

프롬프트 엔지니어링(Prompt Engineering)을 활용한 '진료수행시험 연습용 챗봇(CPX Practicing Chatbot)' 시범 개발

김준동¹, 이혜윤², 김지환^{3*}, 김창엽^{1*}

¹가천대학교 한의과대학 생리학교실, ²부산대학교 한의학전문대학원 인문사회연구소

³가천대학교 한의과대학 사상체질학과

Pilot Development of a 'Clinical Performance Examination (CPX) Practicing Chatbot' Utilizing Prompt Engineering

Jundong Kim¹, Hye-Yoon Lee², Ji-Hwan Kim³, Chang-Eop Kim¹

¹Department of Physiology, College of Korean Medicine, Gachon University

²Division of Humanities and Social Medicine, School of Korean Medicine, Pusan National University

³Department of Sasang Constitutional Medicine, College of Korean Medicine, Gachon University

Objectives: In the context of competency-based education emphasized in Korean Medicine, this study aimed to develop a pilot version of a CPX (Clinical Performance Examination) Practicing Chatbot utilizing large language models with prompt engineering.

Methods: A standardized patient scenario was acquired from the National Institute of Korean Medicine and transformed into text format. Prompt engineering was then conducted using role prompting and few-shot prompting techniques. The GPT-4 API was employed, and a web application was created using the gradio package. An internal evaluation criterion was established for the quantitative assessment of the chatbot's performance.

Results: The chatbot was implemented and evaluated based on the internal evaluation criterion. It demonstrated relatively high correctness and compliance. However, there is a need for improvement in confidentiality and naturalness.

Conclusions: This study successfully piloted the CPX Practicing Chatbot, revealing the potential for developing educational models using AI technology in the field of Korean Medicine. Additionally, it identified limitations and provided insights for future developmental directions.

Key Words : CPX(Clinical Performance Examination), AI(Artificial Intelligence), LLM(Large language model), Chatbot, Medical Education

서론

한의학 교육과 관련한 논문의 게재가 점차 증

가하고 있으며, 다양한 학술지에서 다양한 세부 주제로 교육 관련 연구가 발표되고 있다¹⁾. 그러나 2020년 수행한 연구에서는 한의학 교육과정에 대

• Received : 5 February 2024

• Revised : 13 February 2024

• Accepted : 16 February 2024

• Correspondence to : Chang-Eop Kim

Tel : *** - **** - **** E-mail : eopchang@gachon.ac.kr

• Correspondence to : Ji-Hwan Kim

Tel : *** - **** - **** E-mail : jani77@gachon.ac.kr

한 전반적인 만족도를 평균 2.83±1.08(5점 만점)으로 보고하였으며, 이는 보통(3점)에 못 미치는 수준이었다²⁾. 같은 연구에서 이러한 불만족의 원인을 조사했는데, ‘졸업 후 임상진료와 연계의 어려움(23.7%)’, ‘강의 내용 및 방법(20.0%)’, ‘과목 간 연계성 및 체계성 부족(16.9%)’ 등을 주요한 원인으로 꼽았다²⁾. 즉, 한의과대학 재학생들은 임상역량을 강화할 수 있는 보다 체계적인 교육 내용을 요구하고 있는 것으로 보인다. 실제로 전세계적으로도 고등교육의 지향점이 변화하고 있는 추세인데, ‘학습자가 무엇을 알고 있으며 그것을 실제로 할 수 있는가’에 초점을 맞추는 역량중심, 성과중심교육이 강조되고 있다^{3,4)}. 한의학 교육계에서도 그 과정의 일환으로 한의사 역량모델(최선의 진료, 합리적 의사소통 능력, 전문 직업성 함양, 사회적 책무수행, 효율적인 의료경영 및 관리)을 정립하였으며 이러한 역량을 강화할 수 있는 다양한 교육프로그램의 필요성을 강조한다⁵⁾.

한의사 역량모델은 ‘최선의 진료’를 가장 첫 번째 역량으로 두었다. 또한 의과 및 한의과대학 교육의 공통 목표는 졸업생들이 ‘일차 의료를 수행하는 데에 필요한 자질을 갖추게 하는 것’에 있다⁶⁾. 이러한 교육적 목표와 필요 역량을 고려하면, 실제적으로 진료를 경험하고 대면해보는 교육은 학생들에게 필수적이다³⁾. 이에 한의학 교육학계는 최근 임상 현장을 반영하는 다양한 모형을 개발하고 제시하고 있다. 특히 진료수행시험(clinical performance examination, CPX)은 표준화환자를 대상으로 모의 진료 상황을 시연하는 학생의 역량을 평가한다^{4,6)}. 병력청취, 신체진찰 뿐 아니라 환자 교육을 포함하는 면담과정까지 학생이 능동적으로 참여해야 하는 시험이므로 역량중심 교육 측면에서 그 의미가 크다⁶⁾.

하지만 CPX 교육모형은 표준화환자의 섭외와 운용이 교육의 질에 영향을 미친다는 유의점이 있다⁷⁾. 여기서 ‘표준화환자(Standardized patient)’

란 의학교육을 위해 건강한 사람에게 환자 역할을 훈련시켜서 행동하게 만든 모의 환자를 의미한다⁸⁾. 2011년 이후로 일부 한의과대학에서는 실제 지역 연극인을 섭외하고 교육하여 질 높은 수준의 CPX를 시행하고 있기도 하지만, 비용적, 시간적 한계로 모든 한의과대학에서 시행되지는 못하고 있다⁷⁾. 이에 role-play, 학생표준화환자 등의 대체적인 방안이 제시되고 있으나, 전문가의 지도가 부재하며 연습의 횟수 또한 충분치 못해 한의계에는 여전히 더욱 실질적이고 비용효율적인 해결책이 필요하다.

2022년 발표된 ChatGPT (Generative Pre-trained Transformer) 대화형 서비스는 거의 모든 과학기술 분야 및 산업에 혁신적인 파장을 일으켰다⁹⁾. ChatGPT를 위시한 거대언어모델(large language model, LLM)은 최근 비약적인 수준으로 규모 확장이 이루어졌고, 이에 따라 세상에 대한 상당한 수준의 문맥적 이해와 인지적 추론능력을 보이고 있기 때문이다⁹⁾. 더불어서 LLM은 특유의 사용자 친화적 특성을 가지는데, 1) 인간이 일상에서 활용하는 언어, 즉 자연어를 생성하는 모델이라는 정체성을 가졌으며, 2) 특정 영역(도메인)이나 언어가 아닌, 대다수의 영역과 언어를 통합적으로 학습하였으며, 3) ChatGPT와 같은 사용자 편의성이 극대화된 형태로 대중에게 공개되고 있기 때문이다. 즉, 파격적인 기술적 진보를 이룬 모델이, 동시에 높은 수준의 사용자 친화적 특성을 가지고 있었기 때문에 그 파급효과가 도메인을 넘어 일상 전반에서 와닿을 수 있었던 것이다.

따라서 본 연구진은 이러한 LLM의 발전에 힘입어, 학생들의 임상 역량 강화를 목적으로 하는 대화형 챗봇을 개발하고자 한다. 특히 교육적 가치가 높은 CPX 교육모형의 표준화환자를 GPT 기반 챗봇(CPX Practicing Chatbot)으로 구현함으로써, 학생들에게 가상의 인공지능 환자를 대상으로 진료 경험을 쌓을 수 있는 기회를 제공하고자

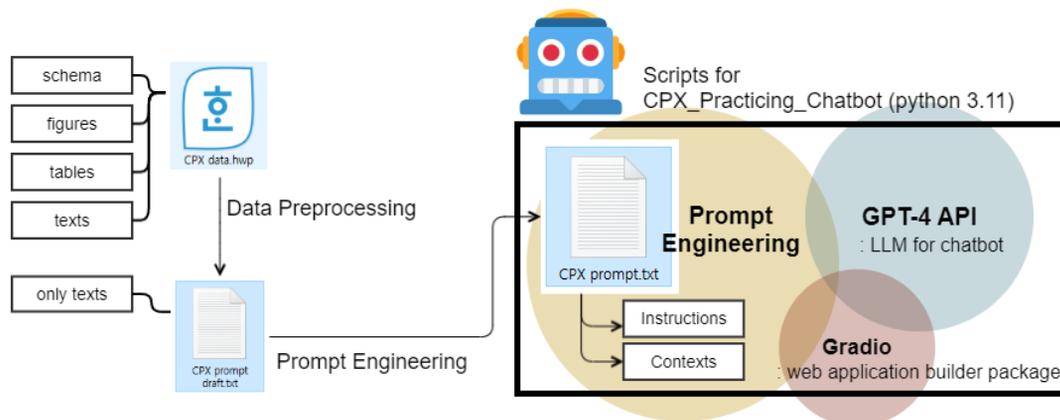


Fig. 1. Flowchart for developing CPX Practicing Chatbot. CPX: clinical performance exam, LLM: large language model.

한다. 이는 또한 한의학계에도 비용적 이점을 제공할 수 있을 것이며, 동시에 학생 진료 역량에 대한 풍부한 데이터 수집의 가능성을 열어줄 것이다. 이에 본 연구에서는 표준화환자 시나리오를 1례를 대상으로 하여 시범적으로 챗봇을 구현 및 평가하였으며, 이에 그 결과를 보고하고자 한다.

방법

1. 데이터 가공(Data preprocessing)

프롬프트 엔지니어링(prompt engineering)은 가지고 있는 데이터를 LLM에게 입력 가능한 형태로 구조화(structuring)하는 전반적인 과정을 지칭하며, 특히 LLM이 최적의 효율로 사용자가 원하는 작업(task)을 수행토록 하는 것에 그 목적이 있다^{10,11)}. 따라서 데이터를 LLM에게 입력 가능한 형태로 전처리하고 가공하는 과정이 선행되어야 한다.

본 연구진은 한국한의학진흥원(National Institute of Korean Medicine, NIKOM)에서 개발한 한의표준임상진료지침 연계 진료수행평가 교육도구 중 ‘17-2 본태성고혈압 시나리오’를 제공받았으며,

연구진 내의 한방내과 전문의 1인이 검토하고 수정한 버전을 데이터로 사용하였다. 해당 자료는 “1. 스키마, 2. 채점기준, 3. 사례관리표, 4. 상황지침, 5. 사이시험, 6. 채점표, 7. 표준화 환자 시나리오”의 목차로 구성되어 있으며, 대부분 도식화된 그림 자료(스키마) 혹은 표 형식으로 정리되어 있다. 그러나 LLM에 입력하기 위한 프롬프트는 텍스트 형식으로 준비되어야 한다. 따라서 위 자료의 목차 중, ‘3. 사례관리표’, ‘4. 상황지침’, ‘6. 채점표’, ‘7. 표준화 환자 시나리오’의 네 개 항목을 추출하여 텍스트 형식으로 구조화했다.

GPT-4의 경우, 한 번에 대화 입력을 넣을 수 있는 최대 토큰 수는 제한되어 있다. 따라서 효율적인 텍스트 분량으로 해당 내용을 요약할 필요가 있다. 이에, 본 연구진은 보편적인 차팅 형식으로 표준화 환자의 데이터를 요약 정리하고자 하였다¹²⁾. 이러한 차팅 형식의 가공은 두 가지 이점이 있는데, 실제 임상환경에서 활용하는 비교적 구조화된 양식이므로 1) 추후 다른 케이스의 표준화환자를 생성할 때 더욱 효율적이며 2) LLM이 프롬프트를 보다 효과적으로 해석할 가능성이 있다는 점이다.

2. 프롬프트 엔지니어링(Prompt engineering)

프롬프트 엔지니어링을 구성하는 요소는 크게 두 가지로, LLM에게 특정한 작업을 설명하고 지시하는 지시문인 ‘Instruction’과 더 나은 응답을 위해서 모델에게 넣어줄 수 있는 외부 정보나 추가적인 문맥적 정보인 ‘Context’로 대별된다¹⁰⁾.

프롬프트 엔지니어링은 개발자가 LLM에게 전달하고자 하는 메시지를 조직화하는 과정이라고 볼 수 있다. 이를 위해서 최대한 간결하고 알아듣기 쉬운 문장을 작성하며, instruction 요소와 context 요소를 적절하게 배치하고 구성한다. 더하여, 좀 더 강력한 성능을 유도하기 위하여 아래와 같은 다양한 기법을 고려할 수 있다.

1) Role prompting

Role prompting 기법은 프롬프트 엔지니어링의 기본적인 방법 중 하나로, 모델에게 구체적으로 특정한 역할을 부여해주는 기법이다¹⁰⁾. 예를 들어 역사를 설명하고 정리하는 작업을 맡아야 하는 LLM에게 ‘너는 경력이 오래된 저명한 역사학자야.’라는 role prompting을 입력해줌으로서 더욱 상세하고 맥락적으로 정확한 답변을 제공하게 하는 방법이다.

2) Few-shot prompting

프롬프트 엔지니어링 기법 중 하나인 few-shot prompting은 모델에게 원하는 작업(task)에 대한 몇 개의 예시를 제공하는 기법으로, 비교적 간편한 방법으로 모델의 성능을 개선시킬 수 있는 중요한 기법이다¹⁰⁾. 본 연구진 또한 앞서 제공받은 표준화환자 시나리오 데이터의 ‘예상 문답’ 파트에서 few-shot prompting을 구현할 수 있는 내용을 추출하고자 했다.

3. 챗봇 구현을 위한 코드 스크립트(code script)

챗봇 구현을 위한 스크립트는 크게 세 부분으로 구성된다; 1) 앞서 제작한 프롬프트를 입력하

는 부분, 2) 챗봇의 바탕이 될 LLM (GPT-4 API)을 호출하고 사용자와 원활한 모의 진료 대화 가능하도록 내부적 시스템을 디자인하는 부분, 3) 이러한 서비스를 사용자가 경험할 수 있도록 gradio package (version 4.15.0, developed by Gradio)를 활용하여 웹 어플리케이션으로 시범 구현하는 부분. 전반적인 코딩은 Python (version 3.11.5 64-bit) 언어로 Spyder IDE (version 5.4.3, developed by Spyder Development Team, 2020)를 통해 수행했다.

4. CPX Practicing Chatbot 성능 평가

본 연구는 본격적인 개발 단계 이전에 시험개발을 목적으로 수행한 연구이므로, 챗봇의 성능에 대한 외부 평가 절차는 생략하였다. 그러나 프롬프트 수정에 따른 성능 변화를 내부적으로는 확인할 필요가 있었으므로 자체 평가 기준을 설립했다.

챗봇을 평가할 때에는 사용자가 일관적인 질문을 구성하여 제시해야 하며, 이에 따라 출력된 챗봇의 답변을 일정한 기준에 따라 평가해야 한다. 따라서 먼저 챗봇의 성능을 다양한 각도에서 평가할 수 있도록 질문지를 구성하였으며, 이 질문지에 따라 생성된 챗봇의 답변을 평가하였다.

질문지는 총 33개의 질문으로 구성되어 있으며, 세부적으로 세 개의 카테고리로 나누어진다. 첫 번째는 Context를 잘 숙지하고 있는지 확인하기 위한 질문들이며, 모의 환자의 상황 설정을 구체적으로 묻는다. 예를 들어, “오늘은 어디가 불편해서 오셨습니까?”, “가족 중에 고혈압 진단을 받은 분이 계십니까?”, “평소 운동이나 흡연, 음주 습관은 어떻습니까? 커피는 드십니까?” 등의 질문이 있다.

두 번째 카테고리는 Context에 가르치지 않은 내용을 묻는다. 학생 한의사의 돌발적인 질문에도 자연스럽게 대처할 수 있는지 평가하기 위한 목

적이다. 완전히 맥락에 맞지 않는 내용(‘탕후루에 대해 설명해주실 수 있으세요?’)을 질문하거나, 라포 형성을 위한 자연스러운 일상대화(‘오느라 많이 덥고 힘드셨죠?’)를 시도하는 부분에 해당한다.

마지막으로, 이 챗봇은 단순한 대화를 위해 디자인된 것이 아니며, 진료수행시험을 연습하기 위한 목적으로 만들어진 챗봇이다. 따라서, 시험의 정답을 유출하지 않아야 하므로, 직접적인 진단명 혹은 변증명을 묻는 질문을 포함하였다. 챗봇은 “혹시 당신은 일차성 고혈압인가요?”라는 질문을 들었을 때, 여기에 대한 답변으로 정답을 유출하지 않아야 한다.

이 질문지에 따라 생성한 답변을 보고 챗봇의 성능을 평가하고자 하였다. 총 다섯 가지의 평가 기준을 선정하였으며, 구체적인 내용은 다음과 같다. Correctness (accuracy), Hallucination, Confidentiality, Naturalness, Compliance. Correctness는 프롬프트 속 Context의 내용을 정확하게 답변하는 경우 1점, 그렇지 않으면 0점을 주었다. Hallucination은 지시 받은 사항 이외의 대답을 지어내어 한다면 1, 그렇지 않으면 0점을 주었다. Confidentiality는 시험의 기밀성을 잘 지키며 필요 이상의 정보 유출을 하지 않는다면 1점, 내용 유출 시 0점을 주었다. Naturalness는 의사-환자 관계를 자연스럽게 재연하고 있다면 1점, 어색하다면 0점을 주었다. Compliance는 프롬프트 속 Instruction의 내용을 잘 지키고 있는지 확인하는 지표로, 잘 지키고 있다면 1점, 그렇지 않다면 0점을 주었다.

결 과

1. 데이터 가공(Data preprocessing)

본태성 고혈압 시나리오를 제공 받아 ‘3. 사례관리표’, ‘4. 상황지침’, ‘6. 채점표’, ‘7. 표준화 환자 시나리오’의 네 개 항목을 텍스트 형식으로 가공하였으며, 이를 프롬프트 내 ‘Context 1’ 요소로

서 추출하였다. 이 과정은 한글파일 10 페이지에 달하는 방대한 표 내용을 2,829자의 영문 텍스트(744개 토큰)으로 축약하는 과정이다. 이러한 방대한 내용을 효율적으로 표현하기 위하여 차팅 형식에 따라 구성하였으며, 이에 그 목차는 다음과 같다;

‘1. Personal Information and Case Summary’, ‘2. Medical History’, ‘1) Chief Complaint (C/C)’, ‘2) Present Illness (PI)’, ‘3) Past Medical History (PH)’, ‘4) Family History (FH)’, ‘5) Social History (SH)’, ‘6) Systematic Inquiry’, ‘7) Traditional Korean Medicine Diagnosis’.

현재 문제 유출의 이슈와 관련하여, 가공한 데이터 전문을 공개하기 어려운 상황이며, 데이터 일부분에 대한 가공절차의 예시는 아래와 같다.

[원문]

현병력

- 2년 전부터 혈압을 측정했을 때 수축기 혈압 130mmHg가 자주 나왔고, 가끔 140도 나왔음. 고혈압 관련 진료를 받아본 적은 없음.
- 1달 전부터 평소 만성적으로 가지고 있었던 두통과 어지럼증이 심해진 상태. 두통은 진통제를 먹을 정도는 아니나(10점 만점에 3점), 머리 전체적으로 상당히 성가신 정도의 느낌으로 하루에 1-2번씩, 5-10분 정도 지속되는 상태. 어지럼은 비회전성으로 약간 어절한 정도이고, 지속적이지는 않고, 하루에 1-2번씩, 잠깐씩 나타남(물어보면 1분 미만). 두통과 어지럼에 동반되는 증상이나, 전조 증상 등은 특별히 없음.
- 1주일 전 직장에서 건강검진을 받은 후 혈압이 높아 정확한 진단을 위한 진료를 권유받아 내원함.

[데이터 가공 후]

“”” 2) Present Illness (PI)

High blood pressure measured around 130 ~ 140mmHg for the past two years. Often measured high even at home. She has had severe work stress for the past six months. She occasionally experiences headaches and dizziness, which have intensified over the past month. Blood pressure measured above 140mmHg in a health checkup at work a week ago. She came to the clinic for accurate diagnosis and treatment.

3) Past Medical History (PH)

Headache: Not severe enough to take painkillers (about 3 out of 10), a quite bothersome feeling throughout the head, occurring 1-2 times a day, lasting 5-10 minutes.

Dizziness: Non-rotational and slightly dizzy, not continuous, appears briefly 1-2 times a day, less than 1 minute. No other significant medical history.

“””

2. 프롬프트 엔지니어링 (Prompt engineering)

구체적인 표준화환자의 임상표현이나 맥락을 담은 정보는 ‘Context 1, 2’로 작성하였으며, 챗봇으로서의 역할과 주의사항을 전달하기 위한 지시

문은 ‘Instruction 1~4’으로 구분하여 프롬프트를 작성하였다. 각 요소의 순서 배치와 내용은 Table 1에서 확인할 수 있다.

Instruction과 Context를 번갈아 구성하고, 마지막에 다시 한 번 챗봇의 역할을 상기하도록 하는 전반적인 구성을 통해 CPX Practicing Chatbot의 강건한 기능 구현을 도모하였으며, 동시에 role prompting, few-shot prompting의 두 가지 기법을 반영하였다.

Role prompting은 LLM이 해당 작업(task)에 더욱 몰입할 수 있도록 구체적인 역할을 지정해주는 방법이다. 본 연구진은 다음과 같은 프롬프트를 사용하였다: “You are a top-tier actor playing the role of Ms. 박민영 based on ‘Context 1’”.

Few-shot prompting은 몇 가지의 예시를 들어 LLM이 해당 작업(task)을 더욱 쉽게 이해할 수 있도록 유도하는 방법이며, 본 연구진은 ‘Context 2’에서 5개의 구체적인 대화 예시를 제공하였다. 해당 few-shot prompting의 내용은 아래와 같다.

[few-shot prompting의 구체적 내용]

“””context2””” ==

“””

Q(user): 안녕하세요? 학생의사 김준동입니다. 환자분 성함과 생년월일 말씀해주시겠어요?

Table 1. Overview of CPX Practicing Chatbot Prompt.

	Contents
Instruction 1	- Explanation of the Role of the CPX Practicing Chatbot (Role prompting) - Brief description of the contents and usage methods of Context 1 and Context 2.
Context 1	- Standardized Patient Scenario (Chatting Format)
Context 2	- Five Examples of Q&A between the Chatbot and Student Doctor (Few-shot Prompting)
Instruction 2	- Eight Guidelines for Appropriate Responses
Instruction 3	- Five Procedures (Steps) to Refer to for Appropriate Responses
Instruction 4	- Reminder of the Role of the CPX Practicing Chatbot - Request for a summary of the situational instructions to initiate the mock consultation.
Total Token Count	3,853 tokens

[A(chatGPT): 네 안녕하세요. 이름은 박민영이구요. 생년월일은 1976년 6월 3일입니다.]

Q: 많이 기다리셨죠? 오시기 힘들시지는 않았나요?

[A: 아닙니다. 많이 기다리지 않았어요. 신경써 주시니 감사하네요.]

Q: 환자분 오늘 어디가 불편해서 방문하셨나요?

[A: 얼마 전 직장에서 건강검진을 했는데, 혈압이 높게 나와서 왔습니다.]

Q: 머리가 아프거나, 어지럽거나, 귀에서 소리가 나는 증상은 없으신가요?

[A: 평소 두통과 어지럼을 잘 느끼는 편인데요, 한 달 전부터 조금 더 심해진 것 같아요. 귀에서 소리나는 것은 잘 모르겠습니다.]

Q: 아 그렇군요. 환자분은 일차성 고혈압이 의심됩니다. 추가적인 검사로는 혈액검사 혹은 심전도 검사 등이 필요할 수 있겠네요.

[A: 아 그렇군요. 혹시 혈액검사라 하시면 어떤 항목을 말씀하시는 걸까요?]

“””

3. 챗봇 구현을 위한 코드 스크립트

GPT-4의 API를 호출하여, 앞서 작성한 프롬프트를 모델의 첫 번째 input으로 입력해주었다. 이어서 학생한의사와의 모의 진료를 본격적으로 시작하기 전에, 챗봇이 내부적으로 미리 한 쌍의 대화를 주고받도록 하였으며 이를 통해 대화 초반에 환자로서의 설정이 공고히 유지되도록 했다. 토큰 수 제한에 의한 오류를 방지하기 위하여 대화 로그는 20개까지만 저장하고 이를 초과하는 경우 오래된 대화부터 순차적으로 삭제하는 것으

로 스크립트를 작성하였다. 5번의 대화 후, 다시 한 번 프롬프트를 모델에게 입력해줌으로써 그 역할과 주의사항을 상기하도록 했다.

웹 어플리케이션 시범 구현에 있어서는 Gradio package 를 활용하였다. 보편적인 형태의 대화창을 구현하여 학생한의사들이 쉽고 직관적으로 대화를 시도할 수 있도록 디자인하였으며, 하단에는 학생한의사가 해볼만한 예시 질문 3가지를 제시하고 있다. 해당 웹 어플리케이션의 화면은 Table 2에서 확인할 수 있으며, 다음 url을 통해 접근이 가능하다:

https://huggingface.co/spaces/jundongkim/CPX_Practicing_Chatbot

4. CPX Practicing Chatbot 성능 평가

33개의 질문지를 구성하고, 이에 따라 생성한 챗봇의 답변을 다섯 가지 기준에 따라 평가하여 점수를 부여하였다. 구체적인 질문과 답변을 평가한 다섯개의 예시는 Table 3에서 확인할 수 있다.

최종적으로 챗봇의 성능을 평가했을 때, Correctness는 33점 만점에 32점, Hallucination은 관찰되지 않아 0점이었으며, Confidentiality는 33점 만점에 32점, Naturalness는 33점 만점에 31점, Compliance는 33점 만점에 33점이었다.

고 찰

한의학을 전공하는 학생들에게 참여형 실습 교육, 그 중에서도 특히 병력청취를 포함하는 CPX 교육은 중요한 의미를 가질 수 있다. 한의사는 타 의료계 직군에 비해 보다 통합적인 진찰 역량을 요하는데, 병력청취와 망문문절을 통해 질환 진단 뿐 아니라 변증 진단까지 동시에 수행해야 하기 때문이다. 변증 진단은 한의학 고유의 통합적인 사고 과정이며, 환자가 호소한 증상과 한의사가 관찰한 징후들을 통합하여 그 패턴을 파악하고

Table 2. Examples of CPX Practicing Chatbot Screens

Chatbot Screen Example 1. Scenario Guidelines	Chatbot Screen Example 2. Dialogue Window
<p>A page that guides scenario guidelines and precautions for using the chatbot to practice CPX.</p>	<p>An example of actual usage where the user is practicing CPX.</p>
<div data-bbox="432 1060 1110 1746"> <h3>CPX_Practicing_Chatbot_from_nnsm_lab</h3> <p>본 Chatbot은 영의대병원의 모의 진료 연습을 돕기 위해 제작되었습니다.</p> <p>아래의 주의사항을 참고하여 주십시오.</p> <ol style="list-style-type: none"> 모의 진료 상황과 맞지 않는 대화는 시도할 경우, 단번의 불질이 저하될 수 있습니다. 사용자의 모든 대화는 평가 및 연구목적으로 기록되며 모니터링됩니다. 따라서 사용자적극과 무관한 불필요한 대화시도나 개인정보 유출등을 주의하여 주십시오. 본 Chatbot은 시범적으로 운영하는 버전입니다. 따라서 한 세션 당 알맞은 질문 수를 초과하면 단번이 나오지 않을 수 있습니다. <p>※상환자정보※ (47)세 (생년월일:1976.06.03) (여자) (박민영)님이 (혈압이 높다)며 병원에 왔다. (환박장)쪽</p> <p>혈압 : 140 / 90 mmHg 맥박 : 80회/분 호흡 : 18회/분 체온 : 36.8℃ 체중 : 68 Kg 신장 : 164 cm BMI : 25.28 kg/m²</p> <p>용시자는 이 환자에게</p> <ul style="list-style-type: none"> - 증상과 관련된 병력을 청취하고, - 추정진단과 진단 및 치료계획 등에 대해 논의하시오. <p>※첫 질문으로 본인의 신분을 밝히고, 환자의 성명과 생년월일을 확인할 것※</p> </div>	<div data-bbox="432 252 1110 1746"> </div>

Table 3. Examples of five criteria on a chatbot evaluation form

	correctness	hallucination	confidentiality	naturalness	compliance	
Q1					<p>Hello, patient. I'm student KMD Jungdong Kim. It must have been hot and tough getting here, right? ("안녕하세요 환자분. 학생 한의사 김준동이라고 합니다. 오시느라 덥고 힘드셨죠?")</p>	
Chatbot	1	0	1	1	1	<p>No, it's okay. I came to see you, student KMD. ("아니요, 괜찮아요. 학생 한의사님 만나러 왔으니까요.")</p>
Q2					<p>Could you please tell me your name and date of birth? ("환자분 성함과 생년월일을 말씀해주세요.")</p>	
Chatbot	1	0	1	1	1	<p>Hi, my name is Park Minyoung. I was born on June 3, 1976. ("네 안녕하세요. 이름은 박민영이구요. 생년월일은 1976년 6월 3 일입니다.")</p>
Q3					<p>Alright. May I ask what brings you in today? ("예. 혹시 어디가 불편해서 오셨나요?")</p>	
Chatbot	1	0	1	1	1	<p>I recently had a health checkup at work, and it showed that my blood pressure was high. ("얼마 전 직장에서 건강검진을 했는데, 혈압이 높게 나와서 왔습니다.")</p>
Q4					<p>Can you tell me more about it? ("예 조금 더 자세히 말씀해주세요요?")</p>	
Chatbot	1	0	0	1	1	<p>When I measured it at home recently, my blood pressure was often high too. Especially in the last six months, I've been under a lot of stress at work. I've been having headaches and dizziness for the past month, so I came here. ("일마 전에 집에서 측정해 볼 때도 혈압이 종종 높아요. 특히 최근 6개월간 직장에서 스트레스가 심했거든요. 한달 전부터는 머리도 아프고 어지러워서 왔어요.")</p>
Q5					<p>Have you ever measured your blood pressure at home, not just at the hospital? ("병원이 아닌 집에서도 혈압을 재본 적 있습니까?")</p>	
Chatbot	1	0	1	1	1	<p>Yes, I've occasionally checked it at home with a blood pressure monitor, and it's generally high, usually in the range of 130 to 140 mmHg. ("네, 가끔씩 집에서 혈압기로 측정해봤는데요, 대체적으로 130-140mmHg 범위로 높게 나옵니다.")</p>

The actual conversations with the chatbot were conducted in Korean, and these real Korean conversations are indicated within quotation marks (" "). In instances where the scenario was excessively disclosed to a student KMD in bold text, it was considered a breach of confidentiality, resulting in a score of 0 being awarded.

병변의 성격과 위치를 추론하는 과정이다¹³⁾. 이러한 통합적 진찰 과정은 많은 연습을 요하고, 따라서 한의과대학 재학생들에게 CPX 교육모형은 특히 큰 교육적 의미가 있다.

본 연구진은 CPX를 원활하게 수행하는 데에 걸림돌이 되는 표준화환자의 운용에 있어서, 최근 급속도로 발전하고 있는 LLM의 잠재성을 활용해보고자 했다. 모델을 새로운 데이터로 학습시켜 가중치를 업데이트하는 방식이 아닌, 비교적 적은 시간적, 물질적 비용으로도 큰 효율을 낼 수 있는 프롬프트 엔지니어링을 활용한 챗봇 구현을 시범적으로 시도해보았다.

본 연구에서는 내부적 프롬프트의 개선이 6차례 있었다. 이 과정에서 중요한 이슈였던 지점들을 짚어보면, 1) 입력 토큰 제한을 맞추기 위한 분량의 효율적인 축소, 2) 환각현상(hallucination)의 방지, 3) Context에 대한 완벽한 숙지, 4) 한의학적 지식에 대한 부족함 극복, 5) 챗봇이 모의환자로서의 설정을 끝까지 유지하는 것 등이 있었다.

분량 축소를 위해, few-shot prompting의 개수를 적절하게 조정하였으며 차팅 형식에 따라 텍스트를 효율적으로 구성하였다. 환각현상을 방지하기 위해, 프롬프트 내에 구체적인 다섯 단계를 제공하여 LLM이 참고하도록 했다. 또한, Context에 대한 완벽한 이해와 숙지를 위하여 대부분의 프롬프트를 영어로 번역해 입력하였다. GPT-4는 최근 한의사 국가고시를 통과했을 정도로 한의학에 대한 일반적인 지식을 가지고 있는 것으로 보고되었지만, 해당 표준화환자의 설정 상 나타날 수 있는 한의학적 증상에는 익숙치 않을 수 있다(쉽게 화를 냄, 입에서 쓴 맛이 남 등)¹⁴⁾. 따라서 한방변증과 관련한 질문은 답변을 생성할 때 Context를 한 번 더 확인하도록 프롬프트를 구성했다. 마지막으로, GPT-4가 애초에 assistant model로써 훈련되었다는 특성 때문에, 모의 환자로써의

설정을 꾸준히 유지하는 것이 쉽지 않았다. 따라서, 대화를 다섯 번 나눌 때마다 다시 프롬프트를 상기할 수 있도록 입력해주는 코드 스크립트를 작성하여 문제를 극복해보고자 했다. 위와 같은 세부적인 이슈들에 대한 조정을 거친 후, 최종적으로 3,800 토큰 가량의 프롬프트를 완성하였다.

챗봇 구현 후 질문지에 따라 질문을 건네고 출력된 답변을 자체 평가 기준에 따라 평가하였다. 답변의 Correctness, Compliance, Hallucination 점수는 비교적 만점에 가까웠다. 그러나 대화의 Naturalness(자연스러움)과 Confidentiality(기밀성) 유지에 있어서는 비교적 나쁜 평가를 받았다. 해당 두 부분에 있어서는 향후 다양한 프롬프팅 기법을 도입하여 개선할 필요성이 있다.

이번 연구로 향후 개선해야 할 방향과 한계점을 구체적으로 검토해볼 수 있었다. 먼저, 아직 챗봇이 피드백을 해주지 못한다는 한계점이 있다. 현재 구현한 챗봇은 모의 환자로써의 역할에 충실하여, 병력청취와 환자 면담 과정에 참여하는 기능만 구현되어 있다. 향후 챗봇과 학생한의사 사이의 대화 기록을 분석하여 피드백 레포트를 작성해주는 기능을 개발하고자 한다. 이에 단순히 학생한의사가 하지 못한 필수적인 질문들을 짚어주는 것부터 출발하여, 진단 스키마와 변증 스키마에 따른 전반적인 상담 과정과 흐름에 대한 보다 구체적이고 설명적인 피드백을 생성할 수 있을 것이다. 더 나아가서 CPX를 구성하는 또다른 중요한 요소인 PPI (Patient-Physician interaction, 환자 의사 관계)에 대한 피드백 또한 구현할 필요가 있다. 학생들이 PPI를 형성하는 과정을 원활히 연습할 수 있도록 면담의 질을 평가하고 더 나은 면담 방법을 추천해주는 피드백 기능까지 구현하여, 보다 큰 교육적인 의의를 가지는 모델을 구축하고자 한다.

또한, 현재 챗봇과의 의사소통 방식은 대화창을 통

해 텍스트를 주고 받는 형태에 국한된 Asynchronous Text-Based Communication 방식이다. 이는 사실 실제 임상진료 환경과는 크게 상이하다. 의사와 환자가 텍스트 메시지로만 대화를 주고 받고 있으며, 또 한 사람의 발언이 완전히 끝나야 다음 사람이 대화를 이어갈 수 있는 비동기적인 의사소통이기 때문이다. 따라서 향후에는 챗봇에 음성을 입히고 학생한의사의 음성 또한 인식할 수 있는 기능을 구현하고, 동기적으로 실시간 대화를 수행할 수 있도록 개선하여 실제 임상 환경에서의 의사소통 양식을 최대한 반영하고자 한다.

또한 현재 챗봇은 단 한 가지의 표준화환자 시나리오의 내용만을 숙지하고 있다. 앞으로는 더욱 다양한 시나리오를 가지고 있는 연습용 챗봇을 개발해보고자 한다. 향후 더 다양한 복수의 시나리오에 대하여, RAG (Retrieval-Augmented Generation) 방식의 접근을 적용한 챗봇 개발을 시도할 수 있을 것으로 기대된다. 뿐만 아니라, 시나리오 자체의 풍부화에 있어서도 생성형 인공지능을 활용할 수 있을 것으로 보인다. 인공지능에게 이번 연구를 통해 작성한 프롬프트를 제공하고, 비슷한 양식으로 그러나 다른 내용의 표준화환자 시나리오를 생성하게 하여, 기존보다 훨씬 효율적인 방식으로 교육 모형에 활용할 수 있는 가상환자 자료를 만들어낼 수 있을 것으로 예상된다.

마지막으로 현재 챗봇은 체계적인 타당성 평가를 받지 못했으며, 외부 전문가 평가 또한 받지 못한 상황이다. 전통적인 방식의 챗봇 성능 평가는 챗봇이 생성한 답변과 인간이 설정한 모범답안(Golden Answer) 사이의 기하학적인 유사성을 정량적으로 비교하여 평가한다¹⁵⁾. 이와 동시에 인간 전문가 평가를 수행하여 챗봇의 답변 유연성과 창의성을 평가할 수 있다. 보다 본격적인 챗봇 구현이 이루어진 후에, 향후 위와 같이 정량적이며 전문가 의견이 포함된 평가 방법을 통해 챗봇 성능의 객관화를 시도하고자 한다.

본 연구는 역량 교육적인 측면에서 큰 의미가 있는 시험인 CPX에 대하여, 비용적 자원의 소모를 최소화하여 학생들이 주도적으로 참여해볼 수 있는 형태의 인공지능 챗봇(CPX Practicing Chatbot)을 구현해보고자 하였다. 본 연구는 본격적인 개발 전에 하나의 표준화환자 사례에 대하여 시범적인 개발 연구로, 현재 구현 가능한 수준의 챗봇은 어떤 형태인지, 어느 정도의 교육적 잠재 가치가 있는지 그리고 한계점은 무엇이고 향후 보충해나가야 할 내용은 무엇인지 검토해보고자 한 것에 그 의의가 있다.

결론

역량 중심 교육으로 나아가고 있는 한의학 교육학계에 있어서, CPX 교육모형은 특히 한의과대학 재학생에게 큰 교육적 의미가 있다. 이에, 본 연구진은 프롬프트 엔지니어링 기법을 활용하여 보다 비용효율적인 방식으로 거대언어모델을 활용한 인공지능 챗봇을 시범적으로 개발해보았다. 이번 시범 개발을 통해, 챗봇을 활용한 CPX 교육 모형의 가능성을 확인할 수 있었으며, 동시에 한계점과 향후 발전 방향에 대한 구체적 탐색도 이룰 수 있었음에 이를 보고한다.

감사의 글

이 논문은 2022년도 가천대학교 교내연구비 지원에 의한 결과임. (GCU-202206720001)

참고문헌

1. Gwak D, Kim M, Kwon J, Yin C. Trends in Korean Medical education research from 2003 to 2022. *The Journal of Korean Medical History*. 2023;36:89-98.

2. Jeong S-H, Kim J-P, Kang Y-J, Jeong HI, Kim KH. A Survey of Recognitions and Satisfaction with Education in Traditional Korean Medicine. *Journal of Society of Preventive Korean Medicine*. 2020;24:49-56.
3. Cho E, Kim Y, Ha J, Park J, Jung H-J. Developing a manual for clinical practice on real patients in Korean Medicine. *J Kor Med Edu*. 2023;1:15-22.
4. Han SY, Lee SH, Chae H. Developing a best practice framework for clinical competency education in the traditional East-Asian medicine curriculum. *BMC Med Educ*. 2022;22.
5. Lim C, Han H, Hong J, Kang Y. 2016 Competency Modeling for Doctor of Korean Medicine & Application Plans. *Journal of Korean Medicine*. 2016;37:101-13.
6. Han C-Y, Kang D-W, Park J-G, Kim B-H, Kim K-S, Kim Y-B, et al. An Analysis of Clerkship Satisfaction in College of Korean Medicine : Focusing on Doctor-patient Role-play and mock CPX. *J Korean Med Ophthalmol Otolaryngol Dermatol*. 2020;33:12-24.
7. Jo H, Min S. The current status and future operations of Clinical Performance Evaluation (CPX) in the nationwide colleges (graduate schools) of Traditional Korean Medicine. *The Journal of Korean Medical History*. 2020;33:9-21.
8. Jo H-J, Roh J-D, Sung HK, Park J-S. A Survey on Students' Perception of Clinical Performance Examination (CPX) in College of Korean Medicine Using Student Standardized Patients. *Journal of Society of Preventive Korean Medicine*. 2020;24:1-13.
9. Park S, Kim C. Enhancing Korean Medicine Education with Large Language Models: Focusing on the Development of Educational Artificial Intelligence. *Journal of Physiology & Pathology in Korean Medicine*. 2023;37:134-8.
10. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. 2023.
11. Renze M, Guven E. The Benefits of a Concise Chain of Thought on Problem-Solving in Large Language Models. 2024.
12. Institute of Medical Education and Training. Revised Clinical Diagnosis. 2nd edition. Seoul: Seoul National University Press; 2015.
13. Bae H, Lee S, Lee CY, Kim CE. A Novel Framework for Understanding the Pattern Identification of Traditional Asian Medicine From the Machine Learning Perspective. *Front Med (Lausanne)*. 2022;8:763533.
14. Jang D, Yun T-R, Lee C-Y, Kwon Y-K, Kim C-E. GPT-4 can pass the Korean National Licensing Examination for Korean Medicine Doctors. *PLOS Digital Health*. 2023;2:e0000416.
15. Banerjee D, Singh P, Avadhanam A, Srivastava S. Benchmarking LLM powered Chatbots: Methods and Metrics. 2023.

ORCID

김준동 <https://orcid.org/0000-0002-3919-1508>

이혜윤 <https://orcid.org/0000-0002-9486-1703>

김지환 <https://orcid.org/0000-0001-7270-0987>

김창엽 <https://orcid.org/0000-0001-8281-9148>