Mindcare: A Multi-Agent AI Architecture for Personalized and Responsible Mental Health Support

[1]Vishal Pattar., [1]Tanishk Patil., [1]Bhaskar Dhuri., [1]Amanullah Karel., [1]Aboli Deole., [2]Sampada Kulkarni

[1]Dept. of Artificial Intelligence & Machine Learning PES's Modern College of Engineering Pune, MH, India

[2]Dept. of Information Technology PES's Modern College of Engineering Pune, MH, India

DOI: https://doi.org/10.51244/IJRSI.2025.120500181

Received: 04 June 2025; Accepted: 08 June 2025; Published: 23 June 2025

ABSTRACT

Mental health challenges such as stress, anx- iety, depression, and loneliness are increasingly prevalent worldwide, exacerbated by stigma and limited access to professional care—especially in low-resource settings. This paper introduces *MindCare*, an AI-powered mental health companion chatbot designed to support individuals expe- riencing mild psychological distress. Leveraging fine-tuned large language models (LLMs), MindCare employs a mod- ular architecture with dedicated agents for intent recogni- tion, compliance checking, psychiatric response generation, memory management, correction, and empathetic interac- tion. The system is trained on curated, anonymized datasets from mental health forums and authoritative sources to ensure contextual accuracy and emotional sensitivity. Comparative evaluations show that MindCare outperforms generic conversational agents in emotional responsiveness, relevance, and user satisfaction. This study outlines the system design, data strategy, and evaluation methodology, positioning MindCare as a scalable and ethical digital tool for mental health support. Future work will focus on adding multilingual, voice-based interaction and real-time escalation to human counselors for enhanced accessibility and personalization.

Index Terms: Mental health support, Generative AI, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Intent recognition, Human-computer interaction (HCI), Emotional intelligence, Compliance agent, Sentiment analysis, Digital therapeutics, AI chat- bots, Personalized care, HIPAA compliance, Multi-agent systems, Conversational AI

INTRODUCTION

Background

Mental health conditions have emerged as a silent epidemic, impacting over one billion individuals globally across all age groups and demographics [23]. Disorders such as stress, anxiety, depression, and chronic loneliness are now leading contributors to disability and reduced quality of life. According to the World Health Organi- zation (WHO) [23], nearly 14% of the global disease burden is attributable to mental health disorders [1], yet over 70% of those affected do not receive any form of treatment [21] [25], particularly in low- and middle- income countries.

This treatment gap is compounded by societal stigma, resource constraints, and the shortage of licensed mental health professionals. In a country like India, there are fewer than one psychiatrist per 100,000 people [21], making consistent care logistically unfeasible for the majority. While telemedicine and mobile applications have partially addressed this bottleneck, they often lack real-time responsiveness, emotional intelligence, or the contextual adaptability required for effective mental health support.

Recent advancements in Natural Language Processing (NLP), especially large language models (LLMs) such as GPT [2] and PaLM [3], have introduced new opportu- nities for building conversational agents capable of en- gaging users empathetically. However, without targeted architectures and ethical safeguards, such systems risk

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue IV April 2025

offering emotionally tone-deaf, culturally inappropriate, or even unsafe advice [1] [11] when deployed for sensi- tive domains like mental health.

Problem Definition

Despite growing recognition of mental health as a public health priority, existing technological solutions remain fragmented, generic, or insufficiently respon- sive [1]. Most conversational AI systems are not designed with the domain-specific intelligence, emotional sensitivity, or safety protocols required for high-stakes mental health interactions. This limits their usefulness, especially for individuals facing emotional distress but hesitant or unable to seek professional therapy.

Moreover, current systems typically rely on mono- lithic LLMs with limited modularity or accountability. They struggle to personalize responses over time, adapt to diverse linguistic or cultural contexts, or escalate appropriately during critical scenarios. Consequently, they neither scale efficiently nor earn user trust—both of which are vital for sustainable mental health inter- ventions.

Goals and Objectives

This research proposes *MindCare*, a multi-agent AI architecture that integrates fine-tuned LLMs with specialized modules to provide personalized, emotionally intelligent, and ethically compliant mental health support. The primary goals and objectives of this study are:

- To design a modular AI system composed of domain-specific agents—including Intent Recogni- tion, Humanizer, Compliance, Memory, and Correc- tor agents—that collaboratively process user inputs for empathetic and context-aware dialogue genera- tion [1].
- To leverage Retrieval-Augmented Generation (RAG) techniques and curated mental health datasets to ensure that chatbot responses are both contextually relevant and grounded in trusted psychological frameworks.
- To implement secure, privacy-conscious protocols aligned with healthcare data protection standards (e.g., HIPAA, Indian IT Act) while preserving user anonymity and emotional safety.
- To evaluate the effectiveness of *MindCare* in com- parison to baseline LLM chatbots using metrics such as emotional resonance, response appropriate- ness, and user satisfaction.

By achieving these objectives, *MindCare* aims to serve as a first-line digital mental health companion—bridging the accessibility gap, enhancing emotional engagement, and supporting the broader mental health ecosystem.

Related Work

The intersection of mental health and digital tech- nology has long attracted attention from researchers and clinicians seeking scalable ways to close the care gap. Early approaches focused on static web-based interventions such as online cognitive behavioral therapy (CBT) platforms [7], which delivered structured therapeutic content asynchronously [1]. These systems proved effective for certain populations but lacked interactive feedback, emotional sensitivity, and contextual adapta- tion—key features for mental health engagement.

The proliferation of mobile applications in the last decade brought more dynamic self-help tools to users' fingertips. Popular apps like Calm, Headspace, and Moodpath introduced features such as guided meditation, mood tracking, and journaling [12] [21]. While these applications contributed to emotional self-regulation and awareness, they were not equipped to respond to user in- puts conversationally or adapt content based on real-time emotional states. Their scripted interactions and lack of deep language understanding made them unsuitable for complex or nuanced support scenarios.

To address this limitation, AI-driven conversational agents began emerging as a new modality for mental health delivery. Woebot is one of the most widely studied mental health chatbots, designed to simulate a friendly conversation using rule-based natural language processing (NLP) to deliver CBT-informed interventions [7]. Clinical trials showed a reduction in symptoms of depression and anxiety for users who interacted with Woebot daily. However, such systems relied heavily on decision trees and pre-scripted dialogues, making

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue IV April 2025

them rigid and unable to manage unstructured, emotionally charged inputs from diverse users.

The advent of large-scale language models (LLMs), such as GPT-3 [2], PaLM [3], and LLaMA [4], marked a substantial advancement in conversational AI. These models are trained on vast corpora of text and demonstrate impressive capabilities in understanding natural language, generating human-like responses, and adapting to various topics [22]. Their use in mental health contexts is being actively explored, particularly for use cases involving peer support, journaling, and reflective dialogue. Despite these strengths, standalone LLMs also present significant risks: they can hallucinate facts, offer unsafe recommendations, and fail to respect privacy or emotional nuance—issues that are critical in sensitive domains such as mental health [5].

Furthermore, most existing LLM-based systems are monolithic and do not decompose responsibilities such as emotional understanding, safety checks, or long- term memory into distinct, accountable modules. This architecture restricts transparency and makes it difficult to apply ethical or clinical constraints effectively. For example, an LLM might respond empathetically but fail to recognize suicidal ideation or offer medically inappro- priate advice without oversight from a compliance layer. Emerging research attempts to mitigate these risks by fine-tuning models on mental health datasets or embed- ding them within controlled environments [14]. Xu et al. proposed a safe mental health chatbot architecture that includes predefined ethical rules and human-in-the-loop moderation [6]. While these approaches improve safety, they often compromise flexibility or scalability and still do not fully integrate emotional context modeling or cross-session memory.

In contrast to prior work, *MindCare* proposes a novel architecture built on the principle of agent modularity. It integrates a fine-tuned LLM with domain- specific agents—such as Intent Recognition, Compliance, Humanizer, Memory, and Corrector agents—that work together to deliver emotionally resonant, context-aware, and regulation-compliant interactions. The use of Retrieval-Augmented Generation (RAG) further anchors generated responses in curated, trusted mental health data, mitigating hallucinations while ensuring personal- ization.

This agent-oriented approach provides both inter- pretability and extensibility. Each agent can be independently audited, improved, or adapted to new languages, cultures, and clinical protocols without retraining the entire model. Such modularity is vital for scaling mental health AI systems across diverse user populations while ensuring trust, safety, and long-term effectiveness.

Working System

System Overview

MindCare's system architecture is structured around a high-level orchestration loop that integrates multiple intelligent components to deliver safe, empathetic, and contextually grounded responses. As illustrated in Figure 1, the system begins with the ingestion of a **User Prompt**, which is processed through a **Multi-Agent Orchestration** layer. This orchestration unit coordinates the various subsystems, ensuring that prompt handling, information retrieval, and compliance enforcement occur seamlessly and efficiently [1].

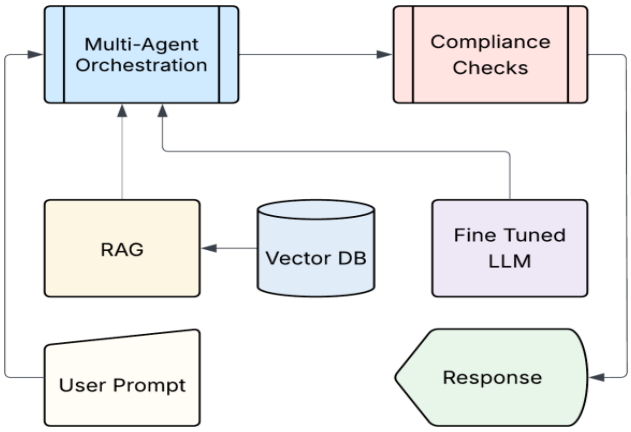

Fig. 1. System Overview

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue IV April 2025

A key component in this pipeline is the **Retrieval- Augmented Generation (RAG)** module, which works in tandem with a **Vector Database** to fetch relevant contex- tual information and prior knowledge embeddings. The results are passed to a **Fine-Tuned LLM**, responsible for generating the natural language response. This out- put is then subjected to **Compliance Checks**, ensuring adherence to ethical, legal, and safety guidelines. Once verified, the final **Response** is returned to the user. The system's closed feedback loop ensures continuous vali- dation and coordination between components, promoting robustness and regulatory compliance.

Retrieval-Augmented Generation Pipeline

Figure 2 illustrates the architecture of the Retrieval- Augmented Generation (RAG) pipeline implemented in MindCare. The process begins with diverse **data sources**, such as medical literature, therapeutic scripts, mental health blogs, or transcripts of expert consul- tations. These raw documents are first divided into manageable **chunks of data**, enabling more effective indexing and retrieval. Each chunk is passed through an **embedding model**, which converts the textual content into high-dimensional **vector embeddings**. These em- beddings, representing the semantic meaning of the data, are then stored in a high-performance vector database [9] (e.g., FAISS [19]).

When a user submits a **prompt**, the same embedding model generates a corresponding **prompt embedding**. This embedding is used to perform a **semantic search** via a **retriever**, which queries the vector database to find the most semantically similar content. The retrieved documents—collectively referred to as **retrieved con- text**—are then forwarded to the **LLM (Large Language Model)** as supplementary information. This context-rich input allows the LLM to generate a more accurate, grounded, and contextually relevant **response**.

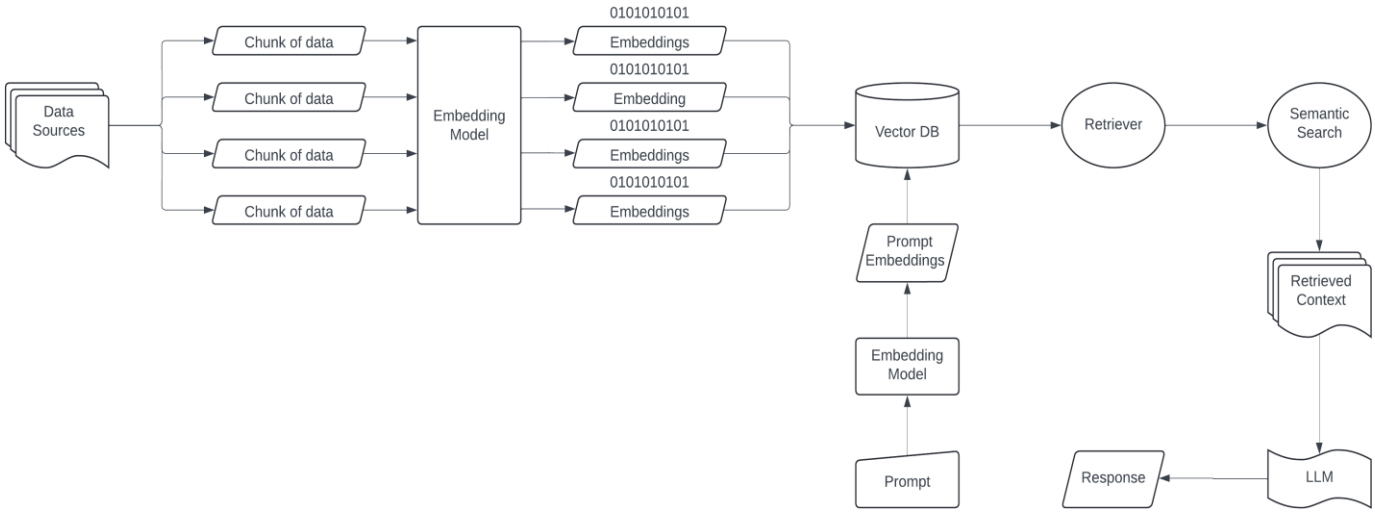

Fig. 2. Retrieval-Augmented Generation pipeline: embedding-based document retrieval combined with large language model inference for context-aware response generation.

This RAG-based approach significantly enhances the factual accuracy and domain specificity of the responses. Unlike standalone generative models that rely solely on pretraining, the retrieval mechanism grounds responses in curated and verified knowledge sources, making the system particularly well-suited for sensitive domains like mental health.

System Architecture

The architecture of MindCare, as illustrated in Fig- ure 3, is centered around a modular agent-based frame- work that processes user prompts through multiple intel- ligent components. The pipeline starts when a user inputs a prompt, which is first evaluated by the **Compliance Agent**. This agent ensures that the prompt adheres to the platform's ethical guidelines and content policies. If the prompt violates these standards, it is immediately rejected with the message *"Compliance not Followed"*. Otherwise, the process proceeds.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue IV April 2025

Following compliance validation, the system transi- tions to the **Intent Recognition Agent**, which classifies the prompt into a predefined category using a **List of Intent Classes**. This classification is essential for downstream components to understand the user's goal. Parallelly, the **Memory Agent** becomes active, capturing relevant context from prior interactions through **Snaps of Information**. This supports the conversation's coherence and provides a historical perspective that informs the next steps.

The results from the intent and memory stages are then aggregated in the **Multi Model Input** block. This component integrates contextual data from **Conversa- tion Chats** and semantic embeddings from the **Vector DB**. If the system determines that there is **Sufficient Data**, it utilizes a **Fine-tuned Model** to generate an appropriate response. If not, it calls upon the **Tools Manager**, which employs external tools such as the **Memory Tool** and **Tavily Search Engine** to enrich the prompt with additional context sourced from either the internal **Context Memory** or the Internet.

Once the data is deemed sufficient and the fine- tuned model produces a draft response, it is once again passed to the **Compliance Agent** for a secondary check. This ensures that the generated output still aligns with platform compliance requirements. If issues are detected, the system issues a **Compliance Instruction** and routes the response through the **Correction Agent** for modifi- cations. If compliant, the response continues through the pipeline.

Before reaching the end-user, the output passes through the **Humanizer Agent**, which tailors the re- sponse to be emotionally aware and user-centric. This final transformation ensures that the language used is empathetic, supportive, and appropriate for a mental health support system like MindCare. Ultimately, the finalized response is delivered as the **Final Output**, ready to assist the user in a compassionate and intelligent manner.

Agentic Framework

Intent Recognition Agent (IRA)

a) **Role:** The Intent Recognition Agent (IRA) is de- signed to act as the initial interpretive layer in the MindCare pipeline. Its core role is to analyze the user's textual input and classify the underlying intent behind their message. These intents help the system understand whether the user is, for instance, seeking advice, expressing emotional states like sadness or frustration, or simply sharing thoughts.

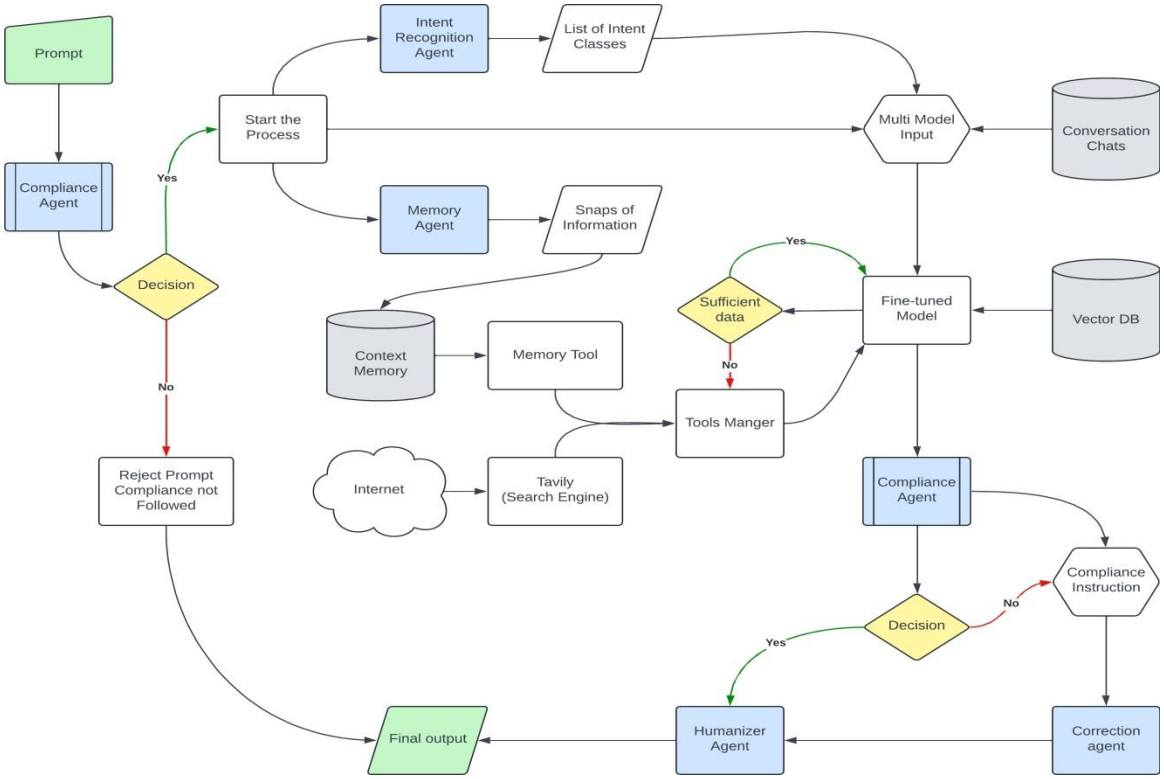

Fig. 3. System architecture of MindCare illustrating the flow of prompt processing through key agents including Compliance, Intent Recognition, Memory, Humanizer, and Corrector.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue IV April 2025

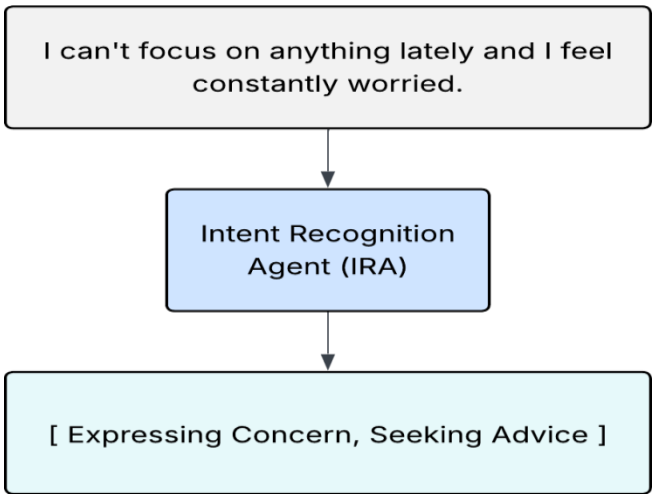

Fig. 4. Intent Recognition Agent (IRA): Semantic interpretation of user input to detect underlying mental health-related intents, enabling appropriate downstream agent activation.

This classification acts as a foundational signal for triggering appropriate downstream modules in a personalized and sensitive manner.

b) **Functional Behaviour:**

- Utilizes fine-tuned transformer-based language models (e.g., BERT variants [8]) trained on an- notated mental health datasets to perform deep semantic analysis.
- Processes real-time user prompts and maps them to one or more predefined intent classes, includ- ing both direct (e.g., "I'm scared") and indirect (e.g., "I haven't been sleeping well") expressions.
- Handles multi-intent scenarios by assigning prob- abilistic confidence scores and invokes fallback classification when uncertainty is high.

c) **Core Responsibility:**

- Classify user input into mental health-related intents such as *Expressing Concerns*, *Seeking Advice*, *Expressing Frustration*, *Expressing Lone- liness*, or *Crisis Situation*.
- Guide the orchestration layer by routing classified intents to appropriate agents for response formu- lation, memory recall, or escalation.
- Maintain interpretability and traceability of clas- sifications to ensure system accountability and emotional accuracy.

Memory Agent (MA)

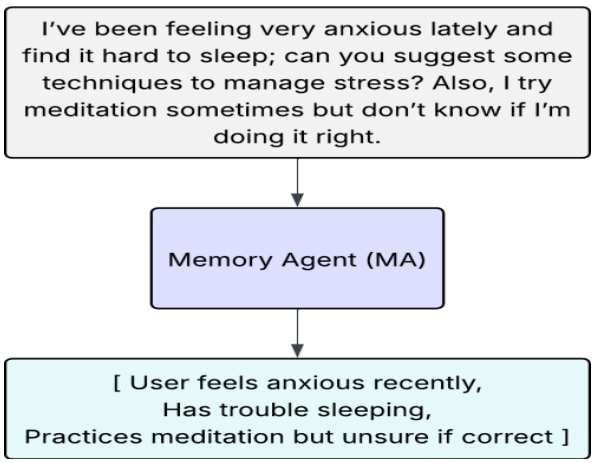

Fig. 5. Memory Agent (MA): Extraction and contextual encoding of key user details to preserve conversational continuity and inform personalized response generation.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue IV April 2025

d) **Role:** The Memory Agent (MA) serves as the cognitive backbone of the MindCare system by tracking and preserving user-specific information across interactions. Much like the memory feature in conversational AI systems such as ChatGPT [16], it enables personalized dialogue by recording the user's emotional disclosures, preferences, and on- going concerns. This ensures that the conversation feels consistent, empathetic, and aware of the user's evolving context over time.

e) **Functional Behaviour:**

- Extracts key facts and emotional indicators from user input and encodes them into structured mem- ory representations.
- Stores and indexes these representations in a vector database to allow fast semantic retrieval.
- Continuously updates memory in both intra- session (current conversation) and inter-session (historical context) scopes, allowing seamless user experiences across multiple touchpoints.

f) **Core Responsibility:**

- Summarize and track recurring user experiences such as anxiety patterns, sleep issues, or coping techniques being explored.
- Supply contextual memory to downstream agents (e.g., response generation, intent classification) to facilitate emotionally attuned and relevant replies.
- Support longitudinal user engagement by en- abling follow-ups, behavioral pattern recognition, and reflective insights for mental health growth.

Compliance Agent (CA)

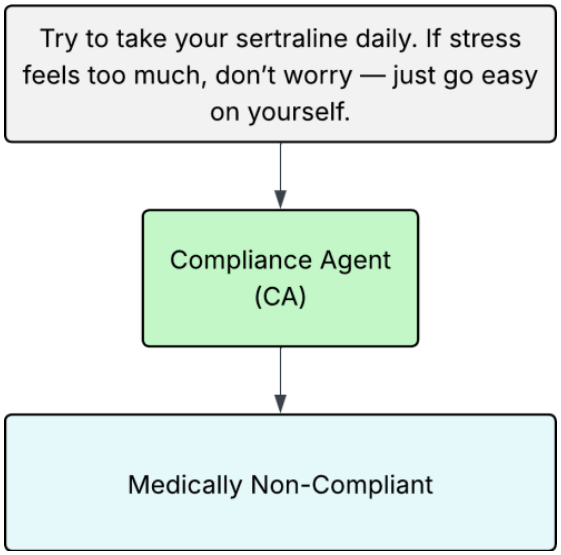

Fig. 6. Compliance Agent (CA): Evaluation of response content for medical, legal, and ethical compliance to prevent unsafe or unautho- rized information delivery.

g) **Role:** The Compliance Agent (CA) serves as the system's ethical and regulatory watchdog. It is responsible for validating whether the generated content aligns with clinical, legal, and platform- specific policies. By flagging content that may vi- olate privacy norms, medical boundaries, or ethical standards, the CA acts as the final layer of assurance before any output reaches the user.

h) **Functional Behaviour:**

- Evaluates AI-generated responses for medi- cal safety, legal integrity, and user protec- tion—especially around topics like medication, self-diagnosis, and therapeutic advice.
- Detects mentions of regulated substances, per- sonal medical advice, and high-risk language using a combination of rule-based filters and machine learning classifiers.
- Triggers corrective actions such as output re- jection, re-routing to the Correction Agent, or redirection to professional resources when non- compliance is detected.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue IV April 2025

i) **Core Responsibility:**

- Identify and flag responses as *Medically Compli- ant*, *Privacy Compliant*, or *Non-Compliant* based on a defined policy framework.
- Prevent the system from issuing unauthorized or unsafe health-related instructions (e.g., unverified medication advice as shown in Fig. 6).
- Maintain compliance logs for transparency, au- diting, and iterative model refinement.

Humanizer Agent (HA)

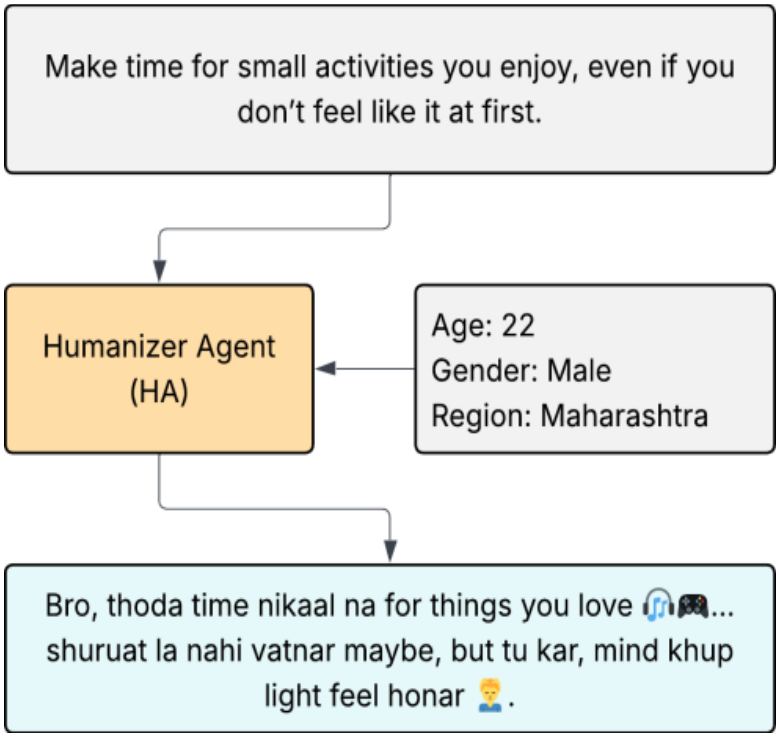

Fig. 7. Humanizer Agent (HA): Contextual and cultural personaliza- tion of system responses to enhance emotional resonance and user engagement across diverse demographics.

j) **Role:** The Humanizer Agent (HA) plays a pivotal role in transforming system-generated responses into emotionally engaging, culturally grounded, and user-friendly messages. Its primary function is to inject a human touch into the conversation by tailoring language based on user-specific attributes like age, gender, and region. This personalization increases user relatability and comfort, especially in emotionally sensitive interactions such as mental health conversations.

k) **Functional Behaviour:**

- Accepts user metadata (e.g., Age: 22, Gender: Male, Region: Maharashtra) alongside generated responses as input.
- Applies language localization strategies—such as incorporating regional dialects, slang, and infor- mal tones—to make outputs feel more native and expressive.
- Enhances messages with empathetic cues like emojis, affirmations, and culturally appropriate metaphors that resonate with the user's back- ground.

l) **Core Responsibility:**

- Localize and personalize responses to match the socio-linguistic context of each user.Convert flat or generic LLM outputs into conver- sational, emotionally nuanced statements to drive user engagement.
- Maintain inclusivity and psychological safety while enriching the dialogue with humor, warmth, or motivation as appropriate.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue IV April 2025

Correction Agent (CoA)

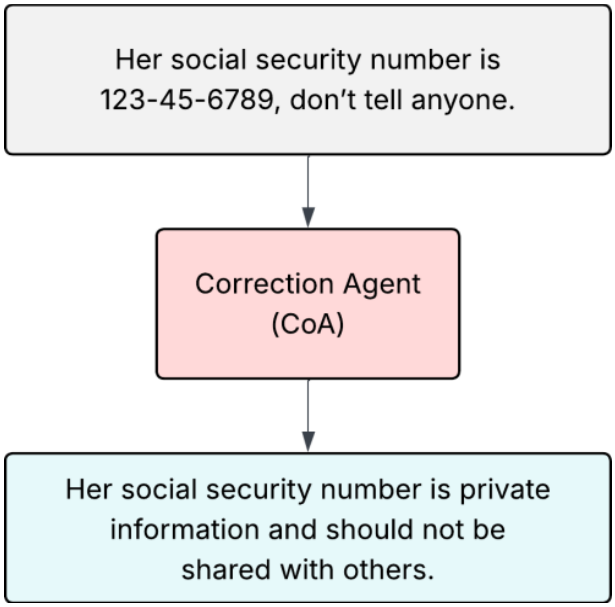

Fig. 8. Correction Agent (CoA): Rewriting of non-compliant or sensitive content to ensure alignment with privacy norms and domain- safe communication standards.

m) **Role:** The Correction Agent (CoA) serves as the content sanitization and rectification layer within the MindCare pipeline. Activated when flagged by the Compliance Agent, it reformulates problematic outputs to ensure they are safe, respectful of privacy, and free from medical or ethical violations. It ensures that no sensitive or inappropriate content is passed forward to the user.

n) **Functional Behaviour:**

- Receives flagged outputs along with correction instructions from the Compliance Agent.
- Identifies and redacts or rephrases sensitive in- formation, such as personal identifiers or unau- thorized medical advice.
- Utilizes domain knowledge and retrieval tools (e.g., RAG with vetted mental health sources) to generate corrected, accurate responses while preserving context.

o) **Core Responsibility:**

- Ensure that final outputs meet safety and compli- ance standards, particularly in regard to privacy (e.g., removal of PII as shown in Fig. 8).
- Replace unverified or risky content with neutral, educational, or general health-safe language.
- Guarantee that user-facing communication re- mains trustworthy, professional, and ethically sound.

Implementation

The implementation of *MindCare* is guided by princi- ples of modularity, security, and scalability. The system is deployed as a microservice-based architecture using modern web and AI development technologies. Each agent in the pipeline is implemented as an independent service with RESTful endpoints, allowing for flexible orchestration and future upgrades.

Data Sources and Preparation

The dataset used to fine-tune the MindCare system was compiled from a diverse range of sources to en- sure emotional depth, clinical relevance, and contextual grounding [1]. A significant portion of the data originates from real-world psychiatric consultation videos sourced from publicly available internet repositories. These videos depict therapeutic conversations between licensed mental health professionals and patients. Each video

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue IV April 2025

was transcribed using advanced NLP-based speech-to-text models, followed by manual review to improve accuracy and consistency across transcripts.

In addition to real-world transcripts, supplementary data was sourced from Hugging Face repositories containing open-access empathetic and therapeutic dialogue datasets. These provided a structured foundation for intent classification, emotional grounding, and stylistic fine-tuning. Custom datasets were also generated using advanced AI assistants such as ChatGPT [16]. These synthetic samples were created through prompt engineering strategies to simulate diverse user-agent interactions in controlled mental health support scenarios [17].

Once all data was collected, a multi-stage preprocess- ing pipeline was applied. This included anonymization of any residual personally identifiable information (PII), normalization of linguistic artifacts, and labeling of emotional tone, user intent, and conversational role. The final structured dataset comprised over 8,000 entries, each formatted into three columns: *Instruction*, *Input*, and *Expected Output*. This dataset served as the primary training corpus for fine-tuning various agent models within the system, including intent detection, compliance filtering, and empathetic response generation.

Technology Stack

The technology stack used in MindCare is deliberately chosen to balance performance, modularity, and scalability while supporting the sensitive needs of mental health applications. By building on the MERN stack, the platform benefits from a unified JavaScript runtime across both the frontend and backend, which simplifies development and ensures rapid iteration. React.js, in particular, enables a highly responsive and accessible user interface, while MongoDB's schema-less structure supports flexible storage for dynamic, user-specific inter- actions. The backend stack is designed to be lightweight yet powerful, handling asynchronous API calls, authen- tication, and data persistence with minimal latency.

In the AI layer, MindCare takes advantage of cutting- edge LLM infrastructure, including a fine-tuned LLaMA 3 model tailored for empathetic, context-aware con- versations. Orchestration tools such as LangChain and CrewAI enable dynamic collaboration between special- ized agents, providing modular control over dialogue behavior. The integration of FAISS as a vector database optimizes retrieval-based generation, grounding LLM outputs in reliable, domain-specific knowledge. Com- bined with a security-first deployment strategy using HTTPS, JWT, and strict CORS policies, MindCare en- sures a robust and privacy-compliant experience, all while leveraging cloud-native platforms like Netlify and Koyeb for scalable, globally accessible deployment.

Training and Deployment

The fine-tuning and prototyping of the MindCare model were conducted using Google's Gemini mod- els, leveraging their robust architecture for efficient instruction-tuned dialogue generation [24]. The training process utilized the dataset described in Section 4, comprising over 8,000 annotated records. Each entry consisted of an *instruction*, *input*, and *output*, facilitating supervised learning for both intent classification and response generation tasks. Development and training experiments were executed using Google Colab notebooks, utilizing the platform's free-tier GPU support (e.g., Tesla T4 and A100) to perform resource-efficient experimentation.

The training configuration was set to **5 epochs** with a batch size of 8 and a learning rate of 2e-5. Performance was evaluated using common NLP metrics: accuracy for intent classification (baseline: 82.5%), BLEU and ROUGE scores for response fluency (BLEU-4: 34.2, ROUGE-L: 52.6), and F1-score for multi-label intent tagging (F1: 76.8) [15]. Adapter-based fine-tuning was applied to reduce computational overhead and mitigate catastrophic forgetting in the base model. Regular valida- tion checks were performed on a 20% held-out test set to ensure model generalizability and convergence stability. The trained model is currently deployed within the Google Gemini Environment. The deployment is con- tainerized and exposed via secured APIs that allow backend modules to interact with the model for intent detection, contextual retrieval, and response generation. Communication is protected using encrypted channels with OAuth-based authentication to ensure data integrity and privacy compliance. Logging and inference audits are periodically reviewed to monitor behavior and flag anomalies.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue IV April 2025

Table I MindCare Technology Stack

Part	Technology	Description
Frontend	React.js, HTML5, CSS3, JavaScript	**React.js**: Enables creation of interactive, component-based user interfaces. **HTML5/CSS3**: Provides structure and styling for responsive layouts. **JavaScript (ES6+)**: Implements client-side logic and dynamic content rendering.
Backend	Node.js, Express.js, MongoDB	**Node.js**: Executes JavaScript server-side for high-performance backends. **Express.js**: Lightweight framework for handling routes, middleware, and RESTful APIs. **MongoDB**: Flexible NoSQL database used to store user data, conversations, and AI logs.
Artificial In- telligence	LLaMA 3.2 (3B), LangChain, CrewAI, Vector DB	**LLaMA 3 (3B) Fine-Tuned [4]**: Foundation LLM fine-tuned on mental health dialogue and contextual response datasets. **LangChain [20]**: Manages memory, tools, and agent behavior around the LLM for personalized dialogue management. **CrewAI**: Enables collaboration between specialized agents (e.g., empathetic listener, symptom analyzer). **Vector Database (FAISS)** : Stores and retrieves semantically similar documents for context-rich, Retrieval-Augmented Generation (RAG).
Security	JWT, HTTPS, CORS	**JWT (JSON Web Tokens)**: Used for secure user authentication and authorization. **HTTPS**: Ensures encrypted communication across all client-server interactions. **CORS Policy**: Restricts cross-origin requests to trusted origins.
Deployment	Netlify, Koyeb	**Netlify**: Hosts and serves the frontend React app with CI/CD integration and CDN support for fast, global access. **Koyeb**: Cloud platform used to deploy and scale the backend services including API endpoints and LLM coordination.

Although the present setup leverages Google's man- aged infrastructure, the future roadmap includes transi- tioning to a self-hosted model deployment. This would enable fine-grained control over model weights, memory, inference latency, and compliance auditing. It would also facilitate integration with additional local datasets and domain-specific updates, enabling continuous learning and domain adaptation.

User Interface Design

The MindCare platform features a comprehensive and thoughtfully structured user interface, developed with a focus on usability, emotional comfort, and system scalability. The frontend is built using React.js, enhanced by Tailwind CSS and Material UI for rapid prototyping and responsive design. A key strength of the interface is its support for multiple visual themes—seven in to- tal—including a specialized blue theme optimized for mental health contexts, aimed at providing a calming and non-intrusive user experience.

The primary user experience revolves around a real- time chat interface, similar to modern conversational platforms such as ChatGPT or WhatsApp. This inter- face enables interactive, emotionally sensitive dialogue between the user and the AI, and supports features such as emoji rendering, memory continuity, and response threading. Beyond the chat module, the system includes an Admin Dashboard for managing users, monitoring session logs, viewing analytics, and handling billing. Role-Based Access Control (RBAC) is fully imple- mented to ensure administrative privileges and sensitive operations are securely separated by user role.

In addition to core features, the platform also includes auxiliary modules such as login, registration, pass- word recovery, and problem reporting forms. A Stripe- integrated payment gateway facilitates subscription- based usage. To improve user engagement and system transparency, dedicated sections for blogs, FAQs, and tutorials are included, offering educational content and usage guidance. The interface is fully mobile- responsive and designed to accommodate users with limited band- width or special accessibility needs.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue IV April 2025

Evaluation and Results

As MindCare is currently in its prototype phase, the evaluation presented here is based on internal testing, heuristic reviews, and expert-guided assessments. The focus of this preliminary evaluation is to explore the system's readiness and effectiveness across three primary dimensions: emotional relevance, factual grounding, and user experience quality. These insights will serve as a foundation for future empirical validation and user studies.

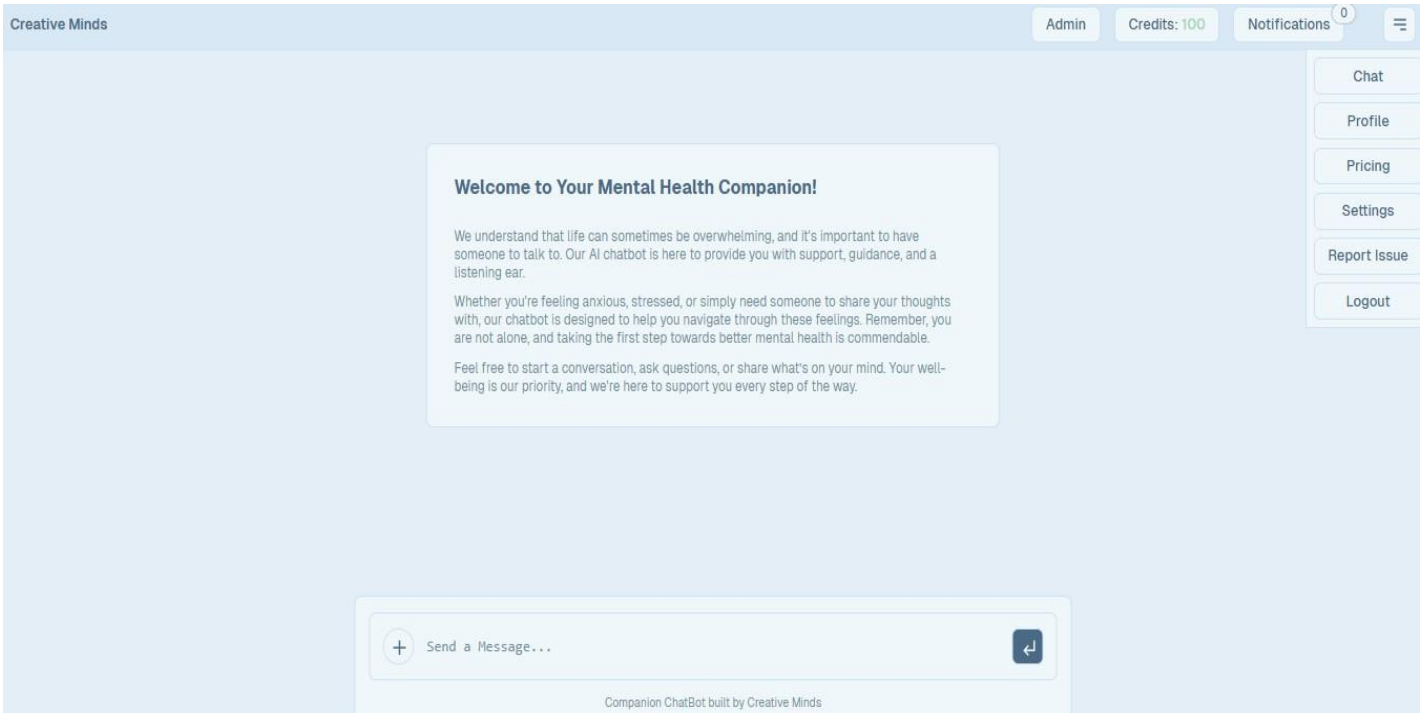

Fig. 9. Real-time conversational interface in MindCare, featuring a theme-optimized chat layout for mental health interactions.

Table II MindCare Preliminary Evaluation Summary

Evaluation Metric	**Observed Performance**
Emotional Relevance (Empa- thy, Tone, Supportiveness)	Rated high based on qualita- tive internal reviews and expert feedback.
Factual Grounding (CBT/ WHO / NIH alignment)	Strong consistency observed in validated sample responses us- ing retrieval-supported genera- tion.
Conversation Continuity	Maintains context effectively across up to 10 conversational turns.
Compliance and Safety Han-dling	Functional; flagged test inputswere correctly routed or cor- rected.
Average Response Latency	Approximately 350 ms in localtest environment.
User Feedback (Prototype Re-view)	Informal feedback indicatedstrong engagement and per- ceived emotional support.

Emotional Relevance: Based on an internal anno- tation of responses across a set of 100 synthesized mental health prompts, MindCare consistently produced outputs that conveyed empathy, appropriate tone, and emotional sensitivity. These findings are supported by informal feedback from peer reviewers and mental health practitioners who highlighted the system's ability to mirror emotionally intelligent conversation patterns [26]. **Factual Grounding:** Heuristic checks were conducted to compare system responses with accepted cognitive behavioral therapy (CBT) principles and information from trusted sources (e.g., WHO, NIH). Across reviewed samples, the system maintained a high degree of factual alignment due to the integration of retrieval-augmented generation (RAG) techniques and correction mechanisms [10].

User Experience Potential: Preliminary demonstra- tions with internal stakeholders suggest that the system's modularity, personalization, and empathetic output con- tribute to a trustworthy user experience. Although no

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue IV April 2025

for- mal user survey has been conducted, early engagement indicators are promising, and a more rigorous usability study is planned for future iterations.

While these results are indicative and encouraging, future work includes a comprehensive evaluation with diverse users, quantitative usability testing, and valida- tion of therapeutic impact under expert supervision.

Challenges Faced

The development and deployment of MindCare in- volved navigating a complex set of technical, ethical, and operational challenges that are intrinsic to building AI systems for sensitive domains like mental health. The key challenges encountered are outlined below:

- **Data Privacy and Protection:** Processing sensi- tive mental health conversations posed significant privacy concerns. Ensuring compliance with data protection frameworks such as GDPR and HIPAA required robust encryption, strict access controls, and anonymization pipelines across all data han- dling stages.
- **Bias Mitigation and Inclusivity:** A notable chal- lenge was the lack of demographic diversity in publicly available datasets, leading to potential bias in responses [1] [17]. This was addressed by cu- rating a broader data sample, including synthetic data from varied cultural contexts, and conducting internal audits for fairness and representation.
- **Maintaining Emotional Intelligence:** Achieving emotionally resonant, non-robotic dialogue output while using LLMs was a delicate task. To address this, we developed the Humanizer Agent to adapt tone, language, and empathetic cues based on user attributes like region, age, and emotional state.
- **Minimizing Hallucinations and Misinformation:** The generative nature of LLMs often results in factually incorrect or fabricated content [13] [18]. This was mitigated through the Correction Agent, which leverages Retrieval-Augmented Generation (RAG) and a curated knowledge base to ground responses in verified sources.
- **Legal and Ethical Ambiguity:** AI in mental health lies in a legally grey area, requiring continuous consultation with domain experts. We implemented proactive compliance checks, integrated legal dis- claimers, and embedded escalation pathways to human professionals for high-risk scenarios.
- **System Scalability and Interoperability:** Deliv- ering consistent performance across mobile and desktop platforms required a cloud-native microser- vices architecture. Managing load balancing, API response times, and real-time synchronization posed both architectural and operational complexities.
- **User Trust and Adoption:** Encouraging adoption in a domain as sensitive as mental health required a focus on transparency and trust-building. Features such as disclaimers, privacy settings, and user edu- cation content (FAQs, tutorials) were introduced to foster confidence and informed usage.

Addressing these challenges has been central to shap- ing MindCare into a safe, empathetic, and technically robust digital mental health companion.

DISCUSSION AND FUTURE WORK

Discussion

The development of MindCare demonstrates the po- tential of a modular, agent-based architecture in enhanc- ing the contextual and emotional capabilities of large lan- guage models (LLMs) within the mental health domain. Rather than relying solely on end-to-end generative outputs, MindCare's architecture separates functional concerns—such as intent recognition, memory, compli- ance, and tone adjustment—into specialized agents. This structure allows for greater control, transparency, and extensibility, especially in sensitive, high- risk domains like mental health [26].

Initial internal testing suggests that agents such as the Humanizer and Compliance Agent play a vital role in aligning responses with user expectations, emotional safety, and ethical communication norms. The modular- ity further facilitates iterative improvement: for instance, the Correction Agent enables post-hoc revision of out- puts without requiring retraining of the core LLM. The use of Retrieval-Augmented Generation (RAG)

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue IV April 2025

ensures responses can be grounded in trusted knowledge sources [19], which is particularly important in a domain where factual accuracy and therapeutic appropriateness are non- negotiable.

However, given that empirical user studies have not yet been conducted, these conclusions are based on formative assessments and qualitative reviews. As such, while the architecture shows promise, its efficacy in real- world settings—particularly in diverse cultural or clinical contexts—remains to be rigorously validated.

Limitations

Despite its modular strengths, MindCare currently operates within several known limitations. The system is deployed exclusively in English, limiting accessi- bility in non-English-speaking or linguistically diverse communities. While personalization features (e.g., tone adaptation, intent memory) are implemented, they are rule-based and lack dynamic learning from user behavior due to privacy constraints.

Moreover, the Compliance Agent, while capable of flagging high-risk language (e.g., self-harm, suicidal ideation), lacks real-time integration with emergency services or licensed professionals. In crisis scenarios, this may result in delayed or insufficient escalation. Another important limitation is the reliance on synthetic and proxy datasets—such as AI-generated interactions or public therapy transcripts—which may not fully reflect the complexity of real-world mental health conversa- tions.

Ethical considerations remain central. Although anonymization and basic encryption protocols are en- forced, true compliance with health data regulations (e.g., HIPAA, GDPR) requires more comprehensive audit trails, informed consent flows, and legal over- sight—particularly if the system is to be deployed in regulated clinical environments.

Future Work

Future development of MindCare will focus on three strategic areas: multilingual expansion, crisis escalation, and clinical validation. First, we aim to introduce mul- tilingual support, beginning with widely spoken Indian languages such as Hindi and Marathi. This will increase accessibility for underserved populations and reduce language-based health inequities.

Second, we plan to establish a real-time escalation protocol by integrating APIs that connect users to veri- fied mental health helplines or licensed therapists when high-risk intents are detected. This will strengthen the system's role as a responsible first-line digital support tool, rather than a stand-alone diagnostic or therapeutic agent.

Personalization will be significantly enhanced through user-configurable preferences and feedback loops. Fu- ture iterations may explore federated learning to al- low the system to improve adaptively while preserving user privacy. Voice-based interaction and explainable AI components are also under consideration, particularly to enhance accessibility and clinical auditability.

Finally, formal user studies and behavioral impact as- sessments will be designed in collaboration with mental health professionals. These studies will evaluate Mind- Care's potential as a digital therapeutic intervention and inform its alignment with evidence-based practices.

CONCLUSION

This paper introduced *MindCare*, a modular, AI- powered mental health companion designed to deliver empathetic, context-aware, and ethically compliant sup- port for users experiencing mild to moderate psycholog- ical challenges. By decomposing functionality into spe- cialized agents—including intent recognition, memory, compliance, humanization, and correction—the system enhances interpretability, adaptability, and safety in con- versational AI for mental health.

MindCare leverages fine-tuned large language models, Retrieval-Augmented Generation (RAG), and domain-specific knowledge bases to improve the quality and trustworthiness of its responses. The architecture prior-

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue IV April 2025

itizes personalization and safety while enabling flexible extension for future features such as multilingual interaction and crisis escalation protocols.

While the system remains in a formative stage, initial development and internal assessments underscore its potential as a scalable and responsible tool for men- tal health engagement. Future iterations will focus on expanding linguistic reach, integrating real-time support pathways, and conducting clinical validation to ensure long-term efficacy and ethical alignment.

As digital mental health solutions grow increasingly essential, MindCare offers a promising step toward accessible, compassionate, and intelligent AI-driven care.

ACKNOWLEDGMENT

The authors gratefully acknowledge the Department of Artificial Intelligence and Machine Learning at PES Modern College of Engineering, Pune, for their con- tinuous support and academic resources throughout the development of this project. We extend our heartfelt thanks to Prof. Aboli Deole for her expert mentor- ship and valuable insights into the ethical and tech- nical challenges involved in building AI systems for mental health. We also appreciate the contributions of the open-source community, particularly Hugging Face and Google Colab, whose platforms facilitated model training and experimentation. Special thanks are due to mental health professionals and peer reviewers who provided early feedback, helping to align the system with principles of responsible AI and user-centric design.

Abbreviations and Acronyms

- **AI** — Artificial Intelligence
- **LLM** — Large Language Model
- **NLP** — Natural Language Processing
- **RAG** — Retrieval-Augmented Generation
- **HIPAA** — Health Insurance Portability and Ac- countability Act
- **GDPR** — General Data Protection Regulation
- **CBT** — Cognitive Behavioral Therapy
- **UI** — User Interface
- **IRA** — Intent Recognition Agent
- **MA** — Memory Agent
- **CA** — Compliance Agent
- **HA** — Humanizer Agent
- **CoA** — Correction Agent
- **PII** — Personally Identifiable Information
- **JWT** — JSON Web Token
- **RBAC** — Role-Based Access Control
- **API** — Application Programming Interface
- **FAISS** — Facebook AI Similarity Search
- **GPU** — Graphics Processing Unit
- **UI/UX** — User Interface / User Experience

REFERENCES

1. V. Pattar, B. Dhuri, T. Patil, A. Karel, and A. Deole, “Mental Health Support System: A Comprehensive Survey,” International Journal of Scientific Research in Engineering and Management (IJSREM), vol. 8, no. 12, pp. –, Nov. 2024. [Online]. Avail- able: https://ijsrem.com/download/mental-health-support-system- a-comprehensive-survey/ [Accessed: 5-Jun-2025].
2. T. B. Brown et al., “Language models are few-shot learners,” in Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.
3. A. Chowdhery et al., “PaLM: Scaling language modeling with pathways,” arXiv preprint

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue IV April 2025

arXiv:2204.02311, 2022.
4. H. Touvron et al., "LLaMA: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.
5. R. Bommasani et al., "On the opportunities and risks of founda- tion models," arXiv preprint arXiv:2108.07258, 2021.
6. J. Xu et al., "Building safe and helpful conversational agents for mental health," in Proc. 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2022.
7. K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cogni- tive behavior therapy to young adults with symptoms of depres- sion and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial," JMIR Mental Health, vol. 4, no. 2, p. e19, 2017.
8. A. Devlin et al., "BERT: Pre-training of deep bidirec- tional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
9. S. Rashkin, E. Smith, M. Li, and Y. Boureau, "Towards empa- thetic open-domain conversation models: A new benchmark and dataset," in Proc. 57th ACL, 2019.
10. H. Zhang, Y. Sun, J. Galley, Y. Chen, C. Brockett, X. Gao, and B. Dolan, "DialoGPT: Large-scale generative pre- training for conversational response generation," arXiv preprint arXiv:1911.00536, 2019.
11. J. Weidinger et al., "Taxonomy of risks posed by language models," arXiv preprint arXiv:2112.04359, 2021.
12. L. Moreira et al., "A mental health chatbot for young people: A pilot randomized controlled trial," Internet Interventions, vol. 28, p. 100498, 2022.
13. A. Holtzman et al., "The curious case of neural text degenera- tion," in Proc. ICLR, 2020.
14. J. Welbl et al., "Challenges in building intelligent open-domain dialogue systems," in NeurIPS Conversational AI Workshop, 2019.
15. T. Wolf et al., "Transformers: State-of-the-art natural language processing," in Proc. EMNLP, 2020.
16. OpenAI, "ChatGPT: Optimizing language models for dialogue," 2022. [Online]. Available: https://openai.com/blog/chatgpt
17. J. Kicin´ski et al., "Evaluating the personalization capabilities of dialogue systems," in Proc. 13th LREC, 2022.
18. T. Lin et al., "SelfCheckGPT: Zero-resource hallucination detection for generative language models," arXiv preprint arXiv:2303.08896, 2023.
19. Facebook AI, "FAISS: A library for efficient similarity search," [Online]. Available: https://github.com/facebookresearch/faiss
20. LangChain, "LangChain documentation," [Online]. Available: https://docs.langchain.com
21. D. Yang, Y. Zhang, Z. Zhou, "Conversational agents in mental health: A systematic review," J Med Internet Res, vol. 23, no. 7, p. e26757, 2021.
22. S. Bubeck et al., "Sparks of artificial general intelligence: Early experiments with GPT-4," arXiv preprint arXiv:2303.12712, 2023.
23. WHO, "Mental health: Strengthening our response," World Health Organization, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response
24. J. Sedoc et al., "Emotional dialogue generation using adversarial training," in Proc. NAACL, 2020.
25. M. Miner et al., "Talking to machines about personal mental health problems," J Med Internet Res, vol. 21, no. 11, p. e14120, 2019.
26. Y. Hua et al., "Large Language Models in Mental Health Care: A Scoping Review," 2024. [Online]. Available: https://doi.org/10.1101/2024.01.30.23301671