IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, May 2026
Impact Factor: 8.2

A Structured and Safety-Aware Conversational Framework Using Large Language Models and Retrieval-Augmented Generation for Mental Health Support

Mr. Varad Deokar[1], Mr. Atharva More[2], Mr. Manav Parmar[3], Prof. Kopal Gangrade[4]
Department of Computer Engineering[1-4]
Pune Institute of Computer Technology, Pune, India
varadvdeokar@gmail.com, atharvadmore20@gmail.com,
mmparmar044@gmail.com, kgangrade@pict.edu

Abstract: *The growing demand for accessible mental health support has led to the development of AI-based conversational systems; however, many existing solutions either lack structure or fail to ensure safety and consistency. This paper presents a structured and safety-aware conversational framework that integrates Retrieval-Augmented Generation (RAG) with Large Language Models (LLMs) in a controlled, multi-stage pipeline. The proposed system incorporates dedicated modules for crisis detection, assessment, cognitive reflection, and intervention, coordinated through an intent-driven orchestration layer. Unlike conventional chatbots, the system emphasises deterministic dialogue flow and psychological structuring using established frameworks. Evaluation using simulated conversational scenarios demonstrates that the framework produces more coherent, context-aware, and safe responses compared to unstructured approaches. The results highlight the importance of architecture-driven design in building reliable and responsible AI systems for mental health support*

Keywords: Mental health chatbot, LLM, RAG, CBT, Structured pipeline, Safety-aware AI

I. INTRODUCTION

A. Rule-Based and Retrieval-Based Chatbots:

In order to provide deterministic and secure responses, early conversational agents relied on rule-based or retrieval-based techniques, which offered little flexibility and personalisation [1, 3, 8]. These systems are useful for providing basic support, but they lack deeper therapeutic engagement and context awareness [7, 10].

B. CBT-Based Conversational Systems:

Structured interventions like cognitive restructuring and mood tracking were introduced by CBT-based chatbots, which demonstrated efficacy in lowering anxiety and depression [1, 7, 13]. However, their lack of explicit multi-stage therapeutic modelling and frequent reliance on predefined flows limit their adaptability [12].

C. LLM-Based Mental Health Agents:

Natural, context-aware dialogues and increased engagement are made possible by LLM-based systems [2, 4]. Knowledge grounding and personalisation are improved by methods such as RAG [5, 6, 11]. These systems, however, lack structured control and are vulnerable to inconsistent behaviour, hallucinations, and safety hazards in delicate situations [2, 5, 13].

DOI: 10.48175/IJARSCT-34979

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, May 2026
Impact Factor: 8.2

D. Safety, Evaluation, and Ethical Challenges:

Risks like inadequate crisis detection and untrustworthy responses continue to make safety a top priority [2, 8]. To increase reliability, recent research suggests evaluation frameworks that use expert assessment and simulated users [5, 11]. Nevertheless, issues such as a lack of clinical validation and standardised assessment remain [1, 8, 10].

E. Summary of Limitations and Research Gap:

From the above survey, it is evident that existing mental health conversational agents suffer from several key limitations:

• lack of structured, multi-stage therapeutic pipelines;
• absence of deterministic dialogue control and state modelling;
• limited integration of safety-aware mechanisms such as crisis detection; and
• over-reliance on either rigid rule-based systems or unconstrained generative models.

Your paper must be in single column format with a space of 4.22mm (0.17") between columns.

II. PROBLEM STATEMENT

Current chatbots for mental health are either LLM-based, which are flexible but unstructured and potentially unsafe, or rule-based, which are secure but rigid and repetitive. This gap highlights the need for a structured therapeutic architecture that integrates strong safety mechanisms to provide reliable and interpretable mental health support, while combining controlled dialogue flow with generative capabilities.

III. METHODOLOGY

A. System Overview:

EmpathAI, the proposed system, is a modular chatbot for mental health support that leverages cognitive-behavioural methods to deliver structured emotional assistance. A central orchestration layer coordinates multiple specialised modules, including FIDO assessment, Hot Cross Bun (HCB) analysis, cognitive reframing, coping strategies, and grounding exercises. The system adopts a hybrid approach combining:

rule-based state machines;
Retrieval-Augmented Generation (RAG);
Large Language Models (LLMs).

Prior research highlights that: (i) LLM-based chatbots may generate unsafe or hallucinated responses without proper grounding [4]; (ii) evaluation frameworks increasingly emphasise safety-first design and clinically informed interaction protocols [2, 13]; and (iii) realistic evaluation requires simulated users and therapist-informed assessment [2]. To address these gaps, we propose the FIDO–HCB framework—an organised, interpretable, and safety-controlled conversational architecture.

B. System Architecture:

The architecture consists of the following components:

User Interface Layer:

- FastAPI-based interaction with a React frontend.;
- A session management module maintains session state and conversation history.

Orchestration Layer:

- Routes user input to appropriate modules.;
- Uses LLM-based intent classification.;
- Maintains active module state and transitions.

Processing Modules:

- FIDO (problem assessment);

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, May 2026
Impact Factor: 8.2

- HCB (CBT cycle analysis);
- Reframing module (thought restructuring);
- Coping module (problem-solving strategies);
- Grounding module (emotional regulation)

Knowledge Layer:

- Vector databases (ChromaDB);
- Embedding model: BAAI/bge-small-en-v1.5;
- Stores curated mental health strategies

LLM Layer;

- Google Gemini (gemini-2.5-flash);
- Used for reasoning, response generation, and routing

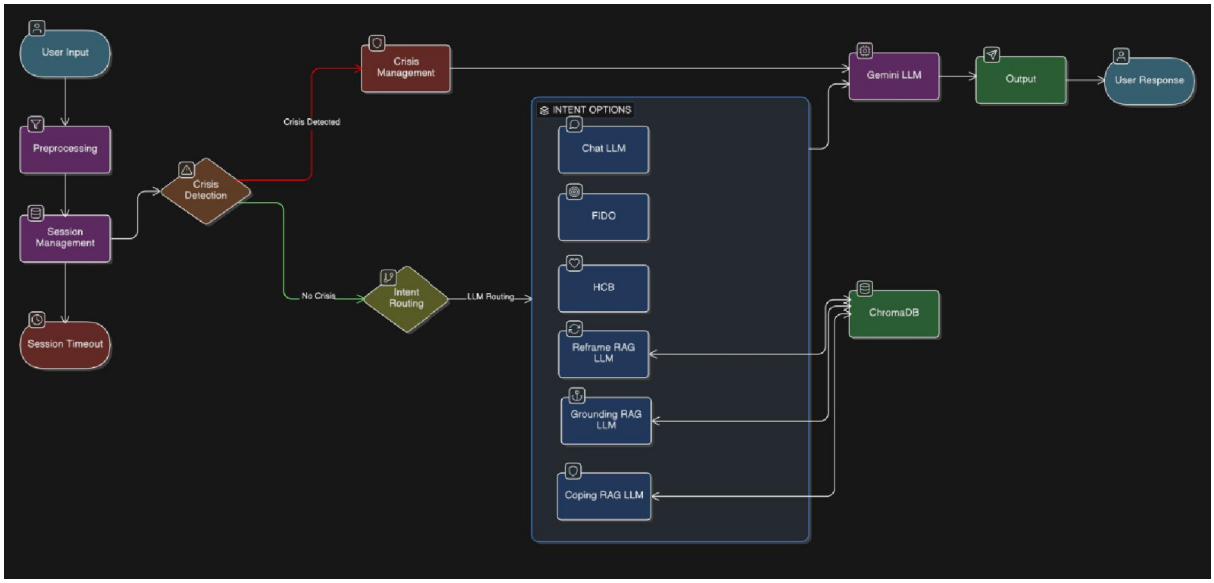

Figure:1 Architecture Diagram

C. Intent Detection and Routing:

The system uses an LLM-based router to classify user input into modules: FIDO, HCB, REFRAME, COPING, GROUNDING, or CHAT. Routing decisions are based on the current active module, conversation context, and semantic meaning of input. A crisis detection module is executed prior to routing; if high-risk intent (e.g., self-harm) is detected, the system immediately provides emergency support resources.

D. FIDO Assessment Module:

The FIDO module guides users through structured stages implemented as a finite-state machine: Frequency, Intensity, Duration, Onset, and Completion. At each stage, the system records user input and uses the LLM to generate empathetic, context-aware prompts. Predefined transition logic ensures controlled progression, transforming subjective emotional experiences into measurable dimensions for better understanding and intervention.

E. Hot Cross Bun (HCB) Model:

The HCB module models cognitive-behavioural relationships through structured steps: Situation, Thoughts, Feelings, Physical Sensations, and Behaviours. User inputs are organised into a coherent CBT cycle, enabling identification of

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, May 2026 Impact Factor: 8.2

links between cognition, emotion, and behaviour. After completion, the system summarises the cycle and suggests appropriate interventions, fostering self-awareness and targeted therapeutic direction.

F. Retrieval-Augmented Generation (RAG) Modules:

1) Embedding and Storage:

Textual knowledge (e.g., coping strategies, grounding exercises) is converted into vector embeddings using SentenceTransformers and stored in ChromaDB across collections such as reframing, problem-solving, and grounding.

2) Retrieval Process:

User input is embedded into vector space, and relevant documents are retrieved using cosine similarity (Top-K retrieval).

3) Response Generation:

Retrieved content is injected into prompt templates, enabling the LLM to generate contextually grounded and personalised responses.

G. Cognitive Reframing Module:

This module facilitates restructuring of negative automatic thoughts using CBT techniques. It identifies the user's thought, generates a balanced alternative perspective, provides emotional validation, and suggests actionable improvements. Relevant examples are retrieved from the knowledge base to reinforce learning.

H. Coping Strategy Module:

The coping module supports structured problem-solving through steps: defining the issue, exploring solutions, selecting actionable steps, and reflecting on outcomes. If retrieval is unavailable, heuristic templates ensure consistent guidance.

I. Grounding Module:

Designed for emotional regulation during distress, this module identifies the user's emotional state and retrieves suitable grounding techniques. It presents one simple, actionable exercise with step-by-step instructions to help reduce emotional intensity and restore focus.

J. Conversation Memory Management:

A session-based memory mechanism maintains contextual continuity and efficiency. It tracks module states (e.g., FIDO/HCB progress), stores recent dialogue in a sliding window, and preserves structured inputs for downstream processing—ensuring coherent long-term interactions without excessive memory usage.

K. Safety and Ethical Considerations:

The system incorporates multiple safeguards to ensure responsible use:

LLM-based crisis detection for high-risk inputs (e.g., self-harm).

Immediate escalation with supportive responses and helpline resources.

Non-diagnostic design to avoid clinical claims.

Encouragement of professional help when necessary.

L. Implementation Details:

The system is implemented in Python, integrating backend and machine learning components. Key technologies include:

ChromaDB for vector storage and retrieval

LangChain for orchestration and prompt management

SentenceTransformers (BAAI/bge-small-en-v1.5) for embeddings

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, May 2026 Impact Factor: 8.2

Google Gemini API for reasoning and language generation
FastAPI powers the backend APIs for routing, session handling, and execution, while React provides a responsive frontend interface. Streamlit is used during development for rapid prototyping and testing.

M. Workflow Summary:
User Input → Crisis Detection → Intent Routing (LLM-based) → Module Execution → Response Generation (LLM + RAG) → Session Update → Output to User

IV. EVALUATION AND RESULTS

In accordance with prior approaches, the system was evaluated using simulated multi-turn dialogues representing common mental health scenarios [4, 5]. Key evaluation metrics included latency, coherence, safety precision, and qualitative response quality. Responses were generated within acceptable latency, enabling smooth and uninterrupted interaction. Coherence was assessed across multiple turns to address the well-known issue of fragmented dialogue in chatbot systems [2]. Safety precision focused on accurately detecting and appropriately handling distress, aiming to reduce risks associated with LLM inconsistency and unsafe outputs [1, 4]. Qualitative examples demonstrated more structured and helpful responses, aligning with existing research suggesting that hybrid, framework-guided systems improve reliability [7, 8]. Consistent with prior work emphasising cautious interpretation of AI-based mental health tools, this evaluation does not make clinical claims [4, 10].

V. CONCLUSION AND FUTURE WORK

To address the limitations of existing conversational systems—particularly in terms of coherence, safety, and reliability—this work presents a hybrid, modular mental health chatbot that integrates structured therapeutic frameworks, retrieval-based grounding, and LLM-driven orchestration [1, 4, 7]. Looking ahead, future work should explore multimodal sensing (e.g., voice and physiological signals) to better capture and interpret user state, as suggested in broader AI healthcare research [9]. Establishing real-world effectiveness and safety will require rigorous clinical trials beyond simulated evaluations [4, 10]. Further improvements include developing hybrid AI–human systems that combine automated support with expert oversight, as well as enhancing personalisation to enable adaptive responses based on user history and preferences. These directions are essential for building responsible, effective, and scalable mental health support systems.

VI. ACKNOWLEDGMENT

The authors thank Prof. Kopal Gangrade and the Department of Computer Science, SCTR's PICT, for their continued support, guidance, and resources throughout this research.

REFERENCES

[1] H. Li, R. Zhang, Y. C. Lee et al., "Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being," *npj Digital Medicine*, vol. 6, p. 236, 2023. https://doi.org/10.1038/s41746-023-00979-5

[2] A. Yuan, E. Garcia Colato, B. Pescosolido, H. Song, and S. Samtani, "Improving workplace well-being in modern organizations: A review of large language model-based mental health chatbots," *ACM Transactions on Management Information Systems*, vol. 16, no. 1, Art. 3, 2025. https://doi.org/10.1145/3701041

[3] E. Kleinau et al., "Effectiveness of a chatbot in improving the mental wellbeing of health workers in Malawi during the COVID-19 pandemic: A randomized, controlled trial," *PLOS ONE*, vol. 19, no. 5, pp. 1–33, 2024. https://doi.org/10.1371/journal.pone.0303370

[4] Y. Wang et al., "Evaluating an LLM-powered chatbot for cognitive restructuring: Insights from mental health professionals," *arXiv preprint arXiv:2501.15599*, 2025. https://doi.org/10.48550/arXiv.2501.15599

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, May 2026 Impact Factor: 8.2

[5] F. O. Kuhlmeier et al., "Combining artificial users and psychotherapist assessment to evaluate large language model-based mental health chatbots," *arXiv preprint arXiv:2503.21540*, 2025. https://doi.org/10.48550/arXiv.2503.21540

[6] X. Fan, L. Yang, X. Wang, D. Lyu, and H. Chen, "Constructing a knowledge-guided mental health chatbot with LLMs," in *Proceedings of the 16th Asian Conference on Machine Learning (ACML)*, vol. 260, pp. 287–302, 2025. https://proceedings.mlr.press/v260/fan25a.html

[7] J. C. L. Chow and K. Li, "Large language models in medical chatbots: Opportunities, challenges, and the need to address AI risks," *Information*, vol. 16, no. 7, Art. 549, 2025. https://doi.org/10.3390/info16070549

[8] A. Martins, A. Londral, I. L. Nunes, and L. V. Lapão, "Unlocking human-like conversations: Scoping review of automation techniques for personalized healthcare interventions using conversational agents," *International Journal of Medical Informatics*, vol. 185, p. 105385, 2024. https://doi.org/10.1016/j.ijmedinf.2024.105385

[9] S. Montagna et al., "Privacy-preserving LLM-based chatbots for hypertensive patient self-management," *Smart Health*, vol. 36, p. 100552, 2025. https://doi.org/10.1016/j.smhl.2025.100552

[10] J. Li et al., "Chatbot-delivered interventions for improving mental health among young people: A systematic review and meta-analysis," *Worldviews on Evidence-Based Nursing*, vol. 22, no. 4, e70059, 2025. https://doi.org/10.1111/wvn.70059

[11] Y. Feng et al., "Effectiveness of AI-driven conversational agents in improving mental health among young people: Systematic review and meta-analysis," *Journal of Medical Internet Research*, vol. 27, e69639, 2025. https://doi.org/10.2196/69639

[12] X. Bai et al., "Application of AI chatbot in responding to asynchronous text-based messages from patients with cancer: Comparative study," *Journal of Medical Internet Research*, vol. 27, e67462, 2025. https://doi.org/10.2196/67462

[13] 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, e19, 2017. https://doi.org/10.2196/mental.7785