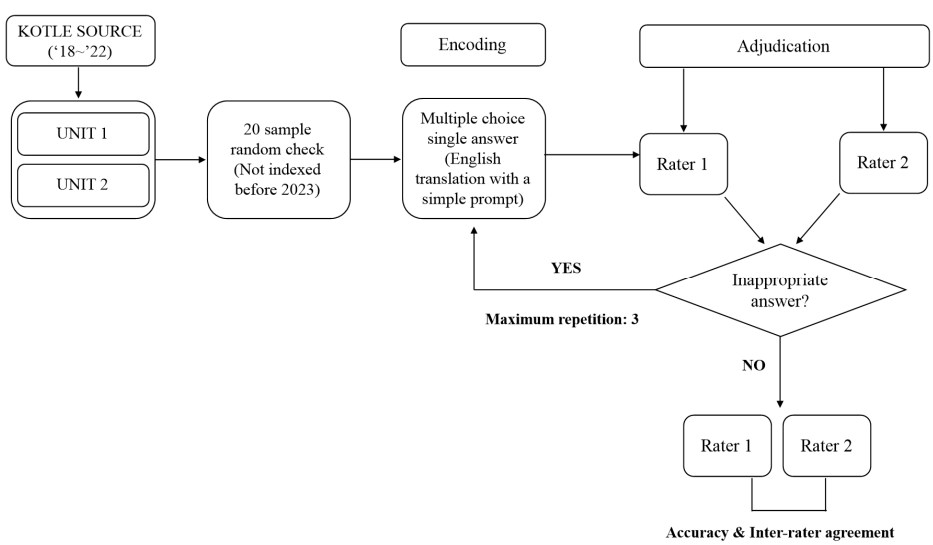

Figure 2. Flow Chart for Sourcing, Encoding, and Adjudicating

5. 통계적 분석

모든 자료는 IBM SPSS Statistics 22.0 버전(IBM Corp.)을 사용하여 분석하였다. 5개년도의 면허시험 문항에 대한 ChatGPT-3.5의 정확도는 기술통계분석을 이용하여 제시하였고 연구자 간 일치도를 구하기 위해 절대합치도를 이용하였다. 정확도와 일치도 간 상관관계를 분석하기 위해 스피어만 상관계수를 이용하였다. 모든 통계적 유의성은 $p < .05$로 설정하였다.

Ⅲ. 연구 결과

1. ChatGPT-3.5 정답률

5개년도 한국작업치료사면허시험에 대한 ChatGPT-3.5 능력을 정답, 오답, 미확인 답변으로 구분하여 제시하였다(Figure 3). 전체 시험 문항에 대한 평균 정확도는 2018년도에는 58.6%, 2019년도에는 58.4%, 2020년도에는 70.0%, 2021년도에는 64.0%, 2022년도에는 79.4%로 나타났다(Figure 3). 2018년도와 2019년도를 제외하고 평균 정확도로 보았을 때 ChatGPT-3.5는 합격권 점수를 보였다. 미확인 답변을 오답으로 간주하고 정확도를 계산하고 각 교실별 합격 기준을 적용하였을 때 2020년도에만 합격을 하였다(Table 1).

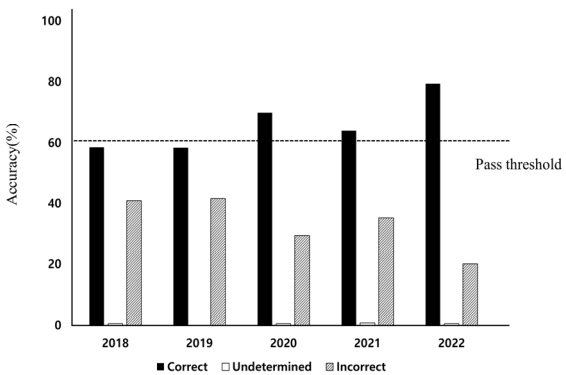

Figure 3. Accuracy of ChatGPT on the National Korean Occupational Therapy Licensure Examination

Table 1. ChatGPT Accuracy Across the Past 5 Years of the National Korean Occupational Therapy Licensure Examination

Year	Unit	Accuracy (%)			Benchmark	Absolute agreement
		First rater	Second rater	Average		
2018	Unit 1-1	77.1	84.3	80.7	Fail	.846***
	Unit 1-2	30.0	20.0	25.0		.710**
	Unit 2	75.0	65.0	70.0		.854***
2019	Unit 1-1	75.7	71.4	73.6	Fail	.819***
	Unit 1-2	45.0	30.0	37.5		.599*
	Unit 2	61.0	67.0	64.0		.718***
2020	Unit 1-1	84.2	80.0	82.1	Pass	.877***
	Unit 1-2	65.0	50.0	57.5		.810***
	Unit 2	70.0	71.0	70.5		.857***
2021	Unit 1-1	78.6	84.3	81.5	Fail	.851***
	Unit 1-2	35.0	35.0	35.0		.006
	Unit 2	75.0	76.0	75.5		.857***
2022	Unit 1-1	80.0	81.4	80.7	Fail	.970***
	Unit 1-2	35.0	35.0	35.0		.345
	Unit 2	73.0	72.0	72.5		.842***

*$p < .05$, **$p < .01$, ***$p < .001$.

2. 사용자 간 일치도

2명의 연구자 간 ChatGPT의 정답 일치도는 .006~.970의 범위를 보였다. 구체적으로 2021년도 2022년도의 1교시의 두 번째 단원에서 통계적으로 유의하지 않은 일치도를 보였고 그 외의 년도 시험 문항에서는 모두 통계적으로 유의한 일치도를 보였다(Table 1).

3. 정확도와 사용자 간 일치도의 상관관계

전반적으로 ChatGPT의 능력과 사용자 간 일치도의 상관관계는 통계적으로 유의하여, ChatGPT의 능력이 낮을수록 사용자 간 일치도가 낮음을 의미한다(Table 1).

Ⅳ. 고 찰

본 연구의 목적은 한국작업치료사면허시험에 ChatGPT-3.5의 적용 가능성을 조사하는 것이었다. 이를 위해 최근 5개년도의 한국작업치료사면허시험 문항을 ChatGPT-3.5에 입력하여 정확도를 평가하였다. 연구 결과, ChatGPT-3.5는 2020년 한국작업치료사면허시험에만 합격권의 능력을 보였다. 구체적으로 1교시의 두 번째 단원을 제외하고 5개년도 모두 60% 이상의 정확도를 보였는데, 이는 특정 문항에 대한 ChatGPT의 적용이 제한적임을 시사한다.

ChatGPT는 자연어 처리 분야에서 뛰어난 발전을 보이고 있고(Biswas, 2023), 이는 상업, 교육, 의료 등 다양한 분야에 깊은 영향을 미치고 있다(Hacker et al., 2023). 게다가 ChatGPT는 방대한 텍스트 데이터 학습을 통해 인간과 유사한 텍스트를 작성할 수 있다(Biswas, 2023). 특히 의학 분야에서는 맞춤형 데이터 학습을 통해 학생들의 시험 풀이 도구로 활용될 수 있는데(Jeblick et al., 2023), 이러한 맥락에서 ChatGPT는 의료 분야의 다양한 시험 문항을 통해 그 능력을 평가하고 있다.

본 연구에서 ChatGPT-3.5는 2018년부터 2022년까지의 한국작업치료사면허시험 중에서 2020년을 제외한 나머지 4개년도에서는 합격권의 정확도를 보여주지 못하였다. 통계에 따르면 한국작업치료사면허시험의 5개년도 평균 합격률은 88%였는데(Korea Health Personnel Licensing Examination Institute, 2023b), 이는 3년 또는 4년의 교육을 받은 작업치료학과 졸업예정자 또는 졸업자들과 비교하면 ChatGPT-3.5의 능력이 현재까지는 부족하다는 것을 시사한다. ChatGPT는 자연어 처리에 높은 능력을 보이기 때문에 모두 객관식 문항으로 이루어진 한국작업치료사면허시험 문항에서 높은 능력이 기대되었다. 하지만 이런 기대와 달리 본 연구에서 확인한 ChatGPT의 낮은 정답률은 다음과 같은 이유로 설명될 수 있다. 첫째, ChatGPT의 자연어 학습이 영어나 영어권 문화를 기반으로 한 국가의 데이터를 기반으로 이루어졌다는 점을 고려했을 때, 문화에 영향을 받는 의료법규에 관한 문항이 포함된 1교시의 두 번째 단원에서 높은 정확도를 기대하기 어렵다는 점이다(Huh, 2023). 본 연구 결과, 1교시의 두 번째 단원이 다른 1교시의 첫 번째 단원과 2교시 문항의 정확도에 비해 5개년도 모두 낮은 경향을 보였다. 둘째, 선행연구에서 ChatGPT는 지식을 기반으로 한 언어모델이기 때문에 영어권 문화의 법률 지식에 관한 질문에서도 부정확한 답변을 제시하여 법률 지식 영역 내에서 ChatGPT-3.5 제한점을 시사하였다. 종합적으로 단순한 법 지식뿐만 아니라 법의 해석에 있어서 ChatGPT-3.5의 답변은 신뢰하기 어렵다는 결론을 내릴 수 있다(Tan et al., 2023).

본 연구 결과는 ChatGPT-3.5가 문화에 영향을 받지 않는 면허시험 문항들에서는 합격권의 정답률을 보였던 선행연구 결과와 대조적이다(Wang et al., 2023). ChatGPT뿐만 아니라 영어는 자연어 처리를 활용한 다양한 응용 분야에서 가장 풍부한 데이터를 제공하는 언어이므로 영어를 사용하는 문화권의 데이터가 학습의 핵심 원천이다(Névéol et al., 2018). 즉, 영어로 학습된 데이터를 처리할 때 ChatGPT는 높은 능력을 발휘할 수 있다. 실제로 본 연구에서 ChatGPT-3.5는 1교시의

첫 번째 단원(작업치료학 기초)과 2교시 문항(작업치료학)에서 비교적 높은 정확도를 보였는데, 이 문항들은 작업치료학 기초 및 전문 지식을 묻는 것으로 문화에 영향을 덜 받는다. 영어나 영어권 문화의 면허시험에 적용된 ChatGPT의 정확도는 평균 60% 이상으로 1교시의 두 번째 단원을 제외한 문항들의 정확도와 유사해 문화의 차이가 본 연구의 낮은 정답률 설명할 수 있다(Kung et al., 2023; Taira et al., 2023; Wang et al., 2023). 둘째, 한국작업치료사면허시험의 모든 문항은 가장 적절한 답을 선택하는 것을 요구하는 한편, 일부 문항에서는 정답이 아닌 선택지라도 임상적인 관점에서 적절한 것으로 간주할 수 있다(Huh, 2023). 실제 선행연구에서는 ChatGPT-3.5가 객관식 문항에서 최선의 답 대신 2개 이상의 답을 선택하는 때도 있었다(Huh, 2023). 이러한 결과는 현재까지 ChatGPT-3.5가 최선의 정답만을 선택하도록 최적화되지 않았음을 시사한다.

한편, 5개년도 전체 한국작업치료사면허시험에서 두 명의 사용자 간에 높은 일치도가 관찰되었는데, 이는 사용자와 관계없이 ChatGPT-3.5는 신뢰할 만한 결과를 제공할 수 있음을 보여준다. ChatGPT-3.5는 프롬프트 입력에 따라 다른 출력물을 생성할 수 있어서, 본 연구에서는 시험 문항과 함께 동일한 프롬프트를 사용하였다. 이는 통제된 프롬프트를 사용하면 신뢰할 수 있는 결과를 얻을 수 있다는 것이며, 선행연구에서도 통제된 프롬프트를 입력할 것을 제안하였다(Liu et al., 2020; Taira et al., 2023). 또한 사용자 간 일치도는 정확도와 양의 상관관계를 보였는데, 이는 선행연구 결과와 일치한다(Kung et al., 2023). 이러한 결과는 낮은 정확도를 보인 문항의 경우 ChatGPT-3.5가 잘못된 답변을 선택했다기보다는 학습 데이터의 부족으로 인해 일관된 정답을 선택하지 못하였다는 것을 보여준다.

한국작업치료사면허시험은 가장 적절한 1개의 정답 선택을 요구할 뿐 정답 선택에 대한 이유를 평가하지 않기 때문에 본 연구에서 ChatGPT-3.5의 정답 선택에 대한 이유는 평가하지 않았다. 따라서 정답 선택에 대한 이유까지 평가하기도 하는 타 직종 면허시험 결과와 달리 작업치료학을 전공하는 학생들의 교육을 지원하는 데 ChatGPT-3.5가 적절한지 파악하기는 제한적이다. 그럼에도 불구하고 1교시 두 번째 단원의 낮은 정확도를 고려한다면, 아직까지 ChatGPT-3.5가 교육 현장에서 사용되기에는 제한적이라는 것을 유추해볼 수 있다. 하지만 ChatGPT-3.5는 사용자의 피드백에 따라 빠르게 그 답변을 개선해나가며 동일한 질문을 사용하고 답변을 학습하면 그 이후에는 다른 결과를 보여줄 수 있는 대규모 언어 모델을 기반으로 하고 있기 때문에 현재 본 연구의 결과를 일반화하기는 어렵다(Taira et al., 2023). 따라서 향후 연구에서는 ChatGPT가 작업치료 관련 질문에 대한 답변 제시와 함께 설명을 제대로 작성할 수 있는지에 대한 검증을 하여 다른 매체를 추가적으로 사용하지 않아도 온전히 학습 도구로 사용할 수 있는지에 대한 확인도 필요할 것이다(Kung et al., 2023).

본 연구에는 몇 가지 제한점이 있다. 첫째, 3교시 문항은 공식적으로 접근할 수 없으므로 이에 대한 ChatGPT-3.5의 능력을 평가할 수 없었다. 결과적으로 2020년 시험의 경우에도 최종 합격 여부를 확인할 수가 없었다. 그럼에도 본 연구에서 사용한 ChatGPT-3.5의 경우 3교시에 문항에 포함되어 있는 그림이나 표를 해석하지 못하기 때문에 직접적인 평가는 제한적일 수 있다. 추후 그림이나 표 분석도 가능한 유료 버전인 ChatGPT-4를 사용하여 그 능력을 평가할 필요가 있다. 또한 ChatGPT-4는 ChatGPT-3.5에 비해 더 많은 양의 데이터를 학습하였기 때문에 3교시에 대한 정답률에도 차이를 보일 수 있다. 둘째, ChatGPT의 능력을 높이기 위해 원래의 문항과 더불어 최적의 답을 유도할 수 있는 비교적 간단한 프롬프트만을 사용하였다. 그러나 프롬프트를 최적화하기 위한 프롬프트 엔지니어링 기법을 적용하지 않았기 때문에 본 연구에서 ChatGPT의 능력을 낮게 보고하였을 가능성도 존재한다(Kasai et al., 2023). 반면, 본 연구에서는 실제 문항과 거의 차이가 없이 프롬프트를 구성하였기 때문에 높은 재현성을 가지고 있다는 장점이 있다. 또한 본 연구는 작업치료학과 학생들이 대중적으로 ChatGPT를 학습 도구로서 사

용될 수 있는지에 대한 검증이 ChatGPT 정답률 향상보다 더 중요한 목표로 삼고 있었기 때문에 프롬프트 엔지니어링을 최소한으로 하였다. 그럼에도 불구하고 추후 ChatGPT가 답을 확신하지 못할 때 오류 답변을 하지 않도록 프롬프트를 구성해야 할 필요는 있다.

V. 결 론

현재까지 한국작업치료사면허시험에 대한 ChatGPT 3.5의 답변은 한국의 작업치료를 전공하는 학생들에게 적용하는 데 제한이 있다. 그러나 ChatGPT의 성능은 계속해서 개선되고 있고 향후 프롬프트 엔지니어링과 함께 답변의 축적 등을 통해 본 연구에서 보고한 정답률보다 높은 정답률을 기대할 수 있다. 특히 가장 최신 버전인 ChatGPT-4는 더 많은 양의 데이터를 학습하였고 이에 따라 최신의 데이터를 포함할 수 있기 때문에 더 정확한 답변과 함께 이론적인 배경을 제시할 수 있어 ChatGPT-3.5보다 더 학습 도구로서 적합할 수 있다. 따라서 가까운 미래에 작업치료학과 교수와 작업치료학을 전공하는 학생들은 인공지능 기반의 대화 시스템의 잠재력을 주목하고 이를 활용한 교육 및 학습을 고려할 필요가 있다.

Conflicts of interest

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

Acknowledgements

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A3A2A02096338).

This work was supported by the Soonchunhyang University Research Fund.

References

American Occupational Therapy Association. (2020). Occupational therapy practice framework: Domain and process-Fourth edition. *American Journal of Occupational Therapy, 74*(Supplement_2), 7412410010p1-7412410010p87. https://doi.org/10.5014/ajot.2020.74S2001

Biswas, S. (2023). ChatGPT and the future of medical writing. *Radiology*, *307*(2), e223312. https://doi.org/10.1148/radiol.223312

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Hesse, C. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877-1901.

Castelvecchi, D. (2022). *Are ChatGPT and AlphaCode going to replace programmers?* Nature. https://www.nature.com/articles/d41586-022-04383-z

Hacker, P., Engel, A., & Mauer, M. (2023). *Regulating ChatGPT and other large generative AI models*. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1112-1123). Association for Computing Machinery. https://doi.org/10.1145/3593013.3594067

Huh, S. (2023). Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: A descriptive study. *Journal of Educational Evaluation for Health Professions*, *20*, 1. https://doi.org/10.3352/jeehp.2023.20.1

Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Stüber, A. T., Topalis, J., Weber, T., Wesp, P., Sabel, B. O., Ricke, J., & Ingrisch, M. (2023). ChatGPT makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *European Radiology*, Advance online publication. https://doi.org/10.1007/s00330-023-10213-1

Kasai, J., Kasai, Y., Sakaguchi, K., Yamada, Y., & Radev, D. (2023). Evaluating GPT-4 and ChatGPT on Japanese medical licensing examinations. *arXiv*. https://doi.org/10.48550/arXiv.2303.18027

Korea Health Personnel Licensing Examination Institute. (2023a). *Information on the national Korean occupational therapy licensing examination*. https://www.kuksiwon.or.kr/subcnt/c_2015/1/view.do?seq=7&itm_seq=13

Korea Health Personnel Licensing Examination Institute. (2023b). *Performance data*. https://www.kuksiwon.or.kr/peryearPass/list.do?seq=13&srchWord=13

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, *2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, *8*, 726-742. https://doi.org/10.1162/tacl_a_00343

Névéol, A., Dalianis, H., Velupillai, S., Savova, G., & Zweigenbaum, P. (2018). Clinical Natural Language Processing in languages other than English: Opportunities and challenges. *Journal of Biomedical Semantics*, *9*(1), 12. https://doi.org/10.1186/s13326-018-0179-8

Nisar, S., & Aslam, M. S. (2023). Is ChatGPT a good tool for T&CM students in studying pharmacology? *SSRN*. https://doi.org/10.2139/ssrn.4324310

Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, *11*(6), 887. https://doi.org/10.3390/healthcare11060887

Taira, K., Itaya, T., & Hanada, A. (2023). Performance of the large language model ChatGPT on the national nurse examinations in Japan: Evaluation study. *Journal of Medical Internet Research Nursing, 6*, e47305. https://doi.org/10.2196/47305

Tan, J., Westermann, H., & Benyekhlef, K. (2023). *ChatGPT as an artificial lawyer*. Artificial Intelligence for Access to Justice. https://ceur-ws.org/Vol-3435/short2.pdf

Wang, Y. M., Shen, H. W., & Chen, T. J. (2023). Performance of ChatGPT on the pharmacist licensing examination in Taiwan. *Journal of the Chinese Medical Association*, *86*(7), 653-658. https://doi.org/10.1097/JCMA.0000000000000942

Wang, W., & Siau, K. (2019). Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda. *Journal of Database Management*, *30*(1), 61-79. http://doi.org/10.4018/JDM.2019010104

Abstract

Can ChatGPT Pass the National Korean Occupational Therapy Licensure Examination?

Hong, Junhwa[*], Kim, Nayeon[*], Min, Hyemin[*], Yang, Hamin[*], Lee, Sihyun[*], Choi, Seojin[*], Park, Jin-Hyuck[**], Ph.D., O.T.

[*]Dept. of Occupational Therapy, Soonchunhyang University, Student

[**]Dept. of Occupational Therapy, College of Medical Science, Soonchunhyang University, Professor

Objective : This study assessed ChatGPT, an artificial intelligence system based on a large language model, for its ability to pass the National Korean Occupational Therapy Licensure Examination (NKOTLE).

Methods : Using NKOTLE questions from 2018 to 2022, provided by the Korea Health and Medical Personnel Examination Institute, this study employed English prompts to determine the accuracy of ChatGPT in providing correct answers. Two researchers independently conducted the entire process, and the average accuracy of both researchers was used to determine whether ChatGPT passed over the 5-year period. The degree of agreement between ChatGPT answers of the two researchers was assessed.

Results : ChatGPT passed the 2020 examination but failed to pass the other 4 years' examination. Specifically, its accuracy in questions related to medical regulations ranged from 25% to 57%, whereas its accuracy in other questions exceeded 60%. ChatGPT exhibited a strong agreement between researchers, except for medical regulation questions, and this agreement was significantly correlated with accuracy.

Conclusion : There are still limitations to the application of ChatGPT to answer questions influenced by language or culture. Future studies should explore its potential as an educational tool for students majoring in occupational therapy through optimized prompts and continuous learning from the data.

Keywords : Artificial intelligence, Large language model, Occupational therapy licensure examination