PHQ-9 기반 우울증 자가 평가를 위한 LLM 기반 대화형 챗봇 시스템 설계 및 평가 연구

주세진 [1], 강미경 [1], 김도하 [1], S. Shyam Sundar[1,2], 한진영 [1*]
[1] 성균관대학교, [2] 펜실베니아 주립대학교

sjju01@g.skku.edu, gy77@g.skku.edu, doha.kim@g.skku.edu,
sss12@psu.edu, jinyounghan@skku.edu*

A Study on an LLM-based Conversational Chatbot for PHQ-9-based Depression Self-assessment

Saejin Ju[1], Migyeong Kang[1], Doha Kim[1], S. Shyam Sundar[1,2], Jinyoung Han[1*]

[1]Sungkyunkwan Univ., [2]Pennsylvania State Univ.

요 약

본 연구는 온라인 우울증 자가 평가 환경에서 대화형 챗봇에 대한 자가 평가 정확도 비교평가를 수행하여 자가 평가 정확도를 검증하는 것을 목표로 한다. 이를 위해 PHQ-9 기반의 대규모 언어모델 기반 온라인 우울장애 챗봇을 구축하고 기존 온라인 설문지 방식 간의 자가 평가 정확도를 실제 사용자를 대상으로 비교·평가하였다. 구축한 챗봇 시스템은 목적 지향 대화 시스템 구조를 적용하여 문항 간의 맥락적 연결성을 유지하여 자연스러운 질문을 사용자에게 제시하고, 사용자의 응답을 유도하도록 설계되었다. 실제 사용자를 대상으로 한 실험에서 본 연구에서 구축한 챗봇 기반 방식에서 산출된 PHQ-9 총점은 기존 PHQ-9 자가 평가 결과와 높은 일치율을 보이는 것으로 나타났으며, 이는 두 방식 간의 평가 정확도가 유사함을 의미한다. 본 연구의 결과는 LLM 기반 챗봇이 임상 환경에서 신뢰할 수 있는 보조 수단으로 활용될 수 있도록 하는 근거를 제시하며, 그 타당성과 실용성 확보에 기여한다

Ⅰ. 서 론

기존 대면 정신건강 서비스 이용 과정에서 발생하는 비용적 부담과 사회적 낙인문제로 인해, 최근 온라인 기반 정신건강 자가 평가 서비스의 활용이 점차 증대되고 있다. 이에 따라 PHQ-9(Nine-item Patient Health Questionnaire)이나 Beck 우울척도(Beck Depression Inventory)와 같은 자가 평가지를 활용하여, 사용자가 정신건강 기관을 직접 방문하지 않고도 자신의 상태를 간단히 점검할 수 있는 온라인 기반의 자가 평가 방식이 활발히 활용되고 있다[1].

그러나 기존의 전통적인 설문지 기반 자가 평가 방식은 획일적인 문항배열과 고정된 응답형식으로 인해 질문 간의 흐름이 단절되고 사용자의 심리적 맥락이나 감정의 뉘앙스를 충분히 반영하지 못한다는 한계가 지속적으로 지적되어 왔다. 또 낮은 몰입감과 비대화형 경험으로 인해 자기개방 수준이 저하되어 사용자의 심리상태를 정교하게 파악하는데 제약이 있다[2]. 이러한 한계를 보완하기 위해 최근 자연스러운 대화를 기반으로 한 대화형 인터페이스 챗봇이 주목받고 있다. 선행 연구에 따르면 챗봇 기반 접근은 사용자의 몰입감과 자기개방을 향상시키며, 전통적인 설문지 방식에 비해 보다 상호작용적인 경험을 제공할 수 있음이 확인되었다[3].

그러나 기존의 대화형 자가 평가 챗봇은 자가 평가 점수의 신뢰성과 정확성 측면에서 구조적 한계가 지적되어 왔다. 선행 연구들에 따르면, 이러한 시스템은 사용자의 자연어 응답을 기반으로 빈도나 강도 정보를 AI 모델이 추론해야 하므로, 언어 표현의 뉘앙스나 은유적 발화를 정확히 해석하지 못할 경우 점수 산출의 일관성이 저하될 가능성이 있다. 이는 AI 모델이 맥락 기반 의미를 충분히 파악하지 못할 때 심리 평가의 신뢰성에 영향을 미칠 수 있음을 시사한다[4]. 또한, 선행 연구들은 주로 영어권 환경에서 이뤄졌으며, 한국어 사용자 환경에 적합한 챗봇 연구는 부족한 실정이다. 문화적·언어적 맥락은 정신건강 평가에 중요한 영향을 미치므로 한국어 기반 챗봇 설계 연구의 필요성이 제기된다[5].

이에 본 연구는 우울증 자가 평가 척도 중 하나인 PHQ-9 을 기반으로 한국어 사용자 환경에 적합한 대화형 챗봇을 설계·구현하고 그 평가의 정확도를 정량적으로 검증하고자 한다. 이를 통해 한국어 기반 우울증 자가 평가 챗봇의 임상환경에서의 실질적 활용가능성과 기술적 개선 방향을 제시한다.

Ⅱ. 실험 설계

2.1 챗봇 시스템 설계

본 연구에서는 PHQ-9 기반 우울증 자가 평가환경에서 LLM 기반 대화형 챗봇의 자가 평가 정확도를 검증하기 위해 챗봇 시스템의 총점과 PHQ-9 온라인 설문지방식의 총점을 비교하는 실험을 설계했다.

2.1.1 설계 개요

본 연구의 PHQ-9 기반 우울증 자가 평가 챗봇은 TOD(Task-Oriented Dialogue)구조에 입각하여 설계되었다. TOD 구조는 특정 작업을 완수하기 위한 대화형 시스템에 효과적인 것으로 알려져 있으며 본 시스템은 이를 챗봇 설계 시 적용하였다. 챗봇은 크게 NLU(자연어 이해), DST(대화 상태 추적), DP(대화 정책 결정),

NLG(자연어 생성)의 4 개의 핵심 모듈로 구성되며 각 모듈은 독립적으로 동작하지만 유기적으로 연결되어 있다.

2.1.2 챗봇의 동작 과정 및 설계 세부내용

NLU 모듈에서는 사용자의 발화와 이전 대화 이력을 입력으로 하여 맥락 기반의 의도 11 가지(증상에 대한 서술, 빈도 표현, 강도 표현, 주제 이탈 등)를 분석하게 하였다.

DST 단계에서는 사용자의 발화로부터 PHQ-9 의 9 개 항목에 해당하는 증상 정보를 추출하고 슬롯 구조를 업데이트한다. 슬롯 구조는 각 PHQ-9 항목별로 항목명, 응답 상태, 점수, 원문 발화와 빈도나 강도표현 등을 포함한다. NLU 모듈에서 분석된 의도가 PHQ-9 관련 의도로 판단되면, 마지막 챗봇 발화, 사용자 메시지, 현재 슬롯 상태와 대화 히스토리를 종합적으로 GPT-5 모델에 입력하고 현재 사용자의 발화가 어떤 PHQ-9 항목과 관련되는지를 분석한다. 그리고 이 과정에서 추출된 증상 정보는 DB 에 실시간으로 저장된다.

DP 모듈에서는 현재 대화상태와 의도 분석 결과를 바탕으로 다음 대화를 이어가기 위한 대화정책을 결정한다. 챗봇은 16 가지 정책(도입 및 라포 형성, 증상 탐색, 대화 종료 등) 중 적절한 첫번째와 두번째 정책의 조합을 선택하는데, 이를 통해 본 챗봇은 사용자의 이전 발화와 의미적으로 연결되는 다음 질문 항목을 선택한다.

NLG 모듈에서는 선택된 정책에 따라 자연스러운 한국어 응답을 생성한다. DP 에서 정의된 정책별로 특화된 프롬프트가 사전 정의되어 있으며 2 개의 정책을 자연스럽게 조합하여 하나의 일관된 응답을 생성할 수 있게 한다. 위 일련의 과정은 모든 PHQ-9 항목에 대한 답변이 완료될 때까지 반복 수행된다.

2.2 실험 결과 및 성능 평가

본 실험은 **총 14 명**의 청소년·청년 대상으로 진행되었으며, 각각 PHQ-9 기반 우울증 자가 평가 챗봇, PHQ-9 온라인 설문지 두 가지 방식을 순차적으로 수행하였다. 두 방식 간의 집단 결과 차이를 비교하기 위해, 참가자별로 산출된 PHQ-9 총점을 병합하여 통계 분석을 하였다.

2.2.1 측정지표 및 분석방법

두 방식의 결과 유사도를 평가하기 위해 피어슨 상관계수를 주요 평가 지표로 사용하였다. 피어슨 상관계수는 두 연속형 변수 간의 선형 관계를 측정하는 통계적 지표로 본 연구에서는 챗봇 방식의 PHQ-9 총점과 기존 온라인 설문지 방식의 총점 간의 일치율을 정량적으로 확인하기 위해 선정되었다.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

수식 1. 피어슨 상관계수 정의

여기서 x_i 는 챗봇 방식에서 산출된 각 참가자의 총점, y_i 는 온라인 설문지 방식에서 산출된 각 참가자의 총점을 각각 변수로 사용하였다. 피어슨 상관계수 값은 1 에 가까울수록 두 변수 간의 강한 양의 상관관계를 의미하며, 통계적 유의성을 검증하기 위한 p-value 를 산출하였고 $p<0.05$ 일 때 통계적으로 유의한 상관관계가 있다고 판단하였다. 표 1 에 따르면 두 방식의 평균 점수는 각각 챗봇 방식 5.07 점, 설문지 방식 3.57 점으로 나타났으며 두 변수 간의 피어슨 상관계수 r 은 0.8644 (p=0.0001)로 통계적으로 매우 유의한 상관관계를 보였다. 이는 LLM 기반 대화형 챗봇이 임상적 도구로서 일정 수준의 자가 평가 정확도를 확보했음을 시사한다.

변수	평균(M)	표준편차(SD)	Γ	p
챗봇 총점	5.07	4.15		
설문지 총점	3.57	2.71	0.864	.0001

표 1 자가 평가 정확도 결과

한편, 일부 사례에서는 챗봇 점수가 설문지 점수보다 다소 높게 나타났는데, 이는 대화형 챗봇의 경우, 사용자의 감정 표현을 반영하여 응답을 유도하여 AI 가 이를 보다 높은 강도로 해석한 것으로 분석된다. 이는 대화형 설계가 자기 개방을 촉진하고 심리적 맥락을 세밀히 반영할 수 있음을 보여준다.

Ⅲ. 결론

본 논문에서는 PHQ-9 기반 우울증 자가 평가 환경에서 LLM 기반 대화형 챗봇과 기존 온라인 설문지 방식 간의 자가 평가 일치율을 비교하였다. 실험 결과, 두 방식의 총점 간 피어슨 상관계수가 0.8644 로 나타나 챗봇 기반 방식이 기존 설문지 방식과 높은 상관관계를 가지는 것을 확인했다. 이는 LLM 기반 챗봇이 기존의 온라인 설문지 방식과 유사한 수준의 평가 신뢰도를 확보할 수 있음을 시사한다. 향후 연구에서는 보다 다양한 연령대와 임상군을 포함한 대규모 검증을 통해 챗봇의 임상적 활용 가능성을 추가로 평가할 예정이다.

ACKNOWLEDGMENT

본 연구는 산업통상자원부 및 한국산업단지공단의 산업집적지 경쟁력강화사업(VCSK2502)의 연구결과로 수행되었으며 과학기술정보통신부 및 정보통신기획평가원의 디지털분야해외석학유치지원 연구결과로 수행되었음(RS-2024-00459638)

참 고 문 헌

[1] Kruzan, K. P., Meyerhoff, J., Nguyen, T., Mohr, D. C., Reddy, M., & Kornfield, R. (2022). "I wanted to see how bad it was": Online self-screening as a critical transition point among young adults with common mental health conditions. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22), Article 328, 1- 14. ACM. https://doi.org/10.1145/3491102.3501976

[2] Lee, E., Cho, Y., Kim, J., & Park, S. (2024). Digital interventions for reducing loneliness and depression in Korean college students: Mixed methods study. JMIR Formative Research, 8(1), e58791. https://doi.org/10.2196/58791

[3] Xiao, Z., Wang, T., & Wang, H. (2020). Tell me about yourself: Using an AI-powered chatbot to conduct conversational surveys with open-ended questions. ACM Transactions on Computer-Human Interaction, 27(3), Article 25. https://doi.org/10.1145/3381804

[4] Pichowicz, W., Kotas, M., & Piotrowski, P. (2025). Performance of mental health chatbot agents in detecting and managing suicidal ideation. Scientific Reports, 15, Article 31652. https://doi.org/10.1038/s41598-025-17242-4

[5] Yang, S. (2022). MICA: 한국 익명 심리건강 플랫폼 기반 심리상담 데이터셋. 한국과학기술정보연구원/Korea Science.