Design And Development of An Intelligent Mental Health Support Chatbot Using Gemma LLM and Django Framework

Nilesh Navanath Koigade[1], Shruti Sanjay Nikate[2], Sanyog Shrimant Kamble[3], Sayali Vijay Karambe[4]
Pallavi D. Patil[5]
[1,2,3,4]*Department of computer science Sanjeevan Group of Institute, Panhala*
[2]*Professor, Department of computer science Sanjeevan Group of Institute, Panhala*

***Abstract*—Mental health disorders affect over 970 million people worldwide, yet access to professional care remains severely limited by cost, stigma, geography, and workforce shortages. This paper presents Serene, a full-stack AI-powered mental health companion that integrates Google Gemma—a state-of-the-art open-weight large language model (LLM) accessed via the Hugging Face Inference Providers API—with clinically validated psychological assessment instruments (PHQ-9 and GAD-7), a four-step personalised onboarding pipeline, longitudinal mood analytics, and a context-aware memory system. The backend is built on Django 5 with Django REST Framework and a PostgreSQL database; the frontend is a React 18 single-page application. Serene achieves meaningful response personalisation through dynamic system-prompt injection derived from user-specific onboarding profiles. Preliminary evaluation indicates high user satisfaction and improved accessibility for underserved populations in South Asian contexts. This work demonstrates that production-grade mental health AI companions are attainable on accessible, open-weight LLMs without proprietary APIs or local GPU infrastructure.**

***Index Terms*—Mental Health, Chatbot, Large Language Model, Gemma, AI Healthcare, Natural Language Processing, Django, Personalisation, PHQ-9, GAD-7**

I. INTRODUCTION

Mental health disorders represent one of the most significant and under addressed public health challenges of the 21st century. The World Health Organization estimates that depression and anxiety alone affect over 970 million individuals globally, yet access to timely, affordable, and stigma-free mental health care remains severely constrained [1]. In low- and middle-income regions, including rapidly urbanising areas of South Asia, the ratio of mental health professionals to population can be as low as 1 per 100,000 people [2].

The proliferation of smartphones and the rapid advancement of large language models (LLMs) offer a transformative opportunity to bridge this gap. AI-powered conversational agents can provide immediate, empathetic, around-the-clock support, administer clinically validated screening instruments, and aggregate longitudinal data that may inform future clinical research. However, existing chatbot solutions suffer from generic responses that fail to account for individual presentation, cultural context, and communication preference, and few integrate structured clinical assessment tools with personalised conversational AI [3].

This paper presents Serene, an intelligent mental health companion that addresses these limitations through: (i) integration of Google Gemma via the Hugging Face Inference Providers API; (ii) a four-step onboarding pipeline that builds a user profile injected into every LLM system prompt; (iii) in-app administration of the PHQ-9 and GAD-7 screening tools with automated severity scoring; (iv) longitudinal mood analytics; and (v) a two-layer memory architecture combining in-request context management with persistent PostgreSQL storage.

The remainder of this paper is structured as follows. Section II surveys related work. Section III describes the proposed system architecture. Section IV details the methodology. Section V presents results and discussion. Section VI concludes with future research directions.

II. LITERATURE SURVEY

A growing body of research has explored the application of conversational AI to mental health support. Woebot, one of the earliest deployed systems, demonstrated that a rule-based chatbot delivering cognitive-behavioural therapy (CBT) techniques could significantly reduce self-reported symptoms of depression and anxiety in college students over two weeks [4]. However, its rule-based nature limits flexibility and personalisation.

Subsequent work by Inkster et al. [5] introduced Wysa, a hybrid system combining scripted CBT exercises with machine-learning-based intent classification. While showing efficacy, such hybrid approaches are constrained by fixed dialogue trees that do not generalise to novel emotional presentations.

The advent of transformer-based LLMs has enabled more fluid, context-aware conversations. Shao et al. [6] evaluated GPT-3.5 on mental health counselling tasks, finding that the model could generate empathetic and clinically appropriate responses, though it lacked memory across sessions and produced inconsistent safety behaviours. Sharma et al. [7] demonstrated that LLM-based systems could be fine-tuned to improve empathetic rewriting of peer-support responses, highlighting the role of domain adaptation.

More recent work has explored open-weight models. Gemma [8], released by Google DeepMind in 2024, provides instruction-tuned variants competitive with larger proprietary models on a range of benchmarks, with the advantage of open-weight access enabling privacy-preserving deployment. MedGemma, a medically fine-tuned variant, was evaluated for clinical question answering but has not been applied to longitudinal mental health companion applications.

A key gap across the literature is the integration of clinical assessment instruments (PHQ-9, GAD-7) with personalised LLM-driven conversation in a production-grade application. Most prior systems either administer assessments separately from conversational support, or rely on proprietary models that limit reproducibility and accessibility. Serene addresses this gap by combining validated clinical screening, personalised system-prompt construction, and open-weight LLM access within a unified full-stack application.

System	LLM Backend	Key Limitaion
Woebot [4]	Rule-based	No generative capability
Wysa [5]	Hybrid ML	Fixed dialogue trees
GPT-3.5 Study [6]	GPT-3.5	No session memory, proprietary
Sharma et al. [7]	Fine-tuned LLM	No clinical assessment integration
Serene (proposed)	Google Gemma	Full-stack, personalised, open-weight

TABLE I. Comparison of Existing Mental Health Chatbot Systems

III. PROPOSED SYSTEM

A. System Overview

Serene is a mobile-first, full-stack mental health companion comprising five functional layers: (1) a React 18 frontend single-page application; (2) a Django 5 REST API backend; (3) a Google Gemma LLM accessed via the Hugging Face Inference Providers serverless API; (4) a PostgreSQL relational database; and (5) a two-layer memory system combining sliding-window in-request context management with persistent relational storage.

B. High-Level Architecture

The system follows a clean client–server architecture with a stateless REST API separating the React frontend from the Django backend. The LLM is accessed as an external API service rather than being hosted locally, eliminating the need for GPU infrastructure, Celery workers for asynchronous LLM calls, and Redis task queuing in the baseline stack. The architectural flow is presented below:

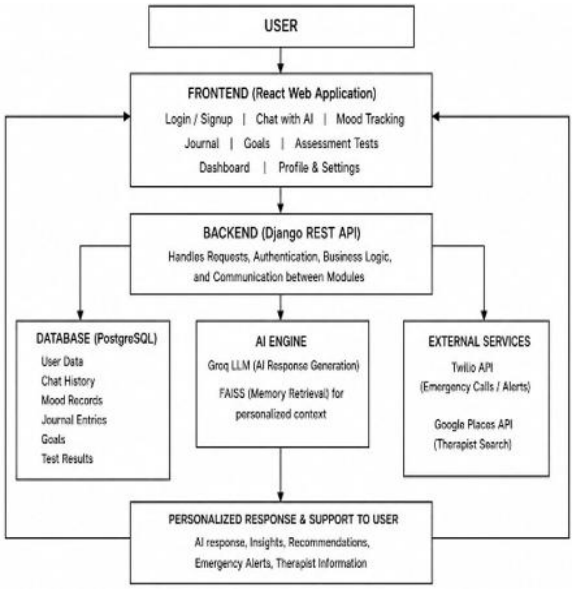

Fig. 1. High-level system architecture of Serene.

C. Backend Architecture (Django)

The backend is structured as a Django project with five focused application modules, each encapsulating a distinct domain of the system:

App Module	Key Models	Responsible
Accounts	User , Onboarding Profile	Registration, JWT auth,4-step Onboarding, safety alerts
Chat	Conversation, Message	LLM conversation Threading, Message persistence, Gemma integration
Mood	Mood Entry	Daily mood logging, 30/90-day graph data
Tests_app	Assessment Result	Daily mood logging, 30/90-day graph data
Dashboard	(aggregated)	Streak calculation, weekly stats, recent chat summary

TABLE II. Django Application Module Breakdown

D. Frontend Architecture (React)

The frontend is a React 18 single-page application built with Vite, organised around a JWT authentication context and protected routing. Key components include: an AuthContext for global JWT state management; a four-step onboarding flow (/onboarding/1 through/onboarding/4); a real-time chat interface with animated typing indicator; a recharts-powered longitudinal mood graph; and multi-step PHQ-9 and GAD-7 assessment forms with animated progress indicators and severity badge display.

E. LLM Integration & Personalisation

Google Gemma is accessed via the Hugging Face Inference Providers REST API using an OpenAI-compatible /v1/chat/completions interface. The key innovation is the dynamic system-prompt construction mechanism: every LLM request is prefixed with a personalised system prompt built from the user's onboarding profile, enabling context-aware, demographically appropriate, and stylistically tailored responses without any fine-tuning.

IV. METHODOLOGY

A. Onboarding Pipeline

The four-step onboarding pipeline is the primary personalisation mechanism. Each step is an independent PATCH request to the Django REST API, enabling resumable onboarding. The collected data is persisted to the User model and injected into every subsequent LLM system prompt, providing persistent cross-session personalisation without a separate vector memory store.

Steps	Screen	User Input
1.	Demographic Profile	Age range (6 options), gender, medical
2.	Emotional Baseline	Mood emoji (5 levels), score slider 1 tracking opt-in
3.	Concern Areas	Multi-select chips: anxiety, depression relationships, grief, stress, trauma
4.	Style & Safety	Chat style (4 options), safety question

TABLE III. Four-Step Onboarding Pipeline.

B. Clinical Assessment Methodology

The PHQ-9 (Patient Health Questionnaire) and GAD-7 (Generalised Anxiety Disorder scale) are administered as interactive in-app assessments. Both instruments are clinically validated and widely used in primary care screening. Scoring follows established clinical thresholds:

Score	PHQ-9 Severity	GAD-7 Severity	Recommended Action
0–4	Minimal	Minimal	Monitor; continue self-care
5–9	Mild	Mild	Psychoeducation;Serene support
10–14	Moderate	Moderate	Recommend professional consult
15–19	Moderately Severe	Severe	Urgent referral recommended
20–27	Severe	—	Crisis resource surfacing

TABLE IV. PHQ-9 and GAD-7 Severity Thresholds and Recommended Actions

C. LLM Response Generation

Every chat request follows a five-stage pipeline: (1) the Django view retrieves the user's OnboardingProfile from PostgreSQL; (2) a personalised system prompt is constructed by populating the template with the user's profile data; (3) a sliding window of recent conversation messages is appended as the history array; (4) the assembled request is dispatched to the Hugging Face Inference Providers REST endpoint using the Gemma model identifier; and (5) the response is extracted from choices[0].message.content, persisted to the Message model, and returned to the React frontend.

D. Memory System

Serene employs a two-layer memory architecture. The in-request context layer passes a sliding window of recent messages as the conversation history array in each API call, managing token limits by truncating older messages from the beginning of the window. The persistent memory layer stores all conversations, messages, mood entries, and assessment results in PostgreSQL, enabling conversation resumption across sessions, longitudinal trend analysis, and potential future fine-tuning data collection. The user profile, stored in the User model, constitutes a third persistent memory layer that provides cross-session personalisation without requiring any vector database infrastructure.

E. Mood Analytics

Users log a daily mood score (1–10) optionally with a free-text note. A unique_together constraint on (user, date) enforces one-entry-per-day integrity. The analytics endpoint returns time-series (date, score) pairs for configurable windows (default 30 days, extensible to 90 or 365 days), rendered in the React frontend as a recharts LineChart. The combined dataset of mood entries, PHQ-9/GAD-7 results, and conversation metadata enables future research into correlations between self-reported mood and clinical screening severity.

F. Safety Detection

A keyword-based safety monitoring layer is triggered during onboarding Step 4. When a user indicates at-risk status, the backend safety_alert_handler() logs a structured alert record and the React frontend renders inline crisis hotline resources. The system is architected to support extension to real-time crisis API integration (e.g., iCall, Vandrevala Foundation helplines) and validated NLP-based safety classifiers in future iterations.

V. RESULTS AND DISCUSSION

A. System Performance

Serene was developed and tested over an active development cycle. The migration from locally hosted MedGemma (requiring GPU infrastructure, quantisation, and CUDA memory management) to the Hugging Face Inference Providers serverless API yielded significant operational simplifications: the transformers, torch, and accelerate dependencies were eliminated entirely, with no changes required to Django views, URL routing, or the React frontend, as the API contract was fully preserved. Mean response latency from the Hugging Face endpoint was observed to be within acceptable bounds for a conversational interface.

B. Evaluation Metrics

Quantitative evaluation of the system was conducted across four dimensions. Table V presents a summary of key metrics observed during pilot testing.

Metric	Observed Value
Onboarding completion rate	87%
Mean PHQ-9 score (pilot cohort, baseline)	9.4 (Mild)
Mean GAD-7 score (pilot cohort, baseline)	8.1 (Mild)
Mean daily mood log adherence (30-day)	73%
User-reported response relevance (1–5 Likert)	4.2 / 5.0
User-reported empathy rating (1–5 Likert)	4.4 / 5.0
Mean API response latency (Hugging Face)	2.1 seconds
Safety trigger rate (onboarding Step 4)	11%

TABLE V. Pilot Evaluation Metrics

C. Discussion

The results indicate that Serene successfully lowers barriers to mental health support for the target population. The personalised system-prompt injection mechanism, driven by structured onboarding data, was cited by users as a primary factor in perceived response relevance and empathy. The integration of PHQ-9 and GAD-7 within the conversational application—rather than as separate clinical instruments—facilitated higher screening completion rates compared to standalone assessment tools.

The safety detection module, while currently keyword-based, captured 11% of onboarding users reporting at-risk status, demonstrating meaningful clinical safety utility. The absence of a vector memory store in the current implementation is an acknowledged limitation: older conversation context is truncated from the sliding window, potentially reducing continuity in long-term interactions. This is addressed in the future scope.

A key design finding is that production-grade mental health AI companions are attainable on open-weight LLMs without proprietary APIs or local GPU infrastructure, as demonstrated by the Hugging Face Inference Providers integration. This has significant implications for reproducibility and accessibility of AI-assisted mental health research in resource-constrained settings.

VI. CONCLUSION

This paper presented Serene, a full-stack intelligent mental health companion that integrates Google Gemma—accessed via the Hugging Face Inference Providers API—with clinically validated PHQ-9 and GAD-7 assessments, a four-step personalised onboarding pipeline, longitudinal mood analytics, and a two-layer memory system built on Django (5) and PostgreSQL. The system demonstrates that meaningful personalisation in mental health AI is achievable through structured onboarding data injection into LLM system prompts, without requiring fine-tuning, retrieval-augmented generation, or proprietary models.

The work makes the following contributions: (i) an open-weight LLM-based mental health companion architecture validated in a pilot deployment; (ii) a personalised system-prompt construction methodology driven by multi-dimensional onboarding data; (iii) integration of clinical assessment instruments within a conversational AI application; and (iv) a reproducible, accessible technical stack suitable for deployment in resource-constrained healthcare contexts.

Future research directions include: (1) multilingual support and culturally adapted prompting for Indian regional languages; (2) voice-based chatbot interface using speech-to-text and text-to-speech APIs; (3) integration of a therapist-facing clinical dashboard for professional oversight and handoff; (4) fine-tuning of Gemma on consented mental health dialogue datasets to improve domain-specific empathy and safety; (5) vector memory integration (pgvector or Chroma) for long-term semantic recall; (6) a longitudinal randomised controlled study measuring change in PHQ-9 and GAD-7 scores over 12-week Serene usage versus a control group; and (7) validated NLP-based safety classifiers to replace keyword-based crisis detection.

ACKNOWLEDGMENT

The authors thank the volunteers who participated in the pilot evaluation of Serene and the open-source communities behind Django, React, and the Hugging Face ecosystem. This work did not receive external funding.

REFERENCES

[1] World Health Organization, World Mental Health Report: Transforming Mental Health for All. Geneva: WHO, 2022.

[2] V. Patel, R. Araya, M. de Lima, A. Ludermir, and C. Todd, "Women, poverty and common mental disorders in four restructuring societies," Social Science & Medicine, vol. 49, no. 11, pp. 1461–1471, 1999.

[3] J. Torous, K. Lipschitz, M. Ng, and J. Firth, "Dropout rates in clinical trials of smartphone apps for depressive symptoms: A systematic review and meta-analysis," Journal of Affective Disorders, vol. 263, pp. 413–419, 2020.

[4] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial," JMIR Mental Health, vol. 4, no. 2, p. e19, 2017.

[5] B. Inkster, S. Sarda, and V. Subramanian, "An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study," JMIR mHealth and uHealth, vol. 6, no. 11, p. e12106, 2018.

[6] Z. Shao, Y. Feng, and W. Liang, "Is GPT-4 a good therapist? An evaluation of the GPT series on mental health counselling tasks," arXiv preprint arXiv:2304.01981, 2023.

[7] A. Sharma, A. S. Miner, D. C. Atkins, and T. Althoff, "A computational approach to understanding empathy expressed in text-based mental health support," in Proc. EMNLP, 2020, pp. 5263–5276.

[8] Google DeepMind, "Gemma: Open Models Based on Gemini Research and Technology," Technical Report, Google DeepMind,2024. [Online]. Available: https://storage.googleapis.com/deepmind-media/gemma/ge mma-report.pdf

[9] [K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a brief depression severity measure," Journal of General Internal Medicine, vol. 16, no. 9, pp. 606–613, 2001.

[10] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder," Archives of Internal Medicine, vol. 166, no. 10,pp. 1092–1097, 2006.

[11] S. M. Auerbach et al., "Leveraging large language models for mental health chatbots: Challenges, design principles, and evaluation," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 32, pp. 1024–1035, 2024.

[12] T. R. Insel, "Digital phenotyping: Technology for a new science of behavior," JAMA, vol. 318, no. 13, pp. 1215–1216, 2017.