AI 챗봇 서비스를 위한 데이터 보안 가이드라인

송현채[1], 이혜인[2], 이일구[3]
[1]성신여자대학교 미래융합기술공학과 석사과정
[2]성신여자대학교 CSE Lab 연구원
[3]성신여자대학교 융합보안공학과 교수
220236166@sungshin.ac.kr, lhynee@sungshin.ac.kr, iglee@sungshin.ac.kr

Data Security Guidelines for AI Chatbot Services

Hyun-Che Song[1], Hye-In Lee[2], Il-Gu Lee[3]
[1]Dept. of Future Convergence Technology Engineering, Sungshin Women's University
[2]CSE Lab, Sungshin Women's Universiry
[3]Dept. of Convergence Security Engineering, Sungshin Women's University

요 약

디지털 헬스케어 기술이 고도화되면서 디지털 치료제와 원격 의료 서비스가 의료 산업과 일상생활에 널리 활용되고 있다. 그러나 빅데이터 기반의 AI 서비스가 보편화될 수 있도록 데이터 수집, 가공, 활용 과정에서 개인정보가 남용되거나 유출되는 보안 위협도 증가하고 있다. 본 논문에서는 AI 챗봇을 활용한 정신건강 서비스를 위한 보안 위협 대응책을 마련하고 개인정보보호 가이드라인을 수립하여 사용자들에게 안전한 서비스를 제공하고 개인정보보호를 강화하는 AI 챗봇 서비스를 위한 데이터 보안 가이드라인을 제안한다.

1. 서론 및 배경

최근 IT 기술의 발달과 함께 병원과 공급자 중심으로 이어오던 헬스케어 서비스가 질병의 진단 및 치료에서 환자와 소비자 중심의 질병 관리 및 예방으로 변화하고 있다. 특히, 디지털 헬스케어 기술이 널리 활용되면서 때와 장소를 구애받지 않고 헬스케어 서비스가 제공되는 일상화된 서비스로 바뀌고 있다.

특히 코로나19로 인해 사회적 거리 두기가 강화되어 원격 의료에 관해 관심이 커지면서 헬스케어 디지털 기술을 활용한 헬스케어 산업이 급성장하고 있다. 그중 하나의 형태인 디지털 치료제 또한 국내외에서 활발하게 개발이 이루어지며 주목받고 있다.

디지털 치료제란 의약품과 같이 건강을 향상시키고 질병을 치료할 수 있는 소프트웨어이다. 일반적으로 모바일 앱, 게임, 가상현실(VR), 챗봇 형식 등의 의료기기로 활용되고 있고, 특히 행동 및 습관의 변화와 같이 꾸준한 관리가 필요한 질환의 환자들에게 유용하기 때문에 정신질환에서 활용할 수 있는 가치가 크다.

2. 정신 건강케어 서비스의 필요성

한국은 OECD 국가 중 자살률 1위이며, 보건복지부의 2021년 정신 건강실태조사 보고서에 따르면 우울장애와 불안장애를 포함한 정신장애를 평생 중 일회 이상 경험할 비율은 27.8%로 매우 높은 것으로 조사되었다[1]. 또한 건강보험공단 2019년 '건강보험제도 국민 인식 조사'에 따르면, 응답자 중 89.2%가 '건강관리가 중요하다'고 응답하였으며, 정신 건강관리에 대한 중요성도 이에 해당하는 것으로 나타났다[2].

그러나 그에 비해 정신 건강 서비스를 이용한 사용자 비율은 12.1%로 46.5%인 캐나다와 20.0%인 일본 등 다른 국가와 비교하면 정신 건강 예방 및 조기 개입 조치가 현저히 낮은 편이다[1].

정신 건강 서비스 이용률이 낮은 첫 번째 이유는 코로나19의 확산으로 관련 시설이 한정적으로 운영되거나 폐쇄되어 의료 서비스 접근이 단절되었기 때문이다. 두 번째 이유는 일반적으로 정신과의 진료 기간이 다른 질환에 비해 길고 주기적인 치료가 필요해서 비용적인 부담이 크기 때문이다.

따라서 기존의 대면 서비스로는 모든 환자의 효율적인 관리가 어렵다. 이를 해결하기 위해 디지털

기술을 활용한 정신 건강 서비스 등 대면 방식 이외의 서비스 제공이 필요하게 되었으며, 최근 몇 년간 진행된 의료 서비스 패러다임의 변화에 주목할 필요가 있다.

3. 디지털치료제와 AI 챗봇

디지털 치료제는 전통적 치료제와는 분명하게 구분되는 특징을 가지고 있다. 먼저 디지털 치료제는 디지털 기기를 통해서 제공되는 소프트웨어 프로그램으로 물리적인 제형이 다르고, 화학적 의약품과는 다르게 독성과 부작용이 거의 없다. 또한, 치료제 개발에 필요한 비용과 시간이 기존 치료제와 비교하여 매우 적고 소프트웨어 복제 비용이 저렴하므로 제품이나 서비스의 단가가 낮게 결정된다. 이러한 이유로 디지털 치료제는 치료 효과를 내면서 의료비용은 대폭 낮출 수 있다.

한편, 디지털 치료제의 주요 도입 분야는 만성질환과 신경정신과 질환이며, 이를 질환 치료제 중심으로 제품 개발이 이루어지고 있다.

미국에서 개발된 심리 인공지능 챗봇인 Tess와 대화한 75명의 대학생 대상 연구에 따르면 챗봇을 일정 기간 사용한 그룹의 불안감이 대조 그룹과 비교하여 유의미하게 감소했으며 이를 통해 챗봇이 인지 행동 요법에 효과적으로 사용될 수 있음을 보여준다[3].

더 나아가 사용자는 챗봇을 어떤 문제라도 논의할 수 있는 대상이라고 인식하고 있으며, 우울증, 불안, 트라우마와 같은 감정의 공유나 업무환경 스트레스에 대해 논의해야 하는 민감한 상황에서 사용자에게 신뢰와 편안함을 제공한다고 인식한다[4].

이러한 챗봇의 장점은 우울증을 비롯해 정신 건강 문제에 대한 사회적 편견이 만연한 지역에서 더욱 유익할 수 있다. Woebot, Wysa, Vivibot과 같은 인지행동치료, 행동 강화, 마음 챙김 등의 솔루션을 제공하는 정신 건강에 초점을 맞춘 상용 챗봇이 그간 출시되었으며, 최근 서울시를 비롯한 국내 지자체에서도 고령자 1인 가구의 외로움이나 고립감을 완화하는 용도로 대규모 언어 모델(Large Language Models; LLMs) 기반의 시니어 케어 서비스를 개시하고 있다.

4. AI 챗봇의 보안 취약점

챗봇의 긍정적인 잠재력에도 불구하고, 디지털 치료제로써 활용할 방안이 많은 AI 챗봇에도 보안 위협은 존재하며 우려의 목소리도 커지고 있다. 대화형 AI는 학습 데이터에 포함된 오류와 편향으로 인해 환각 현상으로 불리는 정확하지 않은 정보를 생성할 수 있으며 이는 허위 정보의 확산으로 이어질 수 있다. 만일 다급한 상황에서 건강관리나 의료 행위에 대한 사용자 질문에 챗봇이 오답을 준다면 이러한 LLM의 환각 현상은 생명을 다루는 분야에서 심각한 문제로 이어질 수 있다.

또 다른 주요한 보안 문제점들을 표1과 같이 정리할 수 있다.

<표 1> AI 챗봇의 주요 보안 위협

해킹 및 악성코드 생성	특별한 지식이 없는 사람들도 챗봇을 이용해 악성코드를 생성하고 해킹을 시도할 수 있다.
모델 악용	허위 정보를 사실처럼 답변하는 문제와 텍스트, 이미지, 음성 등을 생성할 수 있는 능력이 있으므로 악의적 의도를 가지고 허위 정보를 생성하거나 딥페이크를 생성할 수 있다.
사회 공학적 공격	챗봇을 악용하여 피해자의 기밀정보를 유출하는 등 사회 공학적 방법으로 공격할 수 있다.
디지털 도용	생성형 AI가 텍스트, 이미지, 음성 등을 생성할 때 실제 존재하는 인물 또는 조직의 신분을 도용하거나 명예를 훼손할 가능성이 있다.
데이터 유출	챗봇과 데이터를 공유하는 경우 무단 사용자가 데이터에 액세스하여 데이터 유출을 초래할 수 있다.

보안 위협 중 가장 문제가 되는 것은 개인정보 보호와 관련된 데이터를 누출하거나 남용, 오용하는 것이다.

생성형 AI를 활용하기 위해 제공되는 텍스트, 이미지, 음성 등의 데이터는 모두 서버에 저장되기 때문에 여기에 포함된 개인정보가 유출될 우려가 있다. 비식별 처리를 하더라도 AI 모델의 학습 결과와 외부 정보와의 결합을 통해 비식별화된 민감 정보를 추측할 수 있게 하여 심각한 보안 위협 상황을 초래하기도 한다.

더욱이 디지털 치료제로써 활용되는 AI 상담 챗봇일 경우 AI가 학습하는 데이터는 환자의 매우 민감한 정보가 되는데, 이러한 민감한 정보가 누출된다면 치료가 필요한 환자의 2차 피해가 발생할 수 있다.

5. AI 챗봇의 프라이버시 보호

AI 챗봇은 디지털 치료제뿐만 아니라 일상생활과 전 산업 분야에 널리 활용되고 있어서 학습 데이터

및 개인정보를 다루는 안전한 AI 챗봇 개발 가이드 규제책이 마련되어야 할 것이다. 또한, 품목 허가 단계에서도 일반 챗봇보다 강화된 보안 심사를 거쳐야 허가를 받을 수 있도록 해야 한다.

이러한 AI 챗봇 보안 위협을 완화할 수 있는 보안 고려 사항은 아래와 같이 제시할 수 있다[5].

<표 2> AI 챗봇의 주요 프라이버시 보호 고려 사항

입력 데이터 검증	- API 호출 시 입력되는 데이터에 대한 유효성 검사를 수행하여 SQL 주입, XSS 공격 등 위해 행위 방지 - API 호출 시 입력되는 데이터에 대한 기관 내부 정보 혹은 개인정보 포함 여부를 확인
개인정보 처리	- 「개인정보 보호법」등 개인정보의 처리에 관한 규정을 준수하여 데이터 처리 - 민감한 정보를 처리하기 전에 사용자의 동의 수집 - 데이터 가명화 및 익명화를 통해 데이터 누출 시 개인을 직접적으로 식별 또는 연관 불가토록 조치
데이터 오남용 방지	- API에 전달하는 사용자 입력 데이터 최소화 - API를 통해 반환되는 결과를 항상 검토하고, 안전성 및 정확성을 확인 - 불필요한 데이터 저장 및 공유 금지 - 필요한 최소한의 기간만 사용자 입력과 API 응답 결과를 저장하고, 보관 기간 이후에는 기관 및 API 양측 서버 간 데이터 완전 삭제

6. 결론

챗봇 개발 단계에서의 보안 가이드라인도 중요하지만 이미 제시된 고려 사항 중에서도 개인정보 처리에 있어서 체크리스트를 수립하고, 데이터 처리 모니터링 주기를 하루 단위로 설정하는 등 서버에 입력되는 민감 정보 모니터링 솔루션이 필요하다.

AI 챗봇의 보안 위협 대응책과 민감 정보에 대한 처리 방침이 필수적으로 마련되어야 정신 건강 분야의 디지털 치료제로써 심리 상담 AI 챗봇의 개발과 상용화가 가능해질 것이다.

ACKNOWLEDGEMENTS

본 논문은 2024년도 산업통상자원부 및 한국산업기술진흥원의 산업혁신인재성장지원사업(RS-2024-00415520)과 과학기술정보통신부 및 정보통신기획평가원의 ICT혁신인재4.0 사업의 연구결과로 수행되었음 (No. IITP-2022-RS-2022-00156310)

참고문헌

[1] 보건복지부. 2021년 정신건강실태조사보고서. 2021

[2] 국민건강보험 건강보험정책연구원. 2019년 건강보험제도 국민인식조사. 2019

[3] Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. JMIR Mental Health. 2024. 1 정보과학회지 25 5(4):e64 2018 Dec 13

[4] Ta, V. et al. User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. J. Med. Internet Res. 22, e16235. 2020.

[5] 국가정보원, 국가기술연구소. 챗GPT 등 생성형 AI 활용 보안 가이드라인. 2023.6