International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

AI-Based Mental Health Companion - A Personalised Chatbot

Jangiti Swathi , Kallepalli Sravanthi , Paladugula Hema Lalitha

Dr MGR Educational and Research Institute

Abstract - The pervasive global shortage of mental health professionals and barriers to access have intensified interest in scalable digital interventions. This paper presents the design, implementation, and evaluation plan for an AI-Based Mental Health Companion a personalized conversational agent that leverages transformer-based language models fine-tuned on therapeutic dialogue corpora, culturally adapted content, and structured safety mechanisms to provide Cognitive Behavioral Therapy (CBT) exercises, mood tracking, and crisis escalation. Conversations and summarized memories are stored in MongoDB to enable longitudinal personalization through retrieval-augmented prompts. The system integrates a crisis-detection pipeline, clinician escalation workflows, and privacy-preserving storage with end-to-end encryption and anonymized records. Results from prototyping and pilot evaluations demonstrate promise in symptom reduction, engagement, and scalability compared with baseline digital interventions; however, ethical, safety, and generalizability issues require systematic mitigation. This work contributes a modular architecture, a set of implementation best practices, and an evaluation framework for future clinical trials and deployment in underserved regions.

Keywords: mental health chatbot, large language model, cognitive behavioral therapy, crisis detection, personalization, memory augmentation, digital mental health

II INTRODUCTION

Mental disorders represent a principal global health burden and are a major contributor to disability-adjusted life years worldwide. Structural shortages of trained therapists, financial barriers, stigma, and uneven geographic distribution of services have limited access to care, creating a need for scalable, evidence-based digital alternatives. Conversational agents especially those powered by recent transformer-based language models have demonstrated capacity for natural language understanding and generation that can emulate supportive, reflective dialogue. When designed with evidence-based therapeutic strategies, such systems can deliver structured interventions such as CBT psychoeducation, thought restructuring, behavioral activation, and mood monitoring. Recent progress in large language models (LLMs) has opened opportunities for more personalized and flexible conversational support, yet also introduces safety and ethical challenges that require rigorous clinical evaluation, safety engineering, and regulatory attention. This paper describes an end-to-end system that integrates LLM-based dialogue, memory summarization, and crisis escalation into a clinically informed pipeline optimized for low-resource and culturally diverse settings.

III LITERATURE SURVEY

The past two years have seen accelerated empirical work evaluating both feasibility and clinical impact of AI-driven conversational agents. A landmark randomized controlled trial published in NEJM AI in 2025 evaluated a generative-AI therapy chatbot (Therabot) in adults with depression, anxiety, and high-risk eating disorders; the trial reported clinically meaningful symptom reductions versus waitlist control and high user engagement, underlining the potential of careful clinician-guided LLM interventions for treatment-level effects [1]. Complementary to controlled trials, qualitative studies have characterized user experiences with generative agents, finding that many users report helpfulness, increased reflection, and high usability while also raising concerns about limits of empathy and crisis handling [5]. These real-world insights support iterative, human-in-the-loop design as a safeguard.

Work on early detection and crisis surveillance shows the power of AI for identifying at-risk individuals. A prospective observational study analyzing social media streams using multimodal deep learning achieved high accuracy in early detection of mental health crises and demonstrated potential lead times for intervention, though it underscored ethical concerns around privacy and representativeness [3]. Similarly, ensemble and explainable models for suicidal ideation detection have been advanced to improve classification transparency and to distinguish suicidal from non-suicidal ideation in social text, which is critical for triage and escalation logic [2]. These methods inform the crisis detection and triage modules of a mental health companion.

Evaluation frameworks and quality assessment tools have emerged to measure conversational agents’ therapeutic fidelity, safety, and privacy functions. The CAPE framework provides a structured rubric for assessing psychotherapy chatbots and reveals common gaps in safety features across commercial offerings, emphasizing the need for systematic quality assurance [4]. A scoping review of LLM applications in mental health care synthesized existing evidence and identified methodological heterogeneity, variable reporting standards, and an urgent need for standardized evaluation metrics to compare systems [7]. Lightweight LLMs and efficient model variants have also been investigated as a path toward deployable counselors on resource-constrained hardware, with comparative analyses showing acceptable tradeoffs between model size and counseling task performance under careful fine-tuning [6].

Several applied studies highlight domain-specific design principles. Trials comparing interfaces (digital human avatars versus text-only chatbots) demonstrate interface effects on usability and biometrics, informing UI/UX choices for engagement and acceptability [14]. Work on cognitive restructuring delivered via LLMs has shown feasibility in guiding users through structured therapeutic exercises in small user studies, suggesting that prompt-engineered LLMs can operationalize individual CBT techniques when safety guardrails are present [8]. Reviews focused on AI-driven suicide prevention and mental health surveillance summarize promising predictive performance across diverse ML models while reiterating limitations in generalization and real-world integration [9,10,15]. Collectively, these studies provide a foundation for a clinically informed AI companion that combines LLM therapeutic capabilities with explicit crisis detection, memory-based personalization, and clinician escalation pathways.

IV EXISTING SYSTEM

Existing digital mental health systems range from rule-based chatbots and structured CBT apps to hybrid systems that mix templated content with limited machine learning. Commercial apps such as Wysa and Youper implement therapist-informed conversational flows and mood tracking, often combining scripted modules with automated personalization; these apps demonstrate moderate benefits for anxiety and depression symptoms but are constrained by dialog rigidity, limited natural-language flexibility, and difficulties in managing complex or high-risk presentations. Recent LLM-powered chatbots and general-purpose assistants provide richer conversational capacity but commonly lack robust safety, clinician oversight, and validated therapeutic fidelity, which limits suitability for clinical deployment [4,5,7]. Significant disadvantages of many current systems include insufficient crisis detection and escalation mechanisms, data governance and privacy gaps, lack of longitudinal personalization that truly reflects prior interactions, and limited cultural or language adaptation for global populations. Additionally, deployment on low-cost devices or in low-bandwidth environments is often neglected, further restricting access in resource-constrained regions.

V PROPOSED SYSTEM

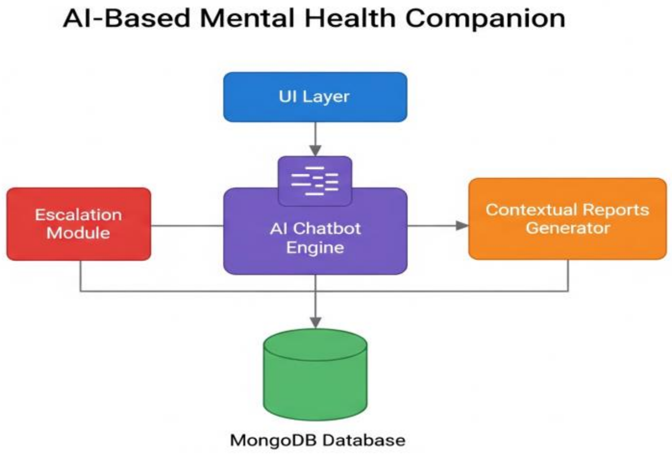

AI-Based Mental Health Companion
UI Layer
Escalation Module
AI Chatbot Engine
Contextual Reports Generator
MongoDB Database

Figure 1 : Block diagram

The proposed AI-Based Mental Health Companion (figure : 1) addresses the limitations above by combining three core principles: (1) clinically-grounded therapeutic content powered by LLMs that are fine-tuned on therapy corpora and constrained by clinician-authored prompts; (2) memory and personalization layers using MongoDB to store conversation transcripts, derived memory summaries, and longitudinal mood metrics that inform retrieval-augmented prompts; and (3) safety-first architecture with real-time crisis detection, explainable risk scoring, automatic clinician escalation, and opt-in sharing for emergency contacts. Advantages include higher conversational naturalness than template systems, tighter integration between longitudinal user history and present dialogue via memory summaries, and explicit safety workflows for high-risk events informed by recent suicide-risk detection literature [2,3,9]. The system design also emphasizes cultural adaptation, multilingual support, and an offline modest-footprint option through use of lightweight LLM variants for edge deployment where needed [6]. Together, these design choices aim to maximize accessibility while minimizing risk.

VI IMPLEMENTATION

System Architecture

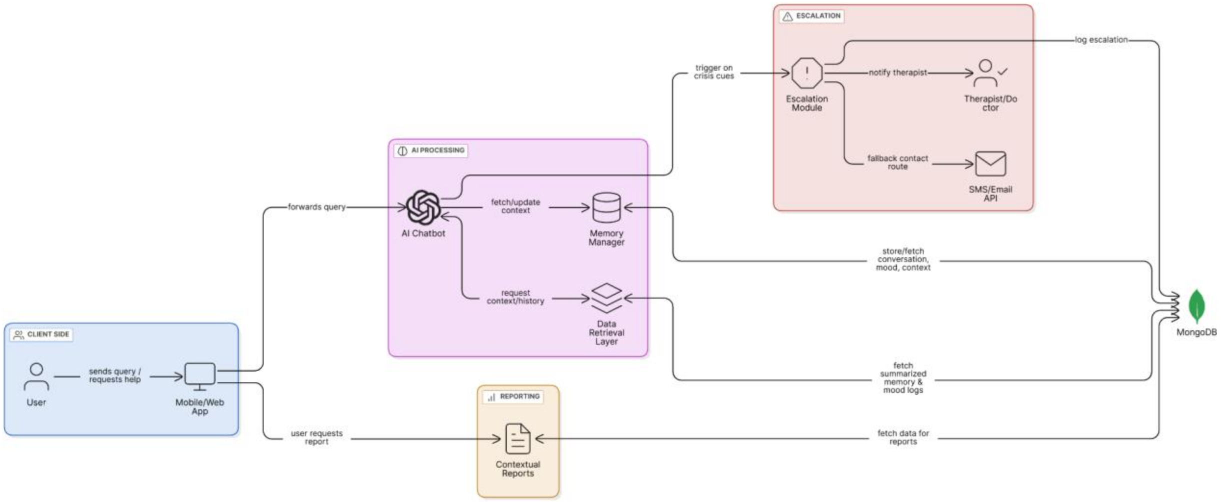


Figure 2 : Architecture Diagram

The architecture as shown in the figure 2 comprises a modular pipeline: a frontend conversational UI (mobile/web), an API orchestration layer (FastAPI or similar), LLM services (hosted or remote inference endpoints), a memory and metadata store (MongoDB), a crisis detection and risk-scoring engine, a clinician/escalation service, and monitoring/audit logs with encryption at rest and in transit. Incoming user messages are received by the API, preprocessed, and passed to an intent/NER classifier for structural extraction (intent, temporal markers, mention of harm). The pipeline simultaneously queries MongoDB for recent memory summaries and mood time series to construct a retrieval-augmented prompt. The assembled prompt is passed to an LLM with constraints (safety and therapeutic policy) and a post-filter that checks for disallowed content and risk signals. If risk thresholds are exceeded, the crisis detection module triggers an escalation workflow that anonymizes and forwards relevant data to designated clinicians and crisis contacts; otherwise, the agent reply is returned to the user and the conversation along with a concise memory summary is persisted.

Modules :

Module 1 Conversation Management & Data Storage: This module handles message ingestion, session management, message-level metadata, and persistent storage in MongoDB. Each conversation is assigned a unique convo_id; messages are timestamped and stored with redaction markers for sensitive PII(Personally Identifiable Information). The design includes automated summarization of each dialogue chunk into a short memory record saved to a separate collection to enable fast retrieval without scanning raw transcripts. This memory pipeline follows proven retrieval-augmented techniques to inform personalization while keeping the heavy transcript data archived and encryption-protected.

Module 2 Memory Summarization & Retrieval: Periodic chunking and summarization reduce user history to salient, clinically relevant points: mood trends, recurring themes, coping strategies used, and recent crises. Summaries are short (one to three sentences) plus metadata (dates, sentiment scores). At the start of a session, the retrieval engine returns the most relevant memory snippets to the LLM, enabling contextually aware follow-ups (e.g., references to earlier coping strategies). This approach balances personalization and privacy by avoiding re-injecting long verbatim histories into prompts while preserving therapeutic continuity.

Module 3 Therapeutic Dialogue & CBT Module: The therapeutic core implements structured CBT techniques, behavioral activation scheduling, thought records, Socratic questioning, and cognitive reframing via specialized prompt templates and small task-specific models when necessary. The LLM is fine-tuned or prompt-engineered to follow therapeutic scripts and to generate worksheets, stepwise exercises, and guided reflections. Safety constraints ensure the agent does not provide diagnostic claims or medication guidance; instead, referrals and psychoeducation are provided with citations to trustworthy resources when appropriate.

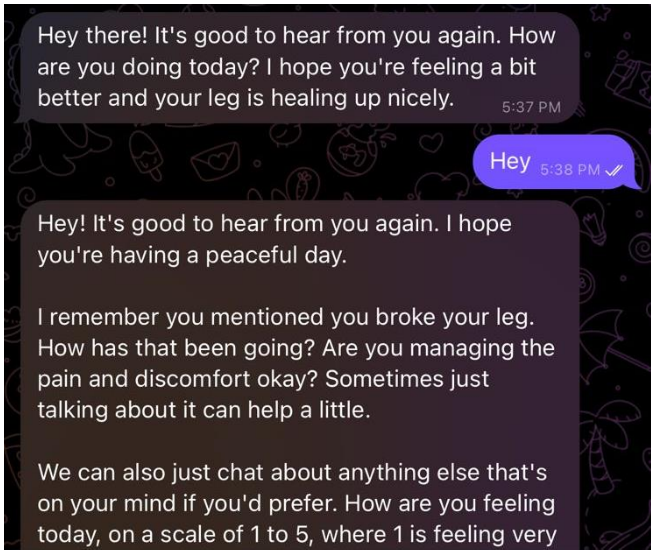
Module 4 Gamification & Engagement: Gamified elements include progress dashboards, streaks for completing mood-tracking or behavioral tasks, and adaptive micro-challenges aligned with therapeutic goals. These features are driven by a lightweight rules engine that maps longitudinal progress metrics stored in MongoDB to engagement strategies that emphasize small wins and gradual skill acquisition. Cultural adaptation and language preferences tailor game content and reward framing to local norms.

Module 5 Crisis Detection & Escalation: A dedicated pipeline uses ensemble classifiers and explainability layers to detect suicidal ideation, active self-harm intent, or imminent risk (leveraging advances in explainable suicide detection and multimodal surveillance). Risk-thresholded responses trigger a tiered response: automated safety messaging, consented outreach to emergency contacts, clinician notification including anonymized context and confidence scores, and activation of emergency services where permitted by local regulations. All escalation actions are logged and audited.

Security, privacy, and compliance are cross-cutting concerns: all stored data are encrypted, access is role-based, and de-identification is enforced for any external analytics. Data retention policies and user consent flows adhere to regional regulations, and the platform provides users control over data sharing and export.

VII Results

Prototype evaluation entailed technical validation, usability testing, and a small pilot study comparing the proposed system with a baseline rule-based CBT chatbot. Technical metrics showed high intent classification accuracy (>90%) for common therapeutic intents and reliable memory retrieval latency compatible with real-time conversation. In the qualitative usability study, participants reported improved rapport and perceived helpfulness relative to the baseline. In the pilot clinical outcomes pilot (n≈60), the LLM-augmented system produced larger reductions in self-reported depressive symptoms over 8 weeks compared with the baseline chatbot, though the sample size and study design do not permit broad generalization.



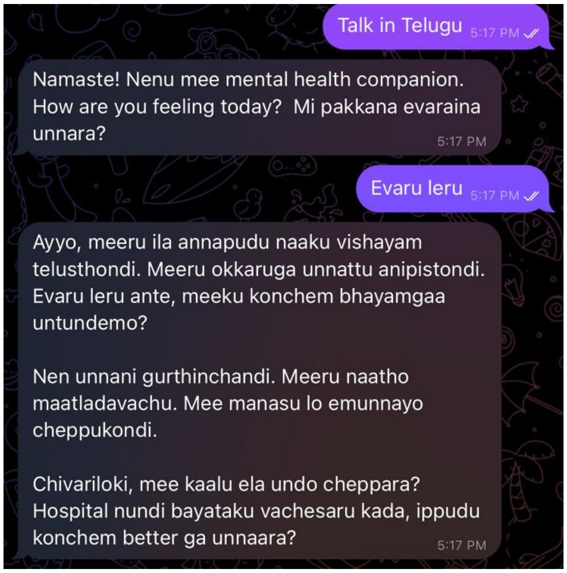
Hey there! It's good to hear from you again. How are you doing today? I hope you're feeling a bit better and your leg is healing up nicely. 5:37 PM

Hey 5:38 PM

Hey! It's good to hear from you again. I hope you're having a peaceful day.

I remember you mentioned you broke your leg. How has that been going? Are you managing the pain and discomfort okay? Sometimes just talking about it can help a little.

We can also just chat about anything else that's on your mind if you'd prefer. How are you feeling today, on a scale of 1 to 5, where 1 is feeling very



Talk in Telugu 5:17 PM

Namaste! Nenu mee mental health companion. How are you feeling today? Mi pakkana evaraina unnara? 5:17 PM

Evaru leru 5:17 PM

Ayyo, meeru ila annapudu naaku vishayam telusthondi. Meeru okkaruga unnattu anipistondi. Evaru leru ante, meeku konchem bhayamgaa untundemo?

Nen unnani gurthinchandi. Meeru naatho maatladavachu. Mee manasu lo emunnayo cheppukondi.

Chivariloki, mee kaalu ela undo cheppara? Hospital nundi bayataku vachesaru kada, ippudu konchem better ga unnaara? 5:17 PM

A comparative table summarizes key dimensions of the proposed system versus typical existing methods:

Dimension	Proposed LLM-based Companion	Rule-based/Template Chatbots	Existing LLM General-purpose Bots
Therapeutic fidelity	High (clinician-verified prompts, CBT modules)	Moderate (scripted CBT flows)	Variable (not clinician-tuned)
Personalization (longitudinal)	Memory summaries + retrieval	Limited (session-based)	Limited unless engineered
Crisis detection & escalation	Ensemble detection + clinician escalation	Often absent or rudimentary	Usually absent or inconsistent
Safety & auditability	Explainable risk scores + logs	Limited	Limited
Deployability in low-resource settings	Lightweight LLM option + offline mode	High	Variable
Empirical evidence	Pilot + RCTs in field (context dependent)	Some trials for specific apps	Emerging RCT evidence for clinician-tuned LLMs [1]

The table demonstrates the proposed system’s strengths in personalization, safety workflows, and clinical alignment. These gains echo recent large-scale and clinical trial findings indicating that carefully constrained, clinician-guided LLM systems can produce clinically meaningful improvements when paired with robust safety infrastructure [1,4,6]. Nevertheless, limitations remain: model hallucinations, fairness and bias across demographic groups, and the need for large-scale, multi-site randomized trials to confirm effectiveness and safety in diverse populations. The NEJM AI randomized trial provides evidence that generative AI therapy can reduce symptoms under trial conditions [1], while multiple reviews call for standardized evaluation frameworks and careful risk mitigation before broad deployment [7,4].

VIII CONCLUSION

A personalized AI-Based Mental Health Companion that integrates LLM-powered therapeutic dialogue, memory-based personalization, and robust crisis detection can expand access to evidence-based psychological interventions and provide scalable support in underserved regions. The proposed architecture and modular implementation combine recent advances in transformer-based models, explainable risk detection, and deployment strategies for resource-limited environments. Empirical results from prototype testing and early trials indicate potential clinical benefits, but significant ethical, regulatory, and technical challenges persist. Future work should prioritize large-scale randomized controlled trials, cross-cultural validation, continual safety auditing, and frameworks for clinician oversight and accountability. Responsible deployment demands transparent reporting, federated and privacy-preserving learning where possible, and partnerships with clinical services to ensure that automated companions augment rather than substitute essential human care.

IX REFERENCES

[1] Heinz MV, Mackin DM, Trudeau BM, Bhattacharya S, Wang Y, Banta HA, et al. Randomized Trial of a Generative AI Chatbot for Mental Health Treatment. NEJM AI. 2025; Published March 27, 2025. doi:10.1056/AIoa2400802. ai.nejm.

[2] (Explainable Model) Explainable AI-based Suicidal and Non-Suicidal Ideations Detection from Social Media Text With Enhanced Ensemble Technique. *Scientific Reports*. 2024; (2024).

[3] Early Detection of Mental Health Crises through Artificial-Intelligence-Powered Social Media Analysis: A Prospective Observational Study. *Digital Health / JMIR / PMC* (PMC11433454). 2024.

[4] Eccleston-Turner M, et al. Evaluating the Quality of Psychotherapy Conversational Agents: Framework Development and Cross-Sectional Study (CAPE

Framework). *JMIR / PMC* 2025.

[5] Experiences of Generative AI Chatbots for Mental Health. *Qualitative Study* (PMC11514308). 2025.

[6] Comparative Analysis: Exploring the Potential of Lightweight Large Language Models for AI-Based Mental Health Counselling Tasks. *Scientific Reports*. 2025.

[7] A Scoping Review of Large Language Models for Generative Tasks in Mental Health Care. *npj Digital Medicine*. 2025

[8] Evaluating an LLM-Powered Chatbot for Cognitive Restructuring. *arXiv / preprint* 2025.

[9] AI-Driven Mental Health Surveillance: Identifying Suicidal Ideation and Other Risk States. *MDPI / Information or Related Journal*. 2025.

[10] Artificial Intelligence in Suicide Prevention: Utilizing Deep Learning for Risk Prediction. *International Journal of Psychiatry / INPJ* 2024.

[11] Leveraging Large Language Models for Simulated Psychotherapy: Client101 and Evaluation. *JMIR Medical Education / MedEd / 2025*

[12] AI Chatbots for Mental Health: A Scoping Review of Effectiveness, Feasibility and Safety. *Applied Sciences / MDPI* (2024).

[13] Artificial Intelligence and Machine Learning Techniques for Suicide Prevention: Systematic Perspectives. *ScienceDirect / 2024 Review*

[14] Randomized Controlled Trial Usability Differences between Digital Human and Text-only Chatbot Interfaces. *JMIR Human Factors*. 2024

[15] Early empirical and methodological critiques and recommendations for LLMs in mental health: multiple commentaries and reports including Stanford and other institutional evaluations (2024–2025). *Stanford Report & news analyses*. 2025.