IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MINDAI: AN AI-POWERED MENTAL WELLNESS ASSISTANT

[1]Om Rajesh Solanki, [2]Abhijit Suresh Kumbhar, [3]Yugandhara Raviraj Desai, [4]Ahana Asifahmed Saraff

[1,2,3,4]B.Tech Student

Department of Computer Science and Engineering,

D.Y. Patil Technical Campus, Talsande, Kolhapur, Maharashtra, India

Abstract: MindAI Therapy is an AI-driven mental wellness assistant that leverages LLaMA 3 — a large language model (LLM) served locally via Ollama — to deliver real-time, empathetic conversational therapy support. The system integrates a Flask-based web backend with a responsive HTML5/CSS3/JavaScript chat interface, browser-based voice recognition using the Web Speech API, and offline text-to-speech synthesis via pyttsx3, enabling fully multi-modal human–computer interaction. LangChain orchestrates prompt construction with gender-personalized system instructions, guiding the model's tone and response structure. Operating entirely on local hardware without any cloud dependency, MindAI Therapy ensures complete user data privacy — a critical requirement for mental health applications. Experimental evaluation across 30 sample interactions demonstrates response quality scores averaging 4.4 out of 5.0 across dimensions of empathy, relevance, actionability, safety, and conciseness. Inference latency ranges from 4.2 to 9.5 seconds on a CPU-only Intel i5 machine. Voice recognition accuracy reaches 96.2% in controlled environments. MindAI Therapy presents a practical, private, and cost-effective pathway toward accessible AI-assisted mental health support.

Index Terms — MindAI, Mental Health AI, LLaMA 3, Ollama, LangChain, Flask, Voice Therapy, Web Speech API, pyttsx3, Human-Computer Interaction, Large Language Models, Natural Language Processing.

I. INTRODUCTION

The rapid advancement of Natural Language Processing (NLP) and Artificial Intelligence (AI) has fundamentally transformed Human–Computer Interaction (HCI), enabling machines to engage in meaningful, context-aware dialogue with humans. In the domain of mental health, this transformation carries profound implications. Globally, over 970 million people live with a mental health disorder, yet access to professional therapy remains limited by cost, geographic availability, and persistent social stigma surrounding mental health care [1].

Conversational AI agents present a compelling complementary solution: they are available 24/7, carry no social judgment, can be deployed at near-zero marginal cost, and — when designed carefully — can provide empathetic, supportive dialogue grounded in evidence-based therapeutic principles. However, existing AI therapy tools predominantly rely on cloud-based large language model (LLM) APIs, which require transmitting sensitive personal conversations to external servers. This creates a significant privacy vulnerability that is particularly concerning in the mental health context.

MindAI Therapy addresses this critical gap by deploying LLaMA 3 — Meta's open-source large language model — entirely on local hardware via Ollama, a lightweight LLM serving framework. Combined with a Flask web backend, an HTML5/JavaScript frontend with voice interaction, and LangChain prompt orchestration, the system delivers empathetic, gender-personalized therapy sessions without any cloud dependency. The system's core contributions are: (1) a fully privacy-preserving AI therapy interface, (2) multi-modal voice and text interaction, (3) real-time response generation on

consumer-grade hardware, and (4) a modular architecture extensible to future enhancements such as emotion detection and memory persistence.

II. LITERATURE REVIEW

This section reviews prior work in AI-assisted mental health support, conversational LLMs, local model deployment, and voice-based HCI systems.

A. AI-Powered Mental Health Systems

The earliest AI therapy interface, ELIZA (Weizenbaum, 1966), demonstrated that rule-based pattern matching could simulate empathetic conversation. Modern systems have advanced substantially: Fitzpatrick et al. (2017) demonstrated that Woebot, a fully automated cognitive behavioral therapy (CBT) chatbot, significantly reduced depression and anxiety symptoms in college students over two weeks [3]. Despite this success, Woebot and similar commercial systems are cloud-dependent, raising data sovereignty concerns.

B. Large Language Models for Therapy

The emergence of transformer-based LLMs — including OpenAI's GPT series and Meta's LLaMA family — has dramatically improved the quality and naturalness of AI-generated therapeutic conversation. Brown et al. (2020) established that GPT-3 achieves strong performance on language tasks through few-shot prompting alone, without task-specific fine-tuning [1]. Touvron et al. (2023) subsequently demonstrated that the open-source LLaMA models achieve competitive results with significantly lower computational requirements, making them suitable for local deployment [2].

C. Local LLM Deployment and Quantization

The feasibility of local LLM deployment has been substantially advanced by model quantization techniques. Dettmers et al. (2022) showed that 4-bit quantization retains over 99% of model accuracy while reducing memory requirements by approximately 4×, enabling deployment of 8-billion-parameter models on standard consumer hardware without a GPU [5]. Tools such as Ollama abstract this complexity, providing a simple API for serving quantized models locally.

D. Sentiment Analysis and Emotion Detection

Sentiment analysis and emotion recognition play important roles in adaptive therapy systems. NLP libraries such as NLTK and spaCy provide foundational tools for preprocessing user input and detecting linguistic markers of emotional states. Research by Calvo and D'Mello (2010) highlights that affective computing systems capable of detecting sadness, anxiety, and stress can dynamically adapt conversational responses to better support users [7].

E. Voice-Based Interaction in Healthcare

Luxton et al. (2016) demonstrated that voice-based telehealth interfaces increase patient engagement and reduce dropout rates compared to text-only alternatives [6]. The Web Speech API, available natively in modern browsers, enables speech-to-text conversion without additional hardware, while pyttsx3 provides cross-platform offline text-to-speech synthesis, together making voice interaction accessible without cloud dependencies.

F. Research Gap

Existing AI therapy systems suffer from one or more of the following limitations: reliance on cloud APIs (privacy risk), absence of voice interaction, requirement for GPU hardware, or lack of open-source accessibility. MindAI Therapy is designed to address all four limitations simultaneously, offering a privacy-preserving, voice-enabled, CPU-deployable, open-architecture therapy assistant.

III. METHODOLOGY

3.1 Research Design

MindAI Therapy employs a system design and experimental evaluation methodology. The system was designed using an iterative, component-based approach, with each module independently developed and tested before integration. Evaluation followed a mixed-methods design: quantitative measurement of latency, accuracy, and throughput; and qualitative assessment of response quality by independent evaluators.

3.2 System Architecture

The system is structured into five primary modules: (1) **Flask Web Server Module** — Manages HTTP routing, user session state, and login/logout functionality; (2) **LLM Inference Module** — Runs the LLaMA 3 (8B parameters, 4-bit quantized) model locally via Ollama; (3) **Prompt Engineering Module** — Constructs context-aware therapy prompts using LangChain's ChatPromptTemplate; (4) **Voice I/O Module** — Speech-to-Text via Web Speech API; Text-to-Speech via pyttsx3; (5) **Frontend UI Module** — HTML5/CSS3/JavaScript chat interface with jQuery AJAX communication.

3.3 Participants and Sampling

System evaluation involved three independent evaluators with backgrounds in computer science and psychology, who assessed 30 simulated therapy interaction samples drawn from 10 thematic categories: stress, anxiety, loneliness, grief, relationship issues, work pressure, sleep disorders, self-esteem, motivation, and general emotional support.

3.4 Tools and Technologies

The complete implementation stack is listed in Table 1 below.

TABLE 1: TECHNOLOGY STACK

Layer	Technology	Role
Language	Python 3.10	Backend + AI logic
Language	JavaScript ES6	Frontend + voice
Language	HTML5 / CSS3	UI structure + style
Backend	Flask 3.0.0	Server & routing
LLM	LLaMA 3 (8B Q4)	AI response generation
LLM Server	Ollama	Local inference
Orchestration	LangChain	Prompt management
STT	Web Speech API	Voice to text
TTS	pyttsx3	Text to voice
AJAX	jQuery 3.2.1	Async requests

3.5 Procedure and Data Collection

The operational workflow of MindAI Therapy proceeds in six sequential steps: (1) User Authentication via Flask-protected login page; (2) Input Capture via keyboard or Web Speech API voice transcription; (3) Prompt Construction using LangChain's ChatPromptTemplate with gender preference and system therapist instruction; (4) LLM Inference via chain.invoke() to local LLaMA 3 model (temperature=0.5, num_predict=150, repeat_penalty=1.1, num_ctx=1024); (5) Voice Synthesis via pyttsx3 at 170 WPM in a background thread; (6) Response Delivery via Flask JSON response rendered in a styled chat bubble.

3.6 Data Analysis

Quantitative metrics — latency, token count, voice recognition accuracy — were recorded across 50 test sessions and analyzed descriptively (mean, range). Qualitative response quality was assessed using a structured rubric scoring Empathy, Relevance, Actionability, Safety, and Conciseness on a 1–5 Likert scale. Inter-rater reliability was computed using Cohen's Kappa coefficient ($\kappa = 0.81$, indicating strong agreement).

IV. RESULTS

4.1 Inference Latency

Response latency was measured as the time between AJAX request submission and complete response delivery. Results are shown in Table 2.

TABLE 2: INFERENCE LATENCY RESULTS

Test Condition	Avg Latency	Tokens
Short query (<10 words)	4.2 sec	~80
Medium query (10-20 words)	6.8 sec	~120
Long query (>20 words)	9.5 sec	~150
Voice input + response	7.4 sec	~110

4.2 Response Quality

Evaluators rated 30 interaction samples across five dimensions on a 1–5 scale ($\kappa = 0.81$). Results are shown in Table 3.

TABLE 3: RESPONSE QUALITY EVALUATION

Dimension	Score (1-5)	Observation
Empathy	4.3	Warm, non-judgmental tone
Relevance	4.5	Context-matched responses
Actionability	4.1	Clear coping strategies
Safety	4.7	No harmful suggestions
Conciseness	4.4	150-token cap effective
Overall Average	4.4	Strong therapeutic quality

4.3 Voice Recognition Accuracy

50 spoken queries were tested via the Web Speech API in Google Chrome: quiet environment (clear speech) — 96.2% accuracy; moderate ambient noise — 88.4% accuracy; accented or rapid speech — 79.6% accuracy.

4.4 System Feature Summary

TABLE 4: SYSTEM FEATURE IMPLEMENTATION

Feature	Implementation
Real-time chat	Flask /get + jQuery AJAX
Voice input (STT)	Web Speech API (Chrome)
Voice output (TTS)	pyttsx3 (offline)
Gender-personalized tone	LangChain prompt injection
Local privacy	Ollama (no cloud)
Video therapy mode	WebRTC + continuous STT
Session management	Flask session (cookie)

V. DISCUSSION

The results demonstrate that MindAI Therapy achieves its core design objectives with meaningful effectiveness. A response quality average of 4.4/5 across five dimensions — with Safety scoring highest at 4.7 — indicates that the system reliably produces therapeutic responses that are not only contextually appropriate but also free from harmful content, a critical requirement in mental health applications.

The inference latency of 4.2–9.5 seconds on a CPU-only machine, while perceptibly slower than cloud-based systems, falls within an acceptable range for a therapy context where thoughtful, considered responses are expected rather than instantaneous replies. In contrast to mental health applications where rapid misresponses could cause harm, this latency may even contribute to a perception of deliberateness in the AI's response.

The voice recognition accuracy of 96.2% in quiet environments confirms that the system is fully viable for single-user deployment in standard indoor settings. The degradation under noise (79.6% with accented speech) identifies a clear area for improvement through noise filtering or alternative STT backends.

A key differentiator of MindAI Therapy is its comparison profile against existing systems. Unlike Woebot or ChatGPT-based therapy tools, it requires no internet connectivity for LLM inference, transmits no user data externally, and operates without GPU hardware — making it deployable in resource-constrained or privacy-sensitive environments such as school counseling services, rural healthcare clinics, or personal devices.

Limitations include the absence of persistent session memory (each conversation starts fresh), a single hardcoded demo user account unsuitable for production multi-user deployment, and pyttsx3's reliance on OS-native TTS engines which may produce robotic-sounding output on some platforms. Future work should address these through vector-database conversation memory, proper user authentication, and cloud TTS fallback options.

VI. CONCLUSION

This paper presented MindAI Therapy, a fully local, privacy-preserving AI mental wellness assistant that combines LLaMA 3 language model inference via Ollama, LangChain prompt orchestration, Flask web serving, and multi-modal voice interaction through the Web Speech API and pyttsx3. The system addresses a significant gap in existing AI therapy tools by simultaneously offering full data privacy, voice-enabled interaction, CPU-only deployment, and open-source accessibility.

Key findings from experimental evaluation demonstrate: (1) response quality averaging 4.4/5 across empathy, relevance, actionability, safety, and conciseness; (2) inference latency of 4.2–9.5 seconds suitable for therapeutic contexts; (3) voice recognition accuracy of 96.2% in standard indoor environments; and (4) a complete feature set including real-time chat, gender-personalized responses, video therapy simulation, and offline operation.

Future development directions include persistent conversation memory using vector databases, multi-user authentication systems, mobile deployment via Edge AI optimization, facial emotion detection for adaptive response personalization, and structured mood tracking with longitudinal analytics. MindAI Therapy establishes a practical and ethically grounded foundation for the next generation of accessible, private AI-assisted mental health support tools.

VII. ACKNOWLEDGMENT

The authors gratefully acknowledge the Department of Computer Science and Engineering, D.Y. Patil Technical Campus, Talsande, Kolhapur, for providing the infrastructure and support necessary to carry out this research. Special thanks are extended to the independent evaluators who participated in the response quality assessment and provided valuable feedback.

REFERENCES

[1] T. Brown et al., “Language Models are Few-Shot Learners,” Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 1877–1901, 2020.

[2] H. Touvron et al., “LLaMA: Open and Efficient Foundation Language Models,” arXiv preprint arXiv:2302.13971, 2023.

[3] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, “Delivering Cognitive Behavior Therapy to Young Adults Using a Fully Automated Conversational Agent (Woebot),” JMIR Mental Health, vol. 4, no. 2, e19, 2017.

[4] J. Weizenbaum, “ELIZA — A Computer Program for the Study of Natural Language Communication Between Man and Machine,” Communications of the ACM, vol. 9, no. 1, pp. 36–45, 1966.

[5] T. Dettmers et al., “LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale,” arXiv preprint arXiv:2208.07339, 2022.

[6] D. D. Luxton, J. D. McCann, N. E. Bush, M. C. Mishkind, and G. M. Reger, “mHealth for Mental Health,” Professional Psychology: Research and Practice, vol. 42, no. 6, pp. 505–512, 2016.

[7] R. A. Calvo and S. D’Mello, “Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications,” IEEE Transactions on Affective Computing, vol. 1, no. 1, pp. 18–37, 2010.

[8] Flask Documentation, Pallets Projects, https://flask.palletsprojects.com, 2024.

[9] Ollama — Local Large Language Model Serving, https://ollama.com, 2024.

[10] LangChain Documentation, https://docs.langchain.com, 2024.