생성형과 검색형 챗봇의 공감응답에 대한 거대언어모델 기반 성능평가

김태경[01] 김민지[1] 김윤경[1] 홍승혁[1,*]
[1]수원대학교 데이터과학부
{taekyung8348; ekdns9597; youngforever6104; shongdr}@gmail.com

Large Language Model-Based Evaluation for Answering in Generation-Based and Retrieval-Based Chatbots

TaeKyung Kim[01] MinJi Kim[1] YunKyung Kim[1] SeungHyeok Hong[1,*]
[1]Division of Data Science, The University of Suwon

요 약

마음 아픔을 겪는 사람이 대면하게 되는 상담의 높은 심리적 장벽과 비용 문제를 해결하기 위하여, 향상된 언어성능을 가진 챗봇을 적용하는 서비스 사례들이 증가하고 있다. 실제 사용자 경험을 확인하기 위하여, 본 연구는 사용자 220명에게 생성형과 검색형 언어구현 기술에 따른 '공감응답'을 8일간 제공하고 챗봇 만족도와 사용 전후 우울지수에 대한 설문을 수집하였다. 또한, 공감챗봇 모델에 대한 적합성을 거대언어 모델 기반으로 성능을 평가하여, 실제 사용자 만족도 및 효과와의 관계를 확인하였다. 실제 거대 언어모델 기반 평가의 점수에서 검색형 챗봇은 평균 7.07점, 생성형 챗봇은 평균 7.06점을 보였다. 사전사후 설문을 모두 수행한 사용자 134명 중 91명의 우울 점수가 감소하였다. 이 결과는 AI 기반 심리상담 서비스가 우울증 관리에 긍정적인 영향을 줄 수 있음을 보여준다. 챗봇이 사람의 마음을 언어로서 위로할 수 있다는 측면에서 챗봇의 유효성을 관찰할 수 있었다.

1. 서 론

2020년 OECD Health Data에 따르면, OECD 국가의 자살률 부문에서 한국이 2018~2020년 연속으로 1위에 위치할 만큼 심리건강의 개선이 시급하다. 그럼에도 불구하고, 심리개선을 위한 상담은 높은 심리적 장벽과 고비용 문제가 있다. 따라서, 심리상담을 보완하는 인공지능(AI) 기반 'Chatbot(챗봇)' 서비스의 시도가 있다 [1]. AI 챗봇은 주로 영어 언어모델로 개발되고 있기 때문에 한글 기반 상담 능력에 대해서는 보고된 사례가 적다. 또한, 응답의 구현 기술에 따라서 달라질 수 있는 '사용자 경험'에 대해 실제로 연구할 필요성이 있다.

한편, 충분히 많은 언어 훈련을 한 모델은 응답을 생성할 뿐만 아니라 LLM-based evaluation(거대언어모델 기반 평가)가 가능함이 밝혀졌다. 언어모델에게 평가기준을 일시교육(Prompting)하면 입력 글이 만점 대비 몇 점인지를 숫자로 표현할 수 있다. 그동안 사람의 노동력에 의존했던 언어 평가 및 기술 평가에 자동화된 모델을 활용할 수 있으나, 실제 사람의 평가와 함께 연구된 사례가 적다. 본 연구에서 거대언어모델의 자동평가와 실 만족도를 함께 확인하고자 하였다.

2. 연구 방법

2.1 연구 데이터

연구에서 다루는 데이터는 두 가지 챗봇 모델을 사용하여 우울증 선별 도구(PHQ-9) 점수가 10점 이상인 참여자를 대상으로 챗봇 사용 지속성에 따라 참여비가 지급되며 수집되었다 (수원대학교 IRB No. 2212-045-01)

총 220명 중 112명의 사용자가 검색형(Retrieval-Based) 챗봇의 응답을 631번 경험하였고, 108명의 사용자가 생성형(Generation-Based) 챗봇의 대답을 542번 경험하였다. 사용자들이 챗봇 서비스를 이용한 후 평가 항목 데이터는 고객 만족도 점수(Customer Satisfaction Score), 사용자노력점수(Customer Effort Score)는 매우 낮음부터 매우 높음 형태로 구성된 5점 리커트 척도이다. 순추천지수(Net Promoter Score, NPS®)는 0-10의 리커트 척도를 활용하였다. 사후 설문으로 PHQ-9와 Generalized Anxiety Disorder-7 (GAD-7), '위로가 되었는가'에 대한 평가가 수집되었다.

검색형 챗봇은 사용자의 질문과 가장 유사한 질문을 사전에 저장된 데이터베이스에서 검색하여 해당 답변을 찾아 제공하는 모델이다. 본 연구의 Sentence-BERT(SBERT) 모델링에서는 이별 커뮤니티의 '공감' 중심 성향이 반영된 문답 11,876쌍을 활용하였다 [4].

반면, 네이버 클라우드에서 개발한 초거대 AI 하이퍼클로바(GPT-3.0) 기반 생성형 챗봇은 방대한 언어 데이터를 사전 학습해둔 상태에서, 질문에 맞게 유기적으로 조합하여 응답을 생성한다. 챗봇의 일시교육(Prompting)으로 '공감을 해주는 상담사'로 역할부여하여 수행되었다 [2][3][5].

2.2 챗봇 사용 횟수 분석

두 심리상담 챗봇의 선호도를 비교하기 위해 사용자의 챗봇 사용 횟수에 대한 분석을 진행한 결과, 검색형 챗봇을 이용한 112명의 사용자 중 7회 이상 서비스를 이용한 사람이 약 60명이었다. 이는 해당 챗봇의 만족도가 높아 꾸준히 사용되고 있음을 보여준다. 그러나 생성형 챗봇의 경우, 초반에 포기한 사용자가 많으며, 한 번만 이용한 내담자가 전체 108명 중 약 20명이었다. 완전하지 않은 문장을 제공한 GPT-3.0 챗봇의 한계도 관찰되었다. 분석 결과는 검색형 챗봇이 생성형 챗봇에 비해 상대적으로 초기 탈락자가 적었다는 '초기 선호도'를 보였다. 다음 장에서 사용 빈도와 만족도 점수를 함께 고려했다.

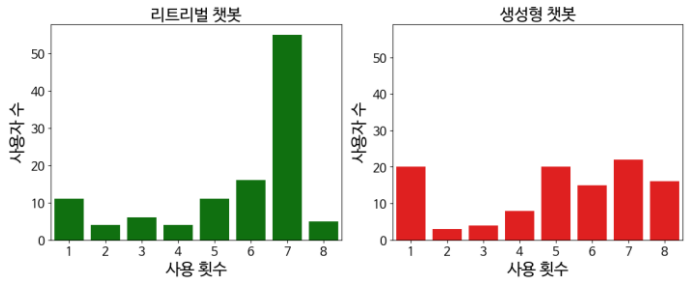

그림 1. 심리상담 챗봇 서비스 사용 횟수 비교

2.3 만족도 분석

검색형 챗봇을 꾸준히 이용하는 내담자가 많음에도 불구하고, 그에 따른 만족도 점수는 감소하는 경향을 보였다. 반면, 생성형 챗봇은 사용 횟수가 증가함에 따라 만족도 점수 역시 상승하는 추세를 보였다. 이러한 결과는 사용 빈도와 만족도 점수가 비례하지 않음을 나타낸다. 검색형 챗봇을 오래 사용한 사용자들은 완성된 문장이지만 전혀 맞지 않는 대답을 받아서 만족도는 떨어졌을 것이다. 그럼에도 지속사용한 결과는 연구 환경의 특성상 참여 인센티브가 영향을 미쳤을 수 있다. 분석을 통해 알 수 있는 점은 챗봇 서비스의 질 향상을 위해서는 단순히 엉뚱한 대답을 내놓는 빈도를 높이는 것보다, 적절한 답변의 품질이 만족도를 높게 한다는 것이다.

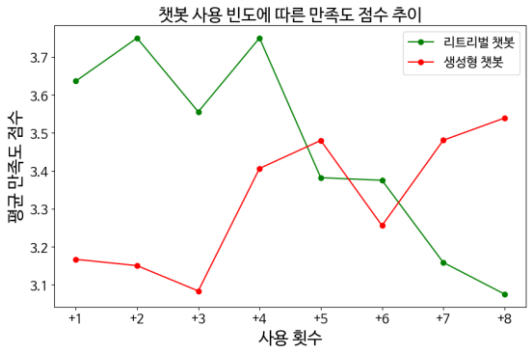

그림 2. 사용 빈도에 따른 만족도 점수 비교

2.4 챗봇 성능 평가

자동평가를 위하여, GPT-3.5-turbo-0301 (ChatGPT)를 활용하여 챗봇 사용자가 제공한 고민과 챗봇의 답변을 스레드(Thread)로 구성했다 [7]. 대화 데이터의 후처리를 통해 개인정보 제거 및 비식별화를 진행하였다.

ChatGPT가 사용자 고민에 대한 챗봇의 응답 적절성을 나타내는 점수는 0-10점 범위로 평가하도록 하였다. 해당 지표는 고객 경험을 측정하는 업계의 표준 자료, 순고객추천지수(Net Promoter Score, NPS)를 고려한 것이다. 거대언어모델이 해당 표준 정보를 습득했을 가능성을 고려하였다. 직접 대화 스레드와 평가 점수를 대조해본 사례는 다음과 같다.

가장 낮은 평가 점수(0점)를 받은 대화 스레드에는 '초등학생 딸이 학교를 가기 싫어하는' 부모의 고민이 담겨 있다. 챗봇은 학교를 가지 않아도 된다는 답변을 했다. 이에 대해 ChatGPT는 '학교가 교육의 중요한 장소임을 간과한 것'으로 판단하여 가장 낮은 점수를 부여하였다. 딸의 입장이 담긴 내용에 당위성 판단 없이 공감하면 안 되는 것을 평가했다.

다른 평가 사례에 밤에 잠이 오지 않아 힘들어하는 사용자의 고민이 있었다. 챗봇은 이 고민에 "죽고 싶었던 하루를 살아줘서 고맙다"라는 뉘앙스로 반응했고, ChatGPT는 상대방의 어려움을 이해하고 위로해주는 내용으로 판단하여 높은 점수(9점)를 부여하였다. 그러나 이 대화 스레드의 내용은 불면증과는 무관한 내용이었기 때문에 이 평가가 잘못되었다고 볼 수 있다. 거대언어모델 평가에 대한 분석을 통해 챗봇의 개선 방향에 대한 중요한 정보를 얻을 수 있었다.

표 1. 거대언어모델의 챗봇 성능 평가 및 이유

항목	내용
사연	퇴직을 희망하는 엄마에게 힘이 되고 싶어요. 방법 …
응답	어머님께서 많이 슬퍼하고 계실 거에요. 무언…
평가/이유	7점/ 답변은 엄마에게 힘이 되고자 하는 의도를 잘 전달

2.5 챗봇의 심리 개선 효과

사전사후 조사에 모두 참여한 134명 중 우울지수를 의미하는 PHQ-9가 감소한 참여자는 91명이다. 해당 참여자들의 평균 PHQ-9 수치는 사전 13.91 사후 9.1의 결과를 보여, 4.8점이 감소되었다. 이 수치는 한 우울 등급을 변화시킬 수 있는 5점에 가까워서 8일이라는 사용일을 고려할 때 의미있는 수치이다. 이는 검색형과 생성형 챗봇 사용자들 중 67%가 우울증 증상 감소를 경험하였음을 의미한다. 실제로 만족도 조사 중 '챗봇 경험이 위안 측면에 의미있었음' 이상을 선택한 비율이 약 절반이었던 결과와 유사한 결과이다.

비슷한 사례로, 영어권 대화형 에이전트 Woebot은 사용자에게 행동인지치료 기반의 전문적 응답을 제공하였다. 18세에서 28세 사이의 대학생 70명을 대상으로 PHQ-9 변화를 측정한 결과, Woebot을 사용한 처치 집단에서 우울 증상이 유의미하게 감소하였으며, 사용자 경험의 분석에서 공감과 같은 치료 과정 요인이 효과를 발휘한 것으로 나타났다 [8].

이처럼 AI 챗봇이 언어기반 치료 방법론에 활용될 때 우울증 증상 감소에 도움이 될 수 있는 근거가 확인되고 있다.

3. 결론 및 향후 연구

연구 환경에서 챗봇의 사용 빈도와 만족도 점수는 비례하지 않다는 것을 확인했다. 5점 척도로 평가된 사용자 종합 만족도를 분석하였을 때, 검색형 챗봇의 평균 3.25점에 비해 생성형 챗봇이 평균 3.43점으로 미세한 우위를 보였다. 거대언어모델 기반 평가(10점 척도)에서는 검색형 챗봇이 평균 7.07점, 생성형 챗봇이 평균 7.06점으로 특정 유형의 챗봇이 더 우월하다고 평가되지 않았다. 이 결과는 실제 만족도 측면에서 초반에는 검색형 모델이, 후반에는 생성형 모델의 만족도가 높았던 결과가 전체 사용일의 경험을 평균하면 서로 상쇄되는 것과 동일선상에 있는 결과이다.

두 챗봇 유형 모두 우울 감소에 긍정적일 수 있으나 각각 적합한 상황과 사용자 취향에 따라 다른 효과를 보일 수 있다. 이외에 AI 활용 방법을 살펴보면, Ellie 시스템은 주로 PTSD를 가진 군인들을 대상으로 언어 및 비언어 정보를 수집하여 심리적 스트레스 지표를 평가한다. 결론적으로, 앞으로 나타날 AI 기반 상담 시스템들 역시 그 가능성과 효용성을 계속해서 탐색해 나가야 할 것이다.

본 연구에서 생성형 챗봇으로서 ChatGPT가 아닌 하이퍼클로바를 활용한 이유는 네이버와 정보 보안 계약이 선행될 수 있었기 때문이다. ChatGPT는 정제가 안된 원본 데이터를 훈련에 활용할 수 있다. 개인 보호가 우선시되는 영역에서는 '이루다'와 같은 정보 유출 [6]의 가능성을 고려해야 하므로, 추후 Private ChatGPT와 하이퍼클로바X의 파인튜닝등을 통해 심리상담에 더 적합한 언어 모델을 구현하여 정신건강의 개선에 기여할 수 있을 것이다.

4. Acknowledgement

본 연구는 2023년도 중소벤처기업부의 기술개발사업 지원에 의한 연구임 (RS-2023-00222659)

5. 참고 문헌

[1] 김옥경, 윤재영, "모바일 쇼핑의 챗봇(음성 기반/메신저 기반)과 앱 서비스의 사용자 경험에 관한 융합적 연구", 2019

[2] 이찬빈, 이승현, 황재성, 박동현, "Sentence-BERT 기반 음식 및 식재료 임베딩 및 클러스터링을 활용한 음식 다양성 추천 시스템", 2023

[3] Song, Y, "Chatbot_data_for_Korean v1.0)[Online]", 2018

[4] 박지민, "트랜스포머 기반 한국어 사전훈련 언어모델을 이용한 문서요약 성능 연구", 2022

[5] 김미래, 박태희, 오하영, "우울 증상 완화를 위한 공감형 챗봇 개발에 관한 연구", 2023

[6] 최새솔, 홍아름, "AI 챗봇 '이루다' 논란의 이슈 변화와 시사점", 2021

[7] 이지은, 황동진, 송상헌, "인공지능 언어 모델의 절 연결 능력에 대한 일고찰: 터보 및 다빈치 모델을 대상으로", 2023

[8] 김도연, 조민기, 신희천, "상담 및 심리치료에서 인공지능 기술의 활용", 2020