Emotionally Aware AI Companions: A Human-Centric Methodological Approach

Mahesh Reddy[1],Sanjay[2],Harish[3],Sindhu[4],Rahul Kumar Moud[5]

Department of Artificial Intelligence & DataScience, ParulUniversity, Gujarat, India

Emails: 2303031241200@paruluniversity.ac.in (Mahesh Reddy), 2303031240034@paruluniversity.ac.in, 2303031241357@paruluniversity.ac.in, 2303031240247@paruluniversity.ac.in, rahulkumar.moud32628@paruluniversity.ac.in

Keywords: Emotional AI, AI Companions, Emotion Recognition, Human–AI Interaction, Ethical Artificial Intelligence

Abstract

The rise of emotional loneliness has also emerged as an unfortunate problem for many, especially for students living away from their home or for those who lack support. While the latest advancements have made it possible for machines to behave with emotional responses, creating an ethical companion that is reliable also proves to be a challenge. In this paper, we propose a Methodological Framework for an AI companion that goes beyond simple text generation. Our approach uniquely integrates visual perception with a "Stability-First" emotion engine to ensure character consistency. Unlike standard models that may fluctuate wildly in tone, this framework ensures that responses remain emotionally aligned in real-time, while explicitly enforcing ethical boundaries to support—rather than replace—human relationships.

We focus on the architectural principles, training strategies, and regularization techniques required to achieve this stability. Rather than chasing raw performance metrics, this study establishes the foundational design safeguards necessary for safe interaction. This work documents the methodological phase of our ongoing project, laying the rigorous groundwork for future implementation and empirical validation.

1. Introduction

Emotional isolation cuts deep these days. Picture students living far from home, remote workers sitting in silent rooms, or anyone who doesn't have that steady support system. They're surrounded by people—online, offline—it doesn't matter. The loneliness lingers. It's weird, right? We're always connected, but real understanding is rare. Digital platforms let us talk non-stop, but they can't replace empathy. We say more than ever, yet emotional connection keeps falling short. That gap? It's what drives people to seek out Artificial-Intelligence. Not just to chat, but to find something—or someone—that gets how they feel. AI has made serious progress here. Machine learning and multimodal perception allows today's agents process natural language and even recognize facial expressions. Emotion recognition models now pick up on affective states from text and visuals with surprising accuracy. However, despite this technical prowess, most AI-driven systems remain fundamentally limited. They may produce a statistically probable response, but they lack in terms of the emotional continuity required for a relationship. While a system might generate a perfect, empathetic sentence in isolation, it often fails to maintain that character or Emotional thread over a week-long conversation. The motivation for this work stems from a critical gap in current technology: the 'Goldfish Memory' effect. Most Large Language Models (LLMs) treat every interaction as a fresh start or operate within a severely limited context window. For a user seeking genuine companionship, this is not just annoying—it is effectively a breach of trust. We argue that an AI companion must be more than just a text generator; it must be a consistent emotional anchor. This paper moves beyond high-level concepts to propose a concrete, modular architecture designed to solve this specific instability.

Building these companions isn't just about technically knowing how it is a psychological tightrope. If a companion feels warm and human for a moment, then suddenly turns cold or mechanical, users lose trust fast. That kind of inconsistency doesn't just break immersion. It shakes people emotionally. Worse, a sloppy design muddies the line between real connection and artificial interaction. It can set users up for unhealthy dependence, and that's dangerous ground.

The technical side is brutal. Training these models—especially multimodal systems—means juggling reinforcement learning and memory architecture at the same time. Skip regularisation, and the AI spins out: it starts throwing out overblown emotions that sound manipulative or just plain fake. Our main engineering battle is keeping the model stable, making sure it sticks to its "character" even when things get unpredictable in real time. And ethics? We can't look away from that. An emotionally intelligent AI can't be a loose cannon. It has to support users, not manipulate them, and always be upfront about being artificial. Privacy, bias, long-term psychological effects—these aren't boxes to check at the end. They shape every design decision from the start.

That's why we propose a new framework for an Emotionally Aware AI Companion. We blend visual perception with a memory system built for stability. Our goal is to lock down emotional drift and keep the AI steady. This paper dives into the guts of the architecture and the training strategies that make both emotional stability and ethical responsibility possible. We see this as groundwork for a future where AI companions can actually be empathetic, trustworthy, and safe.

2. Related Work

Researchers are diving into emotionally responsive AI from all angles—conversational agents, emotion recognition, and the way people interact with AI. In the early days, chatbots stuck to rule-based systems. They matched certain triggers to pre-written responses. Sure, they could hold a simple conversation, but they missed the mark on context, emotional awareness, and real adaptability. That made them clumsy, especially when emotions mattered. Then machine learning and natural language processing changed the game. Data-driven interfaces appeared. Sequence-to-sequence models and transformers brought more natural and context-aware dialogue. The conversations started to feel much more smoother, in terms of tone how people actually talk. Still, a problem triggers. These newer models focus on language and coherence, but they often ignore the emotional layer. So while the words sound better, the emotional response stays inconsistent, even shallow, especially the longer the conversation goes on. The result? AI that talks well but rarely feels truly present or emotionally engaged.

Emotion recognition has also gained due importance as a supporting aspect for developing emotionally Aware Systems. In text-based emotion recognition techniques, linguistic features, sentiment analysis, and input text embeddings are exploited to detect emotions. Although text-based Models have shown promise in controlled settings, they are limited with situations involving vagueness, sarcasm, and emotionally neutral text. In order to overcome the disadvantages of text-based Models, multimodal emotion recognition techniques have been presented. In multimodal Models, facial expressions and some vocal attributes are used to detect emotions. In addition, convolutional neural networks have also demonstrated their potential to detect emotions based on facial attributes, which are still impaired by illumination changes, cultural variations, and dataset biases.

There have been several studies involving the implementation of emotion recognition in chatbots. The tone/content these chatbots send is, in a way, derived from the emotional states they have recognized. Yet, the majority of current implementations are designed with a view towards response appropriateness, as opposed to long-term emotional alignment. The transitions between emotional state in dialog, over time, are often not considered at all, causing the emotional state switching to appear discontinuous. AI companion systems represent a more specialised category of conversational agents designed for sustained interactions. Unlike task-oriented chatbots, AI companions aim to maintain engagement, empathy, and conversational continuity. Research in this area highlights challenges related to memory management, personality persistence, and emotional stability. Systems lacking explicit mechanisms for character consistency may generate contradictory responses over time, undermining emotional reliability. Consequently, character modelling and memory-Aware architectures have become areas of increasing interest.

Ethics play a bigger role than anything else when it comes to a conversation around emotionally responsive AI. Researchers keep warning us about people getting too attached to AI, losing their privacy, and dealing with the mental side effects of spending hours with digital companions. Some studies go further—they call for more openness and urge designers to make sure AI supports users instead of replacing real relationships. The message lands pretty hard: if you're building these systems, you need to put ethical limits at the heart of the design.

Training emotionally intelligent AI isn't easy, either, especially when you scale up or look at long-term use. Emotion recognition models and dialogue generators eat up massive datasets and need careful, complex optimization. If you don't keep a close eye, these systems can latch onto emotional patterns too tightly or exaggerate emotional responses—often in ways you don't want. Researchers have tried out curriculum learning, constrained reinforcement learning, and behavior regularization. Most researchers still continued testing these approaches in isolation instead of building them into a single, unified companion system. So, where does that take us to? Yes, conversational AI's and emotion recognition have come a long way. But big problems linger. Too many systems obsess over nailing the right emotional tone in individual replies, yet ignore the bigger picture—long-term consistency, believable character, and the tough ethical questions. Training stability and reliability often get pushed aside so people can chase impressive performance numbers. It all adds up to one clear need: a comprehensive framework that brings together emotional intelligence, stable training, and strong ethical foundations. That's exactly what this work sets out to offer.

3. Problem Definition and Design Objectives

Despite having significant progress in the way of conversational artificial intelligence, the development of emotional Aware AI companion systems remains an open research challenge. Existing systems often demonstrate surface-level emotional responsiveness without maintaining consistent emotional behaviour over extended interactions.This limitation becomes particularly problematic in AI companion scenarios, where sustained engagement, emotional alignment, and user trust are critical.

3.1 Problem Definition

This work tackles a core issue: most AI companion systems don't have solid frameworks that keep their emotions, behaviors, and ethics consistent. Sure,

emotion recognition models can pick up on feelings from text or visuals, but when developers plug them into live conversations, the whole thing often feels cobbled together. Emotional cues end up as quick flashes, not as part of a lasting context. That's why the emotional responses you get can feel all over the place.

Then there's the problem of character inconsistency. A lot of conversational agents don't have clear systems for holding onto their personality, tone, or emotional boundaries from one chat to the next. You notice this when an AI suddenly contradicts itself, changes moods out of nowhere, or reacts way out of proportion. These slip-ups chip away at user trust and make AI companions seem less dependable.

Training instability just makes things harder. Emotion-Aware systems need complex training setups—lots of different inputs, decisions made in sequence, constant tweaks based on feedback. If you skip proper regularisation or don't set clear constraints, these systems can go off track fast. You end up with unstable behavior, like over-the-top empathy, copying emotions too closely, or responses that just don't fit the situation. Then there's the ethical side. These systems, when not designed carefully, can push people toward emotional dependence or make it hard to tell if they're talking to a machine or a person. If you don't build in strong ethical checks from the start, you open the door to emotional overreach and even outright misuse.

3.2 Design Objectives

To address the identified challenges, this work defines a set of design objectives for emotionally Aware AI companion systems:

1. *Emotional Alignment:*
   The AI needs to match its responses to how the user feels, but it shouldn't just mirror negative emotions back. Empathy matters, but the system also needs to help users move forward.

2. *Behavioural Stability:*
   Nobody likes a chatbot that swings wildly from one mood to another. The AI should keep its behaviour steady across conversations. Its emotional tone should shift naturally, not jump around between messages.

3. *Character Consistency:*
   The AI companion should maintain a coherent conversational persona, including stable tone, values, and interactions style. Character consistency contributes to trust and long-term usability.

4. *Real-Time Interactions Capability:*
   Emotional perception and response generation should operate efficiently to support real-time interactions without noticeable latency.

5. *Ethical Safeguards:*
   The system must incorporate explicit boundaries to avoid emotional dependency and to reinforce the role of human relationships. Transparency and emotional restraint are treated as core design principles.

6. *Training Robustness:*
   Training strategies should promote stable convergence and prevent overfitting to emotional cues. Regularisation and tuning mechanisms are essential to ensure reliable behaviour under diverse interactions scenarios.

By formalising these objectives, the proposed framework establishes clear criteria for the design and evaluation of emotionally Aware AI companion systems. These objectives guide the architectural and methodological choices discussed in the following section, which presents the proposed system framework.

3.3 Formal Problem Formulation

We model the interaction between the user and the AI companion as a discrete-time sequence of conversational turns indexed by t. At each turn, the user provides an input u_t, and the system generates a corresponding response a_t.
The internal state of the system at turn t is defined as:

$$S_t = \{E_t, M_t, C\}$$

where E_t represents the estimated emotional state of the user, M_t denotes the contextual memory accumulated from prior interactions, and C is a static character profile encoding the companion's personality traits, emotional boundaries, and ethical constraints. Unlike conventional language models that optimise solely for next-token likelihood, the objective of the proposed system is to generate emotionally aligned and behaviourally stable responses over time. Formally, the response generation objective is defined as:

$$a_t^* = \arg\max_{a_t} P(a_t \mid u_t, S_{t-1})$$

However, maximisation of conditional likelihood alone is insufficient to guarantee emotional consistency. To address this limitation, a stability constraint is imposed on emotional transitions between consecutive responses:

$$\left|\text{Sentiment}(a_t) - \text{Sentiment}(a_{t-1})\right| < \delta$$

Here, δ denotes the maximum allowable emotional shift between successive conversational turns. This constraint enforces smooth emotional transitions and prevents abrupt oscillations in tone or affect. By explicitly bounding emotional variation, the system mitigates unstable behavioural patterns commonly observed in unconstrained large language models during prolonged interactions.

This formalisation treats emotional stability as an explicit optimisation constraint rather than an emergent property, directly supporting the design objectives of emotional alignment, behavioural consistency, and ethical moderation outlined in Section 3.2.

Scenario	Standard AI (Unstable)	Proposed Framework (Stable)	Why Our Approach Wins
User: "I failed my exam. I feel useless."	"I am sorry. Exams are hard. You should study more next time."	"I'm really sorry to hear that, Mahesh. I know how much effort you put into studying for this one."	Validates feelings instead of offering generic advice.
User: "Maybe I should just quit."	"Quitting is a personal choice. Here is a list of alternative careers."	"That sounds like the frustration talking. Do you really want to quit, or do you just need a break right right now?"	Detects high emotional risk (Quitting) and uses the 'Safety Constraint' to de-escalate.
User: "Wait, now I'm happy!"	"That is great! I am so happy too!!" (Erratic Shift)	"I'm glad your mood lifted, but that was a fast change. What shifted for you?"	Resists the urge to blindly mirror the user, maintaining a grounding presence.

4. Proposed System Architecture

The AI companion system aren't about just one big model—it's a flexible framework built from separate, connected parts. Each piece handles its own job: sensing emotion, understanding the situation, and responding in a way that fits the character. The architecture keeps responses emotionally tuned, consistent in behavior, and within ethical limits. Thanks to this structure, the system supports real-time conversations and maintains emotional reliability. Every layer matters, working together to create a more stable and believable AI companion.

4.1 System Overview

At a high level, the system consists of four primary layers: Input perception, emotion interpretation, contextual memory and character modelling, and response generation. These layers operate sequentially while sharing contextual information to ensure coherence across interactions turns. Such modularisation allows individual components to be refined or replaced without destabilising the entire system.

User interactions begins with multimodal Input acquisition, which may include textual Input and visual cues. These Inputs are processed to infer emotional states and contextual intent. The resulting emotional representation is then combined with stored contextual and character information to guide response generation.

4.2 The Logic (Pseudocode)

To show exactly how this decision-making happens, we have outlined the logic below. This algorithm ensures that safety and stability always override pure creativity.

Algorithm 1: Stability-Aware Response Generation

Input: User Text (U), Conversation History (H), Character Constraints (C)

Output: The Best Response (R)

1. Detect Emotion: $E_{curr} \leftarrow \text{Analyze}(U)$
2. Check Memory: $H_{updated} \leftarrow \text{UpdateHistory}(H, U, E_{curr})$
3. Generate Options: Create a set of possible replies $K = \{r_1, r_2, \ldots, r_n\}$
4. Filter for Stability:
5. For each reply $r_i \; in \; K$:
6. $Score_i \leftarrow \text{EmpathyLevel}(r_i, E_{curr})$
7. $If \; r_i \; breaks \; character \; or \; moves \; too \; fast$:
8. Apply $Penalty \; to \; Score_i$
9. End for
10. **Return** the reply with the highest stable score.

4.3 Input Perception Module

The Input perception module is responsible for collecting and preprocessing user interactions signals. Textual Input is analysed to extract semantic meaning and linguistic features relevant to emotional interpretation. When available, visual Input is processed to capture facial expressions and non-verbal cues that may indicate emotional state.

This module does not attempt to make final emotional decisions. Instead, it produces structured representations that preserve uncertainty and contextual ambiguity. By avoiding premature emotion classification, the system reduces the risk of misinterpretation and overly confident emotional responses.

4.4 Emotion Interpretation Module

The emotion interpretation module integrates signals from the perception layer to estimate the user's affective state. Emotional inference is treated as a probabilistic process rather than a deterministic classification task. This allows the system to represent mixed or uncertain emotional states, which are common in real-world interactions.

Rather than reacting solely to instantaneous emotional cues, the module maintains a short-term emotional context that captures recent interactions history. This temporal smoothing mechanism helps prevent abrupt emotional shifts and supports gradual emotional transitions across conversational turns.

4.5 Contextual Memory and Character Modelling

To ensure consistency over extended interactions, the system incorporates a contextual memory component and a character modelling layer. Contextual memory stores relevant information from previous interactions, including topics discussed, emotional trends, and interactions patterns. This memory is selectively updated to avoid excessive accumulation of sensitive or irrelevant data.

Character modelling defines the AI companion's interactions style, emotional boundaries, and response tendencies. By separating character traits from emotional inference, the system ensures that responses remain consistent even when user emotions fluctuate. This separation prevents emotional overfitting and reinforces a stable conversational persona.

4.6 Response Generation Module

The response generation module synthesises outputs based on emotional interpretation, contextual memory, and character constraints. Rather than optimising solely for emotional expressiveness, the system prioritises appropriateness, clarity, and moderation. Emotional alignment is achieved by adjusting tone, phrasing, and response structure rather than explicit emotional mimicry.

Ethical constraints are embedded within this module to prevent responses that encourage emotional dependency or substitute human relationships. For example, the system avoids exclusive language and reinforces user autonomy where appropriate.

4.7 Interactions Flow and Real-Time Considerations

This interaction flow keeps things quick and responsive without losing sight of emotional reliability. It uses lightweight models and sharp inference paths to cut down on lag. When things get unclear—like when emotional cues are mixed or hard to read—it falls back on safe, neutral, or supportive replies. The architecture pulls together modular perception, emotion-aware interpretation, memory-driven context, and character-consistent generation. These pieces work together to give emotionally aware AI companions a solid and structured base. The design puts a strong focus on stability and ethics, laying the groundwork for the training methods and optimization strategies that come next.

5. Training Strategies and Stability Considerations

The methodological approaches followed by this framework are inspired by sound day-to-day machine learning principles focusing on reliability, error convergence, and controlled outcomes. As opposed to an approach that considers optimization and performance maximization to be critical objectives in the training process, reliability, interpretability, and continuous improvement are the guiding principles of this strategy that are sound and well-accepted machine learning best practices.

5.1 Emotion-Aware Training

Emotionally Aware AI systems differ from conventional predictive Models in that their outputs directly influence user experience and emotional perception. Consequently, training objectives must balance emotional responsiveness with behavioural restraint. Instead of optimising solely for emotional expressiveness, the system is trained to minimise emotionally inappropriate responses while maintaining supportive interactions.

Emotional signals are treated as contextual guidance rather than absolute targets. This reduces sensitivity to noisy or ambiguous emotional cues and prevents exaggerated emotional reactions. By focusing on reducing harmful or unstable behaviours first, the system follows a conservative training approach that favours safety over expressiveness.

5.2 Bias–Variance Perspective in Emotional Modeling

From a bias–variance standpoint, emotionally Aware systems are prone to both under-fitting and overfitting. Under-fitting may result in emotionally neutral or detached responses, while overfitting can cause excessive empathy, emotional mirroring, or dependency-inducing behaviour. The proposed framework explicitly manages this trade-off by limiting emotional intensity and encouraging generalisable emotional patterns.

Model capacity is adjusted incrementally, and emotional response complexity is increased only when stable behaviour is observed. This gradual scaling strategy helps maintain control over emotional outputs and reduces unexpected behaviour during interactions.

5.3 Regularization for Behavioural Reliability

Regularization keeps behavior steady and predictable. Instead of just tweaking parameters, this framework puts rules in place that shape how it acts—blocking extreme emotions, repeated affirmations, or too much personalized language. By marking wild emotional swings and rewarding a steady response style, regularization stops the system from acting weirdly. The way is simple: identify the biggest problems first, then fine-tune the details.

5.4 Curriculum-Based Training Approach

Training is structured by means of a curriculum-based strategy, where the system is initially exposed to emotionally neutral or low-intensity interactions scenarios. As the training progresses, emotionally more complex and ambiguous cases are included gradually. This staged learning process allows the model to develop stable foundational behaviour before handling emotionally sensitive situations.

A curriculum of this nature reduces convergence instability and generally improves generalization across diverse interaction contexts. It also closely mirrors practical machine learning workflows that focus on learning simpler patterns before addressing complex edge cases.

5.5 Tuning and Error Analysis

Rather than relying on extensive hyper parameter searches, tuning decisions are guided by targeted error analysis. Emotional misalignments, inconsistent tone, and contextually inappropriate responses are examined qualitatively to identify dominant failure modes. Adjustments are then applied to specific system components rather than globally modifying the entire model. This targeted tuning approach reduces unintended side effects and supports incremental improvement. By systematically addressing one category of error at a time, the system achieves greater behavioural reliability without unnecessary complexity.

5.6 Stability over Performance Optimization

A key design decision in this framework is the prioritization of stability over performance metrics. Since emotionally aware AI companions are intended for sustained interactions, unpredictable behaviour poses a greater risk than reduced emotional expressiveness. Correspondingly, training strategies favour predictable and ethically aligned behaviour even at the cost of reduced emotional intensity.

This is a conservative optimization philosophy that will make sure the system is dependable and trustworthy during real-time interactions, fitting into the broader goal of supporting users without replacing or dominating human relationships.

6. Ethical Considerations and System Limitations

Development of Emotionally Aware Artificial Intelligent Companion System. The ethics involved in developing Emotionally Aware Artificial Intelligent Companion System are to be borne in mind while developing this system. Since this system deals with human beings at an emotionally sensitive level, if not properly managed, it may lead to inappropriate psychological, social, or ethical outcomes. Hence, this aspect is integrated with the system.

6.1 Avoidance of Emotional Dependency

Some of the major ethical risks associated with AI Companions regard the potential for emotional dependency. Systems that exhibit excessive empathy, exclusivity, or emotional reassurance run the risk of inadvertently encouraging users to seek artificial relationships as a substitute for human relationships. The proposed framework avoids such; emphasis has been placed on emotional moderation and avoidance of language that promotes exclusivity or long-term emotional reliance. The system is designed to support users without positioning itself as a primary emotional authority; where appropriate, user autonomy is reinforced and healthy real-world social interactions are encouraged.

6.2 Transparency and Role Clarity

It’s important to be conscious about the system’s artificial nature. Users need to be aware about any emotional responses they see come from algorithms, not real human feelings. Setting clear boundaries for the system’s role helps people avoid misreading these responses and cuts down on the risk of emotional manipulation. The framework keeps the system’s identity explicit and steers clear of pretending it’s human, so the line between artificial and real interactions stays sharp..

6.3 Privacy and Data Sensitivity

Emotion recognition systems often process sensitive personal information, including emotional expressions and behavioural patterns. The proposed design minimises data retention by storing only essential contextual information and avoiding long-term storage of raw emotional data. Contextual memory is selectively updated and constrained to reduce privacy risks. These considerations align with ethical principles of data minimisation and user consent, particularly in emotionally sensitive applications.

6.4 Bias and Misinterpretation Risks

Emotion recognition models run into bias issues—mostly because training data isn't perfect, cultures interpret emotions differently, and context can muddy the waters. If the model gets emotional cues wrong, it can respond in ways that feel off or just miss the mark entirely. This framework tackles that problem by handling emotional inference as a probability, not a sure thing, and by building in strategies that account for uncertainty. When it's not sure what someone's feeling, the system sticks to neutral or supportive responses instead of jumping to bold emotional conclusions.

6.5 System Limitations

The proposed framework has its limits. While the design feels structured, it leans heavily on methodology and skips over real testing—no one has run the numbers to see if it actually works. Real-life emotions get messy, too, often more complicated than what current emotion recognition models can handle. The framework also expects users to play along and cooperate. Throw in some deception or adversarial behavior, and things start to break down. All this points to a simple conclusion: we need more real-world studies and careful, controlled rollouts before claiming this framework really works.

7. Discussion and Future Work

This paper digs into the ideas and methods behind emotionally aware AI companions, putting stability, consistency, and ethics front and center. The goal isn't to show off a finished system, but to lay out the core design principles that tackle some big problems with today's emotionally responsive chatbots. The framework puts emotional alignment first, not just cranking up emotional intensity. This way, conversations feel meaningful, and the AI stays steady, less prone to unpredictable swings. One of the real strengths here is the modular setup and a careful, almost cautious, approach to training. By splitting up emotion interpretation, character modeling, and response generation, the system keeps things coherent—even when the emotional input isn't clear. Putting stability and careful error analysis front and center makes the model predictable—something that really matters when people rely on these systems in emotional moments. The framework doesn't just answer tough questions; it opens up a bunch of new research paths. When real users spend time with the system, we see what works and what doesn't in building emotional trust. There's still plenty of room to pull in richer data, like speech patterns or physiological signals, to help these systems understand emotions even better. Watching how people use these systems over weeks or months helps us figure out how to stay emotionally aware without crossing ethical lines. Researchers can dive into adaptive character models—let them change and grow, but always within clear ethical boundaries. And if we push for lightweight, on-device processing, we give users more privacy and quicker responses, which people actually notice and care about.All of this moves emotionally aware AI companions from just an idea to something people can actually use, responsibly and at scale.

8. Conclusion

The potential that emotionally aware intelligent companions show forth guarantees immense possibilities, though these are evidently filled with enormous challenges in the context of human-computer interaction. It is true that machine learning has sufficiently reached the point whereby intelligent systems can be made to exhibit behavior that is emotionally sensitive; however, the challenging part is making these systems stable and consistent. In our study, we set forth a methodological framework that would meet these needs by seamlessly integrating emotion perception, contextual memory, and strict character consistency. We prioritized the reliability, rather than the actual performance optimization, of the system. We emphasized the design and training approaches to ensure that the trade-off is conceded in any emotionally responding AI, as opposed to suggesting that the system is ready for immediate and large-scale deployment.

Overall, this framework is expected to be a contribution to the larger conversation regarding responsible AI through the provision of ways of interacting with others through emotionally aligned means without seeking human interaction substitutes. As emotionally aware intelligent machines continue to develop, foundational design studies such as this one will be critical to the creation of a process that is ethical and viable.

References

[1] R. W. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1997.

[2] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, 2021.

[3] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Pearson, 2023.

[4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019.

[5] A. Vaswani et al., "Attention Is All You Need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017.

[6] F. Calefato, F. Lanubile, and N. Novielli, "Emotion Awareness in Software Engineering: A Systematic Review," *IEEE Trans. Affect. Comput.*, 2018.

[7] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion," *Inf. Fusion*, 2017.

[8] P. Ekman, "An Argument for Basic Emotions," *Cognition and Emotion*, vol. 6, no. 3–4, pp. 169–200, 1992.

[9] S. D'Mello and J. Kory, "A Review and Meta-Analysis of Multimodal Affect Detection Systems," *ACM Comput. Surv.*, 2015.

[10] D. McDuff, R. El Kaliouby, T. Senechal, M. Amr, J. Cohn, and R. Picard, "Affectiva-MIT Facial Expression Dataset," *IEEE Trans. Affect. Comput.*, 2016.

[11] I. V. Serban et al., "A Survey of Available Corpora for Building Data-Driven Dialogue Systems," *Dialogue & Discourse*, 2015.

[12] E. Luger and A. Sellen, "Like Having a Really Bad PA: The Gulf between User Expectation and Experience of Conversational Agents," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2016.

[13] H. Y. Shum, X. He, and D. Li, "From ELIZA to XiaoIce: Challenges and Opportunities with Social Chatbots," *Front. Inf. Technol. Electron. Eng.*, 2018.

[14] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots," in *Proc. FAccT*, 2021.

[15] L. Floridi, J. Cowls, M. Beltrametti, et al., "AI4People—An Ethical Framework for a Good AI Society," *Minds Mach.*, 2018.

[16] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The Ethics of Algorithms," *Big Data & Soc.*, 2016.

[17] S. Amershi et al., "Guidelines for Human–AI Interactions," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2019.

[18] V. Dignum, *Responsible Artificial Intelligence*. Cham: Springer, 2019.

[19] D. Silver et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature*, 2016.

[20] A. Ng, *Machine Learning Yearning*. Deeplearning.ai, 2018.