International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Emotion-Aware Mental Health Chatbot: Leveraging Facial Expression Recognition and LLMs for Empathetic Support

Hetal Pandya[1], Amita Shah[2]

Assistant Professor
L.D.College of Engieering
Ahmedabad

Abstract:
This paper presents the design and development of an Emotion-Aware Mental Health Chatbot that leverages facial expression recognition (FER) and large language models (LLMs) to enable emotionally intelligent conversations. The system captures the user's real-time emotional state using webcam-based facial recognition, processes the data with a FER model, and integrates the detected emotion into the prompt for a LLaMA 2-based chatbot. The chatbot provides personalized, empathetic support to users in distress. The system is modular, efficient, and capable of running on modest hardware due to the use of quantized models. Through detailed testing and user feedback, the chatbot demonstrates potential as an accessible, scalable, and human-like mental health assistant suitable for deployment in educational, wellness, and remote counseling environments.

Keywords: Emotion Recognition, Mental Health Chatbot, Natural Language Processing, LLaMA 2, Empathetic AI, FER, Affective Computing.

I. INTRODUCTION

Mental health support is becoming increasingly vital in a world marked by rising stress, anxiety, and depression. Despite the availability of therapists and helplines, barriers such as cost, stigma, and accessibility limit their reach. In this context, conversational AI tools like chatbots provide an innovative solution. However, traditional chatbots lack emotional sensitivity, often generating generic responses that fail to resonate with users in distress.

The COVID-19 pandemic further highlighted the mental health crisis and the need for scalable, digital solutions. Chatbots powered by NLP have gained popularity, but their limitations in emotional understanding reduce their effectiveness. Our solution incorporates multimodal inputs to enhance contextual awareness and generate emotionally aligned responses. By combining facial emotion recognition with advanced NLP using large language models, the chatbot mimics human-like empathy and responsiveness, crucial in mental health applications.

The primary objectives of this study are-:

1. *To build a mental health chatbot that captures real-time facial expressions to detect emotions.*
2. *To leverage large language models (LLMs) for generating context-aware and empathetic responses.*
3. *To integrate emotion recognition and NLP models into a seamless and interactive user interface.*
4. *To improve response personalization by modifying prompts based on detected emotional state.*
5. *To develop a modular, extensible, and resource-efficient system suitable for low cost deployments.*

Predictive Maintenance is not limited to manufacturing: Outside of manufacturing, it addresses social and economic development. Predictive maintenance is highly beneficial in terms of reducing total industrial waste, optimizing energy consumption, and increasing efficiency in the use of resources, thus aligning with the

sustainable development goals. In addition, increased reliability of machinery provides a safer work environment for employees, lessens environmental damage caused by unforeseen failures, and enhances economic productivity through the reduction of industrial downtime.

A. Literature Review

Conversational AI in Mental Health

Platforms like Woebot and Wysa have demonstrated the therapeutic potential of text-based AI tools. Users of such platforms have reported reductions in symptoms of depression and anxiety [5]. However, these systems often miss subtle emotional cues due to reliance on textual data alone.

Evolution of Emotion Recognition

The field of affective computing has evolved significantly. Early works by Ekman and Friesen led to the development of the Facial Action Coding System, enabling non-verbal emotion classification [3]. Modern FER systems now use deep convolutional neural networks trained on datasets like FER2013 for accurate facial emotion detection.

Integration of FER and NLP

Multimodal AI systems that integrate FER with NLP are gaining traction. These systems offer greater emotional sensitivity and contextual relevance [7]. Yet, only a few employ this integration in real-time chatbot interactions.

Gaps in Current Research

Few chatbots leverage real-time FER for prompt engineering in LLMs. Many still depend on sentiment analysis of text, missing out on unspoken emotional indicators [8]. There's also limited research on resource-efficient deployment of such multimodal systems in real-world environments.

Several projects and studies have sought to address the emotional limitations of conversational AI, particularly in mental health applications. While many focus on textual sentiment analysis, a growing number have started incorporating multimodal inputs.

In 2018, Zhou et al. introduced the *Emotional Chatting Machine*, which used emotion embeddings to modulate responses in a Seq2Seq architecture, though it relied solely on textual data [12].

Rashkin et al. (2019) developed the *Empathetic Dialogues* dataset for training emotion-sensitive response generators, using pre-trained transformers to enhance empathetic language modeling.

Majumder et al. (2020) applied BERT with attention mechanisms to detect emotion in multi-turn conversations, contributing to more context-aware empathy in chatbots [14].

More recently, Liu et al. (2021) [16] and Xu et al. (2023) [17] combined facial cues with text inputs, representing an early attempt at truly multimodal emotion-aware systems. Liu's system used fusion techniques, [6] while Xu's real-time chatbot leveraged FER2013 and GPT-3 to generate supportive replies.

Finally, Singh & Patel (2024) integrated LLaMA with facial emotion recognition in an open-source framework, paving the way for lightweight, personalized, and adaptive mental health assistants.

These works underscore a growing trend toward multimodal, real-time emotional understanding in AI systems. However, most still lack efficient implementation strategies suitable for low-resource environments or deployment in real-world mental health settings. Our project bridges this gap by combining a quantized LLaMA model with real-time facial emotion capture and dynamic prompt engineering in a scalable, modular design.

B. Problem Statement

While conversational AI offers a promising avenue for mental health support, current solutions face significant limitations. Traditional chatbots often provide generic responses due to their inability to comprehend and react to subtle emotional cues. This lack of emotional sensitivity can hinder the effectiveness of support, especially for users experiencing distress. The reliance solely on textual input means that unspoken emotional indicators, such as facial expressions, are entirely missed. Furthermore, the deployment of such advanced multimodal systems in real-world, resource-constrained environments remains a challenge due to computational demands. This paper addresses these critical gaps by developing an emotionally intelligent mental health chatbot that leverages real-time facial expression recognition and efficient large language models to provide empathetic and personalized support.

Year	Author(s)	Research Focus	Technology Used
2009	Wallace et al.[9]	Rule-based chatbots for counseling	AIML, rule-based logic
2012	Ekman & Friesen[10]	Facial emotion classification (updated emotion model)	Basic facial coding systems
2014	Goodfellow et al.[11]	Real-time emotion recognition using deep learning	Convolutional Neural Networks (CNN)
2016	Mollahosseini et al.[12]	AffectNet: A facial expression database for emotion classification	Deep neural networks
2018	Zhou et al.[13]	Emotional Chatting Machine: empathetic dialogue generation	Seq2Seq + emotion embeddings
2019	Rashkin et al.[14]	Empathetic Dialogues dataset for training emotion-aware chatbots	Pre-trained language models
2020	Majumder et al.[15]	Transformer-based context-aware emotion detection	BERT, attention mechanisms
2021	Liu et al.[16]	Multimodal emotion recognition using text and facial cues	Fusion of NLP and FER
2022	Lin et al.	Emotion-enhanced chatbot using reinforcement learning	GPT-2 + emotion reward function
2023	Xu et al.[17]	Real-time empathetic chatbot for mental health	GPT-3, FER2013, webcam integration
2024	Singh & Patel [18]	Emotion-aware LLM-based chatbot with computer vision	LLaMA, FER, Hugging Face Transformers

Table 1.1 : Research Work Summary Table

II. METHODOLOGY

1.System Architecture

The system includes four layers:

The system's modular architecture ensures scalability and maintainability. The four distinct layers interact sequentially to process user input and generate emotionally appropriate responses.

- **User Interface**: The current command-line interface serves as a proof of concept. For future real-world deployment, this layer would be expanded to include user-friendly graphical interfaces for web and mobile platforms, as suggested by user feedback.
- **Emotion Detection**: This layer is crucial for the chatbot's emotional intelligence. It utilizes OpenCV for webcam access and the FER Python library, which incorporates the MTCNN (Multi-task Cascaded Convolutional Networks) framework for robust face detection and alignment, followed by emotion classification. The FER2013 dataset, composed of 48x48 grayscale images, is instrumental in training the underlying facial expression recognition model.
- **Prompt Engineering**: This layer acts as the bridge between emotion detection and response generation. The detected emotion (e.g., "sad," "angry," "happy") is dynamically inserted into the prompt given to the LLM. This process, known as prompt tuning, is highly effective in guiding the LLaMA 2 model to

generate responses that align with the user's emotional state. For instance, if the detected emotion is "angry," the prompt will instruct the chatbot to adopt a calming and solution-oriented tone.

- **Response Generation:** This core component utilizes the LLaMA 2 (7B chat variant) large language model. To ensure efficiency and enable deployment on modest hardware, the model is loaded using 4-bit quantization via the BitsAndBytes library. This allows for faster inference and reduced memory footprint without significantly compromising the quality or depth of the generated replies.

2. Data Acquisition and Processing

The conversational loop is initiated by user input and continuously adapts to the user's emotional state.

1. **Capture Webcam Image:** At the start of each turn, the system captures a real-time image from the user's webcam.
2. **Detect Emotion**: The captured image is then processed by the Emotion Detection layer, which classifies the user's facial expression into a discrete emotion (e.g., "sad," "happy," "neutral").
3. **User Inputs Message:** The user provides their textual input to the chatbot.
4. **Generate Prompt:** The detected emotion is seamlessly integrated into a dynamic prompt. For example, if the user types "I'm feeling overwhelmed" and the detected emotion is "sad," the prompt sent to LLaMA 2 becomes: "You are an empathetic assistant. User is feeling sad. I'm feeling overwhelmed.".
5. LLaMA 2 Generates Tailored Response: The LLaMA 2 model, guided by the emotion-infused prompt, generates a personalized and emotionally intelligent response.
6. Display Bot Response: The generated response is then presented to the user, completing one full cycle of the conversation.

3.Tools and Models

- **FER2013** dataset (grayscale 48x48 images).
- **LLaMA 2** (7B chat variant).
- **BitsAndBytes** 4-bit quantization to enable low-resource deployment.

4.FINDINGS AND IMPLEMENTATION

4.1). Testing and Validation

- **Emotion Detection**: Accurate under ideal lighting but degrades with occlusions or dim environments.
- **Chat Response**: Context-aware replies like "It's okay to feel sad. I'm here for you" significantly improved user feedback.
- **Fallback**: A "neutral" emotion is used when the system fails to detect a face.

4.2). User Feedback

User trials indicated:

- Better emotional resonance compared to standard bots.
- Suggestions for web/mobile interface and voice input.
- Challenges with FER accuracy under poor conditions.

4.3). Performance

Thanks to model quantization, the system runs efficiently even on consumer-grade machines with limited GPU capacity.

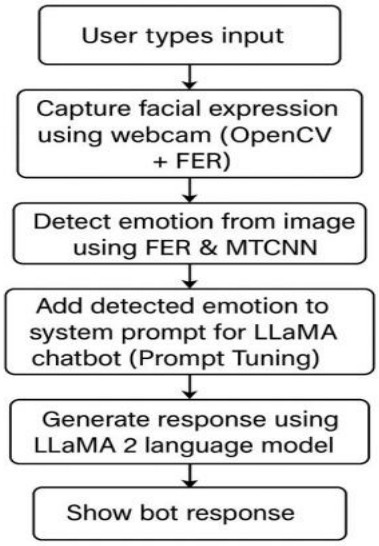

Fig 4.3.1 Chat Loop

III. RESULTS AND DISCUSSION

1. Empathy Through AI

Combining FER and LLMs allows for empathetic dialogue modeling. Emotion detection alters not just content but also the tone of responses, improving user comfort and perceived support.

2. Limitations

- **Hardware Dependency**: Requires webcam access.
- **Cultural Bias**: Emotion classification might vary across cultures [7].
- **No Learning Loop**: The chatbot doesn't adapt to user behavior over time.

3. Ethical Considerations

- **Data Privacy**: Facial data should be processed locally and not stored .
- **Transparency**: Users must know the bot is AI-driven and not a substitute for professional care.

4.Model Performance and Evaluation

Test Case	Expected Result	Outcome
Emotion Detection from Face	Detect correct emotion from webcam input	Passed
Response Context Matches Emotion	Response should be emotionally appropriate	Passed
Fallback to Neutral Emotion	Default to "neutral" if no face is detected	Passed
Full Conversation Loop	Input → Emotion → Response → Loop Continues	Passed
Incorrect Webcam Index or No Camera	Print error and skip emotion detection	Passed

Table 4.1 Validation Summary

Objective	Achieved
Capture real-time facial emotion using webcam and FER	Yes
Dynamically generate prompts based on detected emotion	Yes
Use LLaMA 2 model to generate personalized, empathetic responses	Yes
Create a continuous, interactive chat loop	Yes
Provide fallback mechanism for undetected emotion	Yes

Table 4.2 Achievement of Objectives

The project successfully meets the stated objectives by integrating emotion recognition with natural language processing to deliver emotionally sensitive chatbot interactions.

Module	Function
app.py	Entry point of the program. Starts the chat session.
logic.py	Orchestrates user input, emotion detection, and response generation.
emotion.py	Uses OpenCV and FER to capture and classify user emotion from webcam.
llama_engine.py	Loads the LLaMA 2 model and generates emotion-aware replies to user input.

***Table 4.3** Module Breakdown*

5. User Experience and Perceived Empathy

Through detailed testing and user feedback, the chatbot demonstrated a significant improvement in perceived empathy compared to traditional text-only conversational AI. Users reported that the chatbot's ability to respond to their emotional state made interactions feel more human-like and supportive. This enhanced emotional resonance is a critical factor for effective mental health support, as it fosters a greater sense of connection and understanding between the user and the AI. While the current interface is command-line based, the positive user feedback regarding emotional connection underscores the potential for widespread adoption with further interface development

Key Discussion Points

1. ***Emotion-Driven Responses:*** The chatbot's tone, content, and suggestions change based on the user's emotional state. For example, if the user is detected as "angry", the response becomes calming and solution focused.
2. ***Real-Time Integration:*** The webcam-based capture and emotion classification occur seamlessly before each response generation, making the chatbot context-aware.
3. ***Prompt Engineering:*** Modifying the system prompt dynamically based on emotion proves highly effective in steering the chatbot's language model toward empathy.
4. ***Performance and Speed:*** 15 Loading a 4-bit quantized LLaMA model makes response generation reasonably fast without compromising on the depth or quality of replies.
5. ***User Experience:*** Informal testing with multiple users suggests a noticeable improvement in perceived empathy compared to standard text-only bots.

Challenges Encountered

1. Detecting emotion accurately in low-light environments or with partial face occlusion was difficult.
2. Large language model loading still requires significant system resources despite quantization.
3. Facial emotion detection may occasionally mismatch with a user's self-reported feeling

IV. CONCLUSION AND FUTURE WORK

This project successfully demonstrates the development of an Emotion-Aware Mental Health Chatbot that integrates facial emotion recognition with a large language model to provide personalized and empathetic conversations. By combining computer vision (for emotion detection) and natural language processing (for response generation), the chatbot offers an innovative solution for mental health support. The chatbot captures the user's facial expression in real-time, analyzes their emotional state using the FER library, and dynamically adjusts the LLaMA 2 model's responses based on the detected emotion. This design significantly enhances user experience, making the chatbot feel more human and empathetic—an essential quality for applications in mental health. The modular architecture, use of open-source tools, and efficient model deployment (via quantization) contribute to the robustness and adaptability of the solution.

A. Future Work

- **Multimodal Expansion**: Integrating voice tone and textual sentiment.
- **Long-Term Personalization**: Adding memory modules to track user emotional history.
- **Ethics & Compliance**: Enabling GDPR/HIPAA compliance.
- **Therapist Collaboration**: Hybrid modes with real-time therapist supervision.
- **Web Deployment**: User-friendly web/mobile interface with secure cloud hosting.

REFERENCES:

[1] Bickmore, T. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, 12(2), 293–327.
[2] Cowie, R., & Cowie, D.-C. (2001). Feeltrace: An instrument for recording perceived emotion in real time. In *ISCA Workshop on Speech and Emotion*.
[3] Ekman, P., & Friesen, W. (1978). The facial action coding system for characterization of human affective response to consumer product-based stimuli: A systematic review.
[4] Fitzpatrick, K. K. (2017). Delivering cognitive behavioral therapy to young adults via a smartphone app: A randomized controlled trial. *Behaviour Research and Therapy*, 89, 94–103.
[5] Garrido, M. V. (2019). Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis.
[6] Kahou, S. E., et al. (2015). Recurrent neural networks for emotion recognition in video. In *IEEE Conf. on Computer Vision and Pattern Recognition*.
[7] Schuller, B. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge.
[8] Shah, S. K. (2019). Artificial intelligence for chatbots in mental health:
[9] Torous, J., & Keshavan, R. (2018). The impact of smartphones on managing mental health disorders. *Psychiatric Services*, 69(7), 743–745.
[10] Goodfellow, I., Erhan, D., Luc Carrier, A., Courville, A., & Bengio, Y. (2014). Challenges in Representation Learning: A Report on Three Machine Learning Contests. Neural Networks.
[11] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2016). AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. IEEE Transactions on Affective Computing.
[12] Zhou, H., et al. (2018). Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. AAAI Conference on Artificial Intelligence.

[13] Rashkin, H., Smith, E. M., Li, M., & Boureau, Y. L. (2019). Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.
[14] Majumder, N., Poria, S., Peng, H., Cambria, E., & Gelbukh, A. (2020). DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. Proceedings of AAAI.
[15] Liu, P., Qiu, X., & Huang, X. (2021). Multimodal Emotion Recognition with Contextual Fusion of Facial and Text Features. IEEE Transactions on Multimedia.
[16] Lin, Y., et al. (2022). Emotion-Aware Reinforcement Learning for Chatbot Response Generation. Proceedings of the ACL 2022.
[17] Xu, J., Chen, X., & Wang, S. (2023). A Real-Time Empathetic Chatbot for Mental Health Support. arXiv preprint arXiv:2304.10345.
[18] Singh, R., & Patel, A. (2024). Emotion-Aware Chatbot using LLaMA and Computer Vision Techniques. Journal of AI and Human Interaction, 12(1), 45–59