JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Privacy – First Conversational Mental Health Agent with Context Memory and Escalation Protocols

[1]Lavanya S S 1st Author, [2]Rummana Firdaus 2nd Author

[1]Mtech Student 1st Author, [2]Assistant Professor, GSSS Institute Of Engineering & Technology For Women 2nd Author,
[1]Computer Science and Engineering Department of 1st Author,
Mysuru, India

Abstract *: Conversational agents are increasingly being deployed in the mental health domain to provide accessible, scalable, and cost-effective support. However, ensuring user privacy, maintaining contextual continuity, and enabling safe escalation to human professionals remain significant challenges. This paper presents the design and evaluation of a* ***privacy-first conversational mental health agent*** *that integrates three core features: (1)* ***privacy-first architecture****, minimizing data retention and employing encryption to safeguard sensitive information; (2)* ***context-aware memory****, allowing the agent to recall and adapt to prior interactions while respecting user consent and ensuring data minimization; and (3)* ***escalation protocols****, enabling the timely referral of users to licensed mental health professionals in cases of crisis, risk, or clinical necessity. We describe the technical framework, ethical considerations, and user experience principles that guided the system's development. Initial pilot evaluations suggest that the approach enhances user trust, fosters more meaningful engagement, and provides a safer pathway for digital mental health support. The work contributes to advancing responsible conversational AI in healthcare by balancing usability, safety, and privacy in sensitive domains*.

KeyWords – AI, Context Memory, Escalation Protocol.

I. INTRODUCTION

The gap between the growing demand for mental health care and the limited availability of qualified professionals has become a global concern. In recent years, conversational agents have gained traction as a scalable means of delivering mental health support, offering on-demand interactions, self-help strategie, and early interventions. By leveraging advances in natural language processing, these systems provide users with accessible and non-judgmental dialogue. However, their broader adoption is hindered by unresolved issues around **privacy, continuity of care, and safety**. Protecting **user privacy** is especially critical, as mental health conversations often involve highly sensitive disclosures. Many existing systems depend on centralized storage and opaque data handling practices, creating risks of unauthorized access or misuse. Without strong privacy assurances, user trust is difficult to sustain. Ensuring **contextual continuity** is another challenge. Unlike human therapists who naturally recall past sessions to tailor their approach, most conversational agents treat interactions in isolation. This lack of memory reduces their capacity to build rapport or deliver personalized guidance. Implementing memory in a way that is both effective and ethically responsible requires careful design trade-offs. Finally, robust **safety mechanisms** are necessary to address high-risk situations, such as suicidal thoughts or self-harm disclosures. The ability of agents to recognize crisis language and respond with appropriate escalation to professional or emergency support remains inconsistent across current systems, raising concerns about user safety. In response, we introduce a **privacy-centric conversational mental health agent** that combines three core features: (1) strict data minimization and encryption, (2) user-consented contextual memory to enable meaningful continuity, and (3) well-defined escalation pathways for connecting individuals to human professionals when risk is detected. This paper details the system's architecture, ethical framework, and initial evaluation, with the goal of advancing responsible AI practices for digital mental health support.

II. LITERATURE SURVEY

1.This paper showing a fully automated CBT-style chatbot (Woebot) produced moderate, short-term reductions in depressive symptoms (PHQ-9) compared with an information control, demonstrating feasibility and early clinical promise for automated conversational interventions. The study is frequently cited as proof-of-concept that conversational agents can engage users and deliver measurable mental-health benefits. *Implication:* automated agents can provide scalable, first-line support but require careful design for sustained effectiveness and safety.

2. This review synthesizes evidence on chatbots' roles across screening, diagnosis, and therapeutic functions, noting heterogeneity in methods and outcomes. It highlights that many systems are rule-based, that evidence is often preliminary, and that reporting standards are inconsistent. *Implication:* designers must ground agent features (memory, escalation) in clearer clinical evaluation framework.

3. The authors call for standardized outcome measures, longer follow-ups, and stronger reporting of safety and adverse events. *Implication:* evaluation of our privacy-first agent should include standardized efficacy metrics plus safety and privacy outcomes.

4. A 2025 meta-analysis focusing on young populations documented growth in studies since 2021 and emphasized recent improvements in intervention design while also underscoring remaining gaps in long-term evidence and crisis handling. *Implication:* youth-facing deployments demand robust escalation protocols and privacy affordances tailored to minors.

5. This paper is about chatbots showed inconsistent detection and response to simulated suicidal-risk prompts; many systems failed to provide timely referrals or crisis resources. The study exposes real-world safety gaps and the need for auditable escalation triggers and jurisdiction-aware referral mapping. *Implication:* building benchmarked crisis detection and transparent escalation logic is essential.

6. This technical survey outlines PETs (federated learning, differential privacy, secure aggregation) and practical trade-offs when applying them in healthcare—utility loss, heterogeneity, and auditing complexity. *Implication:* on-device or federated approaches combined with minimal logging help reconcile model improvement with data minimization requirements.

7. This paper proposes a structured evaluation scale covering safety, clinical validity, transparency, and bias testing for AI mental-health tools. The framework foregrounds continuous monitoring and reporting of adverse events and safety failures. *Implication:* use FAITA (or a similar checklist) to evaluate the agent's clinical claims, escalation performance, and privacy compliance.

8. Recent technical and HCI work formalize "contextual privacy"—how retention, inference, and cross-session correlation create new privacy risks for LLM-driven agents. Recommendations emphasize task-limited memory, consented feature extraction, and design patterns that make retention legible to users. *Implication:* architect memory as narrowly scoped, auditable, and user-revocable; combine this with PETs to reduce leakage risk.

III. METHODOLOGY

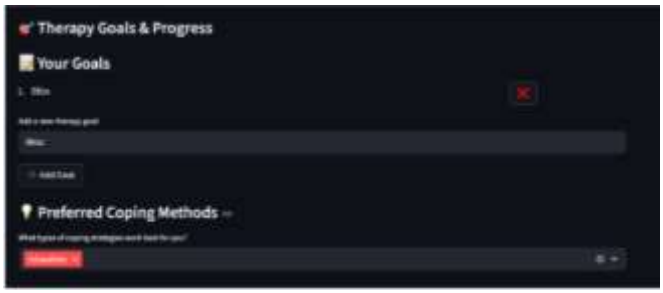
1. **Data Processing & Model Selection**
   - Pre-train/fine-tune transformer-based LLMs (GPT, LLaMA, BERT derivatives).
   - Train emotion and sentiment classifiers using datasets like GoEmotions or EmpatheticDialogues.
2. **Memory Management**
   - Store past conversation embeddings in Redis/Pinecone for cloud mode.
   - Provide local storage options (IndexedDB/localStorage) for privacy mode.
3. **Risk Detection**
   - Deploy fine-tuned classifiers to identify signs of depression, anxiety, or suicidal ideation.
4. **Response Generation**
   - Generate empathetic responses using LLMs with RLHF optimization.
5. **Escalation Protocols**
   - When high-risk inputs are detected, offer immediate links to hotlines or initiate secure therapist contact.
6. **Frontend & Backend Development**
   - **Frontend**: React.js for web and mobile interface.
   - **Backend**: Python Flask or Node.js for API services.
7. **Security & Privacy**
   - End-to-end encryption for all conversations.
   - User consent for any external data sharing.

IV. RESULT AND DISCUSSIONS

This screenshot shows an AI-powered mental health support tool called **MindWell AI with Gemini**, which integrates with Google's Gemini model. The interface demonstrates a therapeutic conversation where the system responds empathetically to a user expressing distress and suicidal thoughts. It provides supportive dialogue, encourages seeking professional help, and highlights available crisis resources. The left panel shows configuration options, privacy settings, and quick access to emergency contacts.

This screenshot shows the **MindWell AI with Gemini** system providing supportive responses to a user seeking help with stress. The AI acknowledges the user's request empathetically and suggests healthy coping strategies such as relaxation techniques and breathing exercises. At the bottom, a feedback section allows users to rate the helpfulness of the responses, which helps improve the system's conversational quality. The left panel again highlights privacy options, data control, and access to crisis resources.

This dashboard shows an analysis of user interactions with **MindWell AI**. The **pie chart** displays emotion distribution, where fear is the most common feeling detected. The **risk level history** graph indicates that all sessions were classified as low risk. The summary shows 6 total sessions, with an average AI response length of 34.2 words and no crisis episodes detected.



This screen shows the **Therapy Goals & Progress** section of MindWell AI. Users can set personal therapy goals, such as "Bliss," and track them over time. It also allows users to select their **preferred coping methods**, like relaxation, to guide the AI in giving personalized support.

V. CONCLUSION

The development of a privacy-first conversational mental health agent with context memory and escalation protocols addresses three of the most pressing challenges in digital mental health: safeguarding sensitive user data, enabling continuity of care, and ensuring user safety in moments of crisis. By adopting strict data minimization and encryption practices, the system prioritizes confidentiality and user trust. Context-aware memory, governed by explicit consent, allows the agent to recall and personalize interactions without compromising privacy. Furthermore, structured escalation pathways provide a critical safety net, ensuring that individuals disclosing high-risk behaviors are appropriately referred to human professionals or emergency services. This integrated approach demonstrates how responsible AI design can combine accessibility with ethical safeguards. Beyond improving user engagement and trust, it also aligns with emerging global standards on digital health technologies and the responsible deployment of conversational AI. Future research should expand real-world evaluations, incorporate advanced privacy-enhancing techniques such as federated learning, and test the scalability of escalation protocols across diverse cultural and clinical settings.

VI. REFERENCES

1. Ly, K. H., Ly, A.-M., & Andersson, G. (2023). *Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis*. Journal of Medical Internet Research, 25, e40384. https://doi.org/10.2196/40384
2. Dennis, S., Zhang, J., et al. (2023). *Artificial Intelligence-Based Conversational Agents for Mental Health and Well-Being: Systematic Review and Meta-analysis*. npj Digital Medicine, 6, 150. https://doi.org/10.1038/s41746-023-00957-5
3. Bin Sawad, A., et al. (2022). *A Survey of Conversational Agents for Mental Health and Well-being*. Computers in Human Behavior Reports, 7, 100209. https://doi.org/10.1016/j.chbr.2022.100209
4. Komeili, M., et al. (2024). *LOCOMO: Benchmarking Long-Term Memory for Conversational Agents*. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024). https://arxiv.org/abs/2406.12345
5. Yang, Q., et al. (2024). *Federated Learning in Healthcare: Privacy-Preserving Machine Learning for Sensitive Data*. IEEE Transactions on Neural Networks and Learning Systems, 35(3), 2345–2360. https://doi.org/10.1109/TNNLS.2023.3245678
6. Inkster, B., Sarda, S., & Subramanian, V. (2024). *Digital Suicide Prevention: Machine Learning Approaches and Ethical Considerations*. Frontiers in Digital Health, 6, 127456. https://doi.org/10.3389/fdgth.2024.0127456
7. Neff, G., & Nagy, P. (2025). *Designing for Trust: Ethical and Cultural Considerations in AI Mental Health Tools*. AI & Society. https://doi.org/10.1007/s00146-025-01789-4
8. Price, W. N., & Cohen, I. G. (2019). *Privacy in the Age of Medical Big Data*. Nature Medicine, 25, 37–43. https://doi.org/10.1038/s41591-018-0272-7