Master Thesis

Conversational AI for Emotion Regulation: Dual Analysis of Self-Report and Biometric Intervention Data

by

Kiara Shivani Kandhai
(kka105)

First Supervisor: Nimat Ullah
Daily Supervisor: Nimat Ullah
Second Reader: Michel Klein

July 18, 2025

Submitted in partial fulfillment of the requirements for
the VU degree of Master of Science in Artificial Intelligence

Conversational AI for Emotion Regulation: Synchronous Analysis of Intervention and Biometric Data

Kiara Shivani Kandhai

Vrije Universiteit Amsterdam, The Netherlands s.a.kandhai@student.vu.nl

Abstract. Current conversational agents for digital mental health focus on clinical interventions to alleviate psychiatric symptoms and use only self-report data. This approach overlooks two key opportunities: circumstantial support of emotion regulation (ER) with conversational AI for subclinical populations, and leveraging objective biometric data to gain understanding in physiological changes in users. The current thesis introduces *Aire*, a novel conversational agent that addresses real-time ER with dialogue grounded in the Extended Process Model of emotions. A connected wearable device collects biometric user data for post hoc analysis. This dual approach allows for a comprehensive view of ER processes. An exploratory pilot study with a one group, pre-test / post-test (N = 17) design examined chatbot interactions after negative affect induction. Physiological data of participants (HRV, HR, and Skin Temperature) were recorded on a Samsung Galaxy Watch 6. The interaction with Aire guided users through a process of identifying and regulating emotions using controlled LLM-powered dialogue. Results showed positive intervention outcomes, demonstrating a statistically significant increase in positive affect and a decrease in negative affect. Biometric analysis of physiological changes showed a decrease in the RMSSD component of HRV, indicating decreased parasympathetic activity. This counterintuitive finding points to ER as an active, cognitively engaging process rather than purely restful. Furthermore, reductions in minimum HR and increases in maximum HR were found, further encouraging a distinction between passive and active ER dynamics. Together, these findings highlight the psychophysiological dynamics of ER, and open up a path to improved digital mental health.

Keywords: Emotion Regulation · Conversational AI · Chatbot · Biometric Data · Wearable Sensors · Heart Rate Variability (HRV) · Arousal.

1 Introduction

The human emotional landscape guides us through everyday life, and having comprehensive awareness of emotional states provides us with insights that allow us to direct our thoughts and responses in a way that results in appropriate self-management, psychological resilience, and strong social connections [1].

Management of these emotional states is called Emotion Regulation (ER), and refers to the use of coping strategies to modulate emotional expressions and experiences [2–7]. Importantly, rather than being related to fixed traits, effective ER results from the integration of robust affective and cognitive skills. This suggests that these regulatory skills can be learned and trained to promote healthy functioning [1, 7–9].

An emerging challenge is the need for accessible real-time support systems for people who experience difficulties in applying these skills, particularly with respect to the management of negative affect in daily life. Conversational agents, also called '*chatbots*', show new, promising solutions from the technological field that provide consistent access to mental health support [10–16]. However, the development and implementation of current systems show a primary focus on severe psychiatric symptoms and crisis management, leaving a distinct gap for tools that can provide immediate support for members of subclinical populations (i.e., expressing complications that fall below the threshold of formal clinical diagnosis) that might not need clinical grade therapeutic intervention, but would rather benefit from circumstantial assistance.

Furthermore, development and analysis of these conversational agents rely on self-report measures, which depend on the user's ability to articulate complex feelings and accurately recognize their own internal states. With the increasing popularity of wearable devices that are capable of capturing physiological information of an individual, this technology can be leveraged to complement self-report designs with objective biometric measurements as representations of internal emotion dynamics. Analysis of metrics such as Heart Rate Variability (HRV), Heart Rate (HR), and skin temperature opens up a path to deeper understanding of physiological fluctuations as users move through affective states.

For these reasons, this thesis addresses the gaps in current research by introducing an exploratory pilot study that examines in-the-moment regulation interventions for sub-clinical use, as well as a post hoc analysis of physiological states during this interaction. The development of conversational agents and wearable technologies is flourishing, and exploratory research that analyzes the impact of these text-based interactions and the corresponding physiological responses creates a bridge to robust and better-informed mental health support.

Two research questions guide this objective:

1. How can a conversational agent be designed to provide emotion regulation support to users?
2. What patterns emerge from the subjective and physiological responses of user interactions with the conversational agent for emotion regulation?

Based on existing literature, the research questions are addressed by exploring the following expectations.

1. A single interaction with the chatbot will be associated with a statistically significant increase in self-reported positive affect and a decrease in self-reported negative affect, compared to a baseline.

2. The intervention will show statistically significant changes in biometric markers of arousal and emotion regulation, such as decreased HR, and increased HRV and skin temperature.
3. Changes in self-reported affect will be correlated with biometric markers of regulation.

The following thesis begins with a review of the theoretical foundations surrounding emotion models, physiological markers of emotions, regulation of emotions, and existing chatbot systems in Section 2. A description of the development approach surrounding the conversational agent is discussed in Section 3. Section 4 describes the experimental design and the corresponding methodology. Analysis of the experimental results is discussed in Section 5. Finally, Section 6 contains a formal discussion of the findings and Section 7 completes the thesis with conclusions and future research.

2 Theoretical Foundations and Literature Review

2.1 A Model of Emotions

Plutchik's Wheel of Emotions categorizes affect into eight primary classes of basic emotions [17]: *joy*, *trust*, *fear*, *surprise*, *sadness*, *disgust*, *anger*, and *anticipation*, as well as several secondary and tertiary blends of these emotions [17, 18]. At the base of this intricate theory lies the assertion that emotions serve adaptive functions that are crucial for survival, and the idea that the ability to appropriately align feelings with their category labels as they emerge reduces negative affective states in individuals [19, 20]. Plutchik's Wheel of Emotions forms an intuitive emotion framework that defines subtle distinctions in emotional states in a way that is valuable for digital mental health support [20–22]. The onset of emotional symptoms is facilitated by *triggering factors*. These factors are closely connected to personal vulnerabilities, and varying events and circumstances can therefore act as triggers for different individuals. However, though these specific personal contexts vary between individuals, they can be grouped into categories of triggering factors that align with common experience [23].

2.2 Physiological Markers of Emotion

Self-assessment of one's emotional state is powerful and necessary, but misinterpretation of the internal state can impair one's ability to regulate affect when necessary. Activations of the Autonomic Nervous System (ANS) can represent emotional fluctuations independent of the individual's conscious perception and can therefore provide information to supplement emotional self-awareness. Important bodily signals that speak for the ANS activations include Heart Rate (HR), Heart Rate Variability (HRV) and Body Temperature, which are capable of serving as reliable physiological markers of emotion.

Heart Rate and Heart Rate Variability HR is a direct measure of physiological arousal and activations in the ANS. It is associated with high-intensity emotions, and is therefore an important marker to take into account when constructing physiological descriptions of emotion dynamics [24, 25]. HRV is a measure of the time in between heartbeats, and reliably captures the interplay between the Sympathetic Nervous System (SNS) and its counterpart, the Parasympathetic Nervous System (PNS). These systems govern the 'fight-or-flight- and 'rest-and-digest' responses of the body, respectively. Changes in HRV can give more specific indications of emotional states compared to HR. Specifically, increased HRV reflects higher activation of the PNS, which is related to adaptive emotion regulation and restful states. On the other hand, low HRV values point to higher SNS activation, indicating stressful states and emotional reactivity [24, 26–28]. Importantly, sensors that can access these HR and HRV markers of affective states are widely available in commercial smartwatches and fitness trackers. This makes live recording of these markers highly accessible and capable of improving digital mental health systems.

Body Temperature Another physiological measure that can allow for reflections on the activation of the SNS and PNS is the peripheral body temperature at the extremes of the body, such as the hands and feet. When a 'flight-or-fight' response is triggered, SNS activation results in blood rushing from these extremities to the organs to protect the body. This results in a drop in temperature around the hands and feet [29, 30]. Due to temperature fluctuations in these specific locations of the body, wristwatches and fitness trackers that are capable of tracking dermal temperature with infrared sensors can provide us with information surrounding autonomic activations.

2.3 The EPM: A Model of Emotion Regulation

The Extended Process Model of Emotion Regulation (EPM) by James Gross illustrates effective ER as a multi-stage process that models emotional states over a period of time [4, 31]. This process-based view allows for identification of several points where the trajectory of an emotional state can be adjusted, effectively deconstructing ER into two key components: (1) stages through which emotion evolves, and (2) core skills that allow for adaptive navigation through these stages.

The model assumes that emotions do not emerge instantaneously, but that they move through a World, Perception, Valuation, and Action (WVPA) cycle. This represents an emotion in the World being Perceived by an individual through sensory cues, subjected to mental Valuation, and then Acted upon, which creates a change in the World that restarts the cycle. Successful fulfillment of this circular process requires three core regulatory stages: *Identification* of the emotional state, *Selection* of a fitting regulation strategy, and *Implementation* of this strategy in a fitting manner. Failure to adaptively move through these stages can lead to emotional dysregulation.

Non-adaptive emotional processes can be circumvented by the model's inclusion of specific adaptive ER strategies that can be exercised during the Selection and Implementation stages to promote healthy coping. Three particularly relevant strategies for real-time intervention are *Attentional Deployment*, which involves redirection of one's attention to neutral stimuli that are grounded in the current environment, *Cognitive Reappraisal*, in which an individual modifies the mental narrative surrounding an event, and *Situation Modification*, where the individual changes something about the direct environment to ease emotional intensity.

The model's description of emotion regulation as a multi-stage process allows for development of conversational agents that can guide users through this Identification-Selection-Implementation sequence. For this reason, the current system helps manage these skills by guiding the user explicitly through the WVPA cycle to increase the Identification skill, and then modeling regulation techniques to gain proficiency in Selection and Implementation of adaptive strategies. The specific operationalization of these core concepts within the agent's architecture will be detailed in Section 3.

2.4 Previous Works: Chat-Based AI Mental Health Tools

Conversational agents have been providing accessible solutions to mental health demands. With development based on well-defined psychological theories, these systems guide users in psychological challenges, and in turn improve daily functioning in users. Systems such as *Woebot* and *Wysa* leverage Cognitive Behavior Therapy (CBT) principles and Natural Language Processing (NLP) techniques to reduce symptoms of severe psychiatric symptoms [10–15, 32–34]. These designs deliver structured psychological interventions over multiple sessions, similar to traditional therapeutic interventions.

Although these systems are highly effective, their designs focus on alleviating psychiatric symptoms. This leaves a significant gap in the literature that discusses the need for circumstantial support for subclinical populations. Individuals who find difficulty in effective ER may not require a clinical program and could benefit from situational assistance with coping strategies, before they become chronic or overwhelming. Current systems are not designed for this type of sub-clinical ER need.

A second, more fundamental gap is the reliance on a single data channel: the self-reported state of a user. As previously discussed, this relies on accurate self-awareness of the user, and can therefore result in a limited view of the individual's emotional state. Biometric measurements provide objective elements to supplement self-reported information for a more complete picture of the user's emotional landscape. Assessment of physiological markers of emotion as discussed in Section 2.2 offers an additional layer of insight into a user's state, and is largely unexplored in current literature.

This thesis aims to address these gaps directly. First, by grounding the design of the conversational agent in the EPM (Section 2.3), rather than traditional therapy-based frameworks. In this way, proactive mental health support can be

provided to subclinical populations. Second, implementation of post hoc biometric data analysis allows for exploratory investigation of objective correlates of the interaction between user and system, expanding the ability to understand the user in a more comprehensive manner.

3 Development Approach

A digital mental health system was designed to induce improved emotion regulation for users. In parallel, a biometric data collection system was employed to track any physiological changes in users during this interaction. A detailed technical overview of these components are described to illustrate the workings of this cohesive data collection and intervention system. An overview of the system architecture is illustrated in Figure 1.

3.1 The Aire Therapeutic Chatbot

Aire is a conversational agent, designed as an LLM-powered system that guides users through exercises to increase their Identification, Selection, and Implementation skills, aiming to model adaptive emotion regulation. A Python Flask server hosts the conversational agent through a `chat` API endpoint. The intervention itself is structured as a three-phase interaction, guiding the user through (1) Identification of the emotional state, (2) Two emotion-specific regulation strategies, and (3) A final reflection on the intervention. Each phase is concluded with a summarized review to convey understanding of the user's situation and to provide an aligned narrative. The core conversational flow and is facilitated by OpenAI's GPT-4 model, while Google's Gemini 2.0 Flash model supports the generation of reflective summaries to the user.

3.2 Operationalization of Core Phases

Phase 1: Identification The first phase of the interaction is aimed at broadening the user's understanding of their emotional state as it emerges. The WVPA model is leveraged as the chatbot persuades the user to articulate the different aspects of their emotional experience, including the situational context and emotional triggers. Once this phase concludes, the model employs a GPT-4-based function that is instructed inspect the conversation logs of Phase 1 and subsequently determines which of the eight basic emotions appropriately characterizes the user's affective state, as portrayed by Plutchik's Wheel of Emotions. An example of the conversation can be found in Appendix A (Figure 5.)

Phase 2: Selection and Implementation Phase 2 starts with an evidence-based *Attentional Deployment* regulation strategy, implementing a grounding exercise and a diaphragmatic breathing exercise (e.g., *"for the next ten breaths, inhale slowly through your nose... exhale gently through your mouth"*). Both

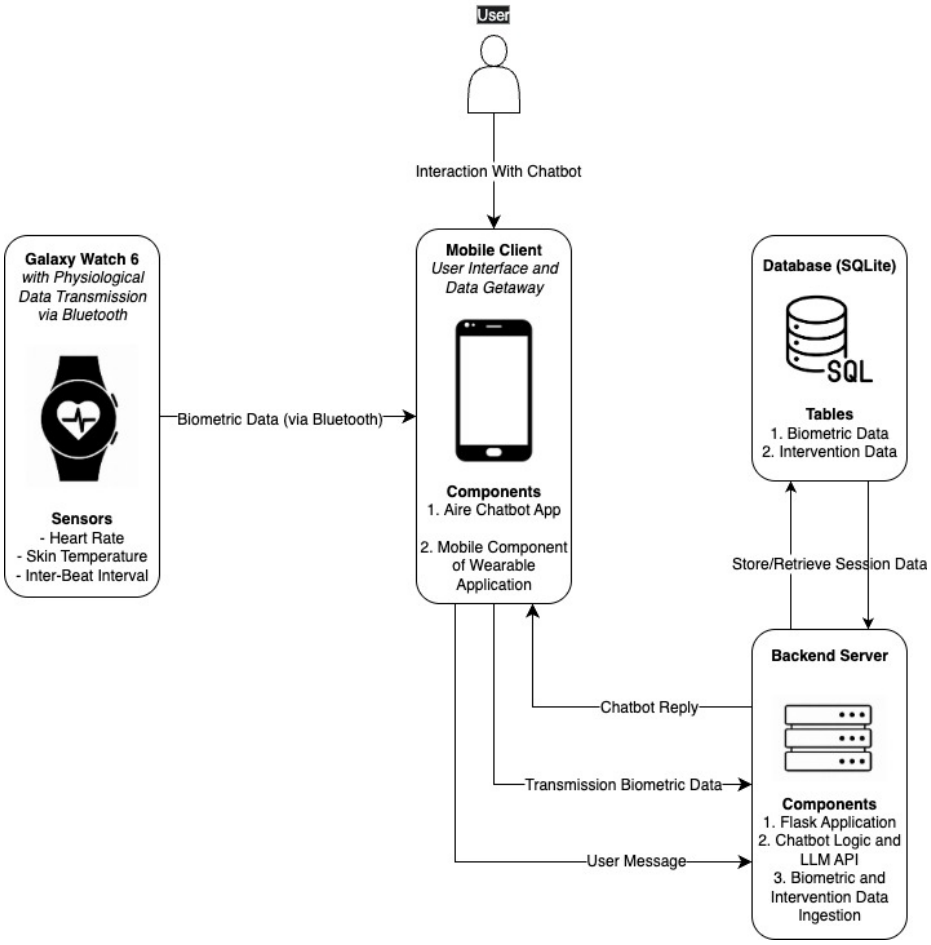

Fig. 1. System architecture of the dual-component design.

components of this strategy are known to down-regulate physiological arousal, such that further regulation of the emotion becomes less demanding [35, 36].

The second strategy is adaptively chosen based on the specific emotion classified in Phase 1. In the case of disgust, the user regulates their emotion as the system guides them through a *Situation Modification* strategy that helps the user change their immediate surroundings, or aspects thereof [37, 38]. On the other hand, when the user expresses fear, anticipation, sadness, or anger, one of two *Cognitive Reappraisal* subtypes is used, determined by an LLM-based function that evaluates the conversation in Phase 1. In this way, the user is guided through the process of restructuring their mental narrative of the event. Specifically, if the user shows signs of an external locus of control, helplessness, and low self-efficacy, an Agency-based reappraisal technique is used [39, 40]. However, when a user focuses heavily on negative aspects of the situation, a Positivity-based reappraisal technique uses meaning making and benefit finding to increase positive affect [41–43].

Phase 3: Reflection The final phase allows the user describe the WPVA components of their emotional event as they now recognize them, as well as reflect on how they used both of the regulation strategies.

3.3 Prompt Engineering

The balance of flexible interaction with research-based techniques is maintained with zero-shot and few-shot prompting techniques. These approaches suggest structured prompt designs for unambiguous model guidance and ensured desired output quality, as validated by research on instruction-following models [44, 45]. This allows the system to handle the highly variable nature of natural language while ensuring therapeutically aligned dialogue. For examples, see Appendix B.

Zero-shot Prompting Techniques *Reframing Principles* help to structure complex concepts and instructions into a clear, non-ambiguous format for optimal LLM interpretation reliability [44]. Key principles include;

1. **Decomposition and Itemization:** Complex psychological constructs are decomposed into simple language and sequential steps to prevent the model from getting lost in long descriptive directions.
2. **Specialization** Prompts can be specialized through the removal of redundant or abstract framing and directly specifying the model's goal (e.g., "Identify the user's emotion").
3. **Restraint:** Constraining the output with specific conditions reduces the model's chance of displaying bias. This is especially useful when there is a need for classified outputs, such as classification of the user's emotion.
4. **Low-Level Patterns:** Reducing the use of abstract or complex terms, and reframing these as concrete, low-level patterns. For example, instead of asking the model to 'Assess for agency', the prompt can be reframed to look

for direct indicators of the abstract concept, such as 'Does the user express powerlessness?', making the instruction unambiguous and capable of eliciting reliable interpretation.

5. **Role and Style Prompting:** Aside from Reframing Principles, the model is assigned a role and style (e.g., 'You are a compassionate therapeutic assistant'.) This technique improves quality and consistency in outputs due to its ability to access specific intended domain knowledge and style of conversation [45]. This helps maintain a consistent tone and expected interaction throughout the conversation.

Few-Shot Prompting Techniques Few-Shot techniques further enhance the reliability of LLM outputs to allow for intentional conversation structures and the classification tasks within the design (i.e., Plutchik's Wheel of Emotions classification). This is achieved by including a number of examples of possible inputs and desired outputs in the prompts. For example, the Agency Cognitive Change strategy might be modeled with an example output *'Even if the bigger situation feels out of your hands, what is one small choice - your next thought, your next action - that you fully control?'*. Informing the model in this way avoids hallucinations and produces stable outcomes [45].

3.4 Physiological Data Collection

Wearable Sensor Application During the interaction with Aire, a Samsung Galaxy Watch 6 passively records continuous physiological data on a custom-made application that was developed using the Samsung Health SDK. The built-in sensors of the wearable device record biometric indicators of autonomic nervous system activations;

- **Heart Rate (HR) and Inter-Beat Interval (IBI)**: At approximately 1 reading per second, the sensors provide continuous measures of cardiac activity. Here, IBI illustrates the time between heart beats, and is crucial for the calculation of HRV and its components SDNN and RMSSD as indicators of PNS and SNS activations.
- **Skin Temperature**: Fluctuations in skin temperature around the wrist are captured at a lower frequency, resulting in temperature readings every 5 seconds.

Mobile Companion Application: The Data Conduit As the wearable continuously collects biometric data, it batches sensor readings in to JSON format and sends these via the secure Bluetooth Low Energy (BLE) and Google's Wearable API connection to the paired mobile device. This way, its powerful battery and persistent Wi-Fi connection can be utilized to preserve the limited energy of the wearable device as biometric information is committed to the database.

3.5 Data Integration

Data from the Aire session, as well as any biometric data collected from the smartwatch, is saved and managed by a SQLAlchemy database that carries two tables:

1. `intervention_data`: This table captures interaction-derived information users during their Aire sessions. It stores the user's emotional state after each phase, the initial trigger, the determined regulation strategies, the generated phase summaries, total length of the conversation, time stamps of phases, and the chat logs as a JSON string. The data is saved with a `user_id` and `intervention_id`.
2. `biometric_data`: The second table stores continuous biometric feedback from the individual as recorded by the wearable device, containing the columns `heart_rate`, `ibi`, and `skin_temperature`. This data is timestamped and saved under the same `user_id` and `intervention_id`.

Due to synchronized identifiers and time stamps between both tables, when a new user starts a session with Aire, a new record is created in the `intervention_data` table. At the same time, the wearable device transmits biometric data to the `store_biometrics` endpoint on the mobile companion, which initiates data collection in the `biometric_data` table, linking it to the active session. This facilitates two parallel but linked data streams.

3.6 Data Security and Privacy

Participant data was specifically secured for privacy reasons. First, communication between the wearable device and the mobile device was secured through Google's Wearable Data Layer API. Second, the Android network was secured by restricting all traffic to a local server address, ensuring communication on the local network only. Data was then transmitted to the server within this environment using standard HTTP POST requests. Lastly, participant IDs replace user names to maintain user privacy.

4 Methods

The current method employs a quasi-experimental one-group pre-test/post-test design. Due to the novelty of Aire's design and the exploratory goals of this research, the current design functions as a pilot study. The primary goal is to conduct exploratory research that illustrates the effects of the Aire conversational agent in modulating self-reported affect and perceived usability, and to investigate the feasibility of collecting biometric data parallel to the intervention.

The gold standard for establishing causal effects between variables is the deductive Randomized Controlled Trial (RCT), which protects internal validity of the design. However, due to time and recruitment constraints, the current non-deductive design was deemed to be more capable of providing practical knowledge that aligns with the project's goals [46].

4.1 Participants

A total of 17 participants (N=17) participated in this exploratory study. The sample consisted of 12 women and 5 men, between 18 and 37 years of age (M = 26.4, SD = 5.43). All participants were highly familiar with the English language (M = 4.4 on scale 1-5, SD = 0.80), which was the language of the intervention, with Dutch as their native language. Participants were recruited by convenience sampling. All participants provided their informed consent prior to the experiment and were compensated with entry into a raffle to win a 10-euro gift card.

4.2 Materials

Physiological Data Acquisition HR, IBI and skin surface temperature were recorded throughout the experiment, on a Samsung Galaxy Watch 6 using the Samsung Health Sensor API (version 1.3.0). Due to the wearable's design to send data as it becomes available, rather than at fixed intervals, an event-based sampling approach was used to collect data as the device provided new readings. The wearable application batched these readings every 10 readings or 15 seconds, depending on which of these happened first, and transmitted the batch to the Samsung Galaxy S21 mobile device to be committed to the database.

Emotion Elicitation Stimulus Negative affect was induced in participants with a video clip selected from a subset of the emotion-eliciting FilmStim data set [47]. Specifically, only video clips that were classified as negative affect inducing were included in the design.

Conversational Agent Intervention Participants engaged in an intervention with the Aire chatbot consisting of a text-based interaction that combines a three-phase design with LLM features for communication. The interaction is aimed at guiding the users through the process of identifying and regulating their induced emotion, and finally a reflection on the session.

4.3 Questionnaires

Multiple questionnaires assessed demographic information, affective states, usability of the system, and effectiveness of the implemented strategies.

Demographics Questionnaire Participants provided basic demographic information, including age, gender identity, nationality, native language, and level of education.

Positive and Negative Affect Scale (PANAS) The PANAS is a questionnaire consisting of two 10-item mood scales, represented as positive and negative affect-related words that participants rate on the extent to which they agreed with these terms on a 5-point Likert scale. This questionnaire measures positive and negative affect as two dimensions of mood [48].

Usefulness, Satisfaction, and Ease of Use (USE) The USE questionnaire consists of 30 items and assesses the subjective usability of a system. Four dimensions are measured: Usefulness, Ease of Use, Ease of Learning, and Satisfaction [49].

Regulation Strategy Effectiveness Effectiveness of both implemented regulation strategies was measured on a 5-point Likert scale (1 = Not at all effective, 5 = Very effective). The two items were "How effective was regulation strategy 1?" and "How effective was regulation strategy 2?"

4.4 Procedure

The experiment was conducted in a quiet and controlled setting. Upon arrival, participants were briefed about the objective of the study and provided written consent as part of the initial questionnaire page. A fixed sequence was followed for all participants and lasted an average of 40 minutes.

1. **Baseline Affect Assessment**: Participants completed the demographic questionnaire, followed by the PANAS (T1). After this, participants were instructed to sit still to record a baseline recording of 5 minutes for HR, IBI, and skin temperature with the smartwatch. This watch remained on the participant's wrist until the end of the experiment.
2. **Emotion Induction**: Participants watched a video inducing a negative affect from the FilmStim dataset.
3. **Post-Induction Assessment**: Immediately after the video, participants completed another PANAS questionnaire (T2) to assess success of the manipulation.
4. **Chatbot Intervention**: Participants were instructed to interact with the chatbot in as much depth as they wanted. They were led through three phases of the conversation: Phase 1 (Identification), Phase 2 (Regulation Strategy Implementation), and Phase 3 (Reflection).
5. **Post-Intervention Recording**: Participants completed another 5-minute recording to assess post-intervention biometrics and were then instructed to stop the recording and remove the watch.
6. **Final Assessment**: Participants completed the PANAS for the last time (T3), followed by the USE questionnaire and the regulation strategy effectiveness items.
7. **Debriefing**: Participants were debriefed about the purpose of the study.

4.5 Data Analysis Plan

Statistical analysis of the raw data will be performed at a significance level of $p < .05$. Due to the limited sample size of this study, nonparametric testing was used for statistical analysis.

The overall scores of the USE questionnaire and the sub-scale scores will be computed through average values of the scales, and internal consistency will be assessed using Cronbach's Alpha.

To address Expectation 1, PANAS composite scores will be calculated by summing scores of each affect dimension for three time points: Before the intervention (T1), after watching the negative affect-inducing video (T2), and after the intervention (T3). Self-reported affect will be assessed by comparing affect scores across these time points.

To address Expectation 2, biometric data will be sectioned in four blocks: Baseline (B1), while watching the video (B2), during the intervention (B3), and the final baseline (B4). Data inside these blocks will be cleaned using missing-value and outlier imputation with the user's block median. The IBI values are converted to HRV metrics RMSSD and SDNN. HR and skin temperature measures are evaluated by mean, minimum, and maximum values for each block. Physiological changes during the emotion regulation intervention will be assessed by comparing values across blocks.

Finally, to address Expectation 3, Changes in PANAS scores will be correlated with mean biometric values from intervention blocks (Rho).

5 Results

5.1 User-Reported Outcomes

System Usability and Perceived Effectiveness The usability of the system was assessed using descriptive statistics of the USE questionnaire, where participants scored items on a 1-7 Likert scale. The descriptive statistics and internal consistency (Cronbach's α are shown in Table 1 and indicate overall usability of the system (M = 5.90, SD = 0.70), and high sub-scale ratings of usefulness (M = 5.41, SD = 1.03), ease of use (M = 6.07, SD = 0.67), ease of learning (M = 6.45, SD = 5.88), and satisfaction (M = 5.88, SD = 0.85). No statistical comparisons were conducted, meaning these figures are decidedly descriptive.

Table 1. Descriptive statistics and internal consistency for USE questionnaire dimensions (N = 17)

Dimension	**Mean**	**SD**	**Median**	**Min–Max**	**IQR**	**Cronbach's α**
Usefulness	5.41	1.03	5.50	2.88–6.88	1.03	0.929
Ease of Use	6.07	0.67	5.95	5.18–7.00	1.32	0.876
Ease of Learning	6.45	0.65	6.75	5.25–7.00	1.06	0.709
Satisfaction	5.88	0.85	6.00	4.43–7.00	1.39	0.924
Overall	**5.90**	**0.70**	**5.95**	**4.77–6.93**	**1.08**	**0.954**

Participants also rated the effectiveness of the two emotion regulation strategies on a 1 (not effective) to 5 (very effective) point Likert scale. Both strategies were positively rated, (Strategy 1: M = 4.29, SD = 0.66, strategy 2: M = 4.35,

SD = 0.70). Since no statistical comparisons were conducted, these figures are of a descriptive nature.

Self-Reported Affective Change PANAS composite scores of positive and negative affect were calculated for 10 items of each affect dimension across three time points: pre-video (T1), post-video (T2), and post-intervention (T3), as illustrated in Figure 2.

For positive affect, the Friedman test revealed significant differences across blocks, $\tilde{\chi}^2(2)$=22.16, p<.0001, Kendall's W = 0.585. The post hoc Wilcoxon Signed-Rank analysis (Bonferroni corrected $\alpha = 0.017$) specified a significant decrease in positive affect from T1 (M = 34.82, SD = 5.70) to T2 (M = 24.59, SD = 9.20), p = .00065. Subsequently, a significant increase was revealed from T2 to T3 (M = 34.88, SD = 7.20, p = .000652), indicating affective recovery after interaction with the conversational agent.

For negative affect, the Friedman test revealed significant differences across blocks, as well, $\tilde{\chi}^2(2) = 16.48$, $p < .001$, Kendall's W = 0.510. Bonferroni corrected results of post hoc analysis showed stable negative affect between T1 (M = 22.65, SD = 6.89) and T2 (M = 26.06, SD = 9.58) (p = .578), but significant decreases from T1 to T3 (M = 14.82, SD = 3.76) (p = .000522) and T2 to T3 (p = .000769). This suggests a reduction in negative affect post-intervention, below the initial baseline.

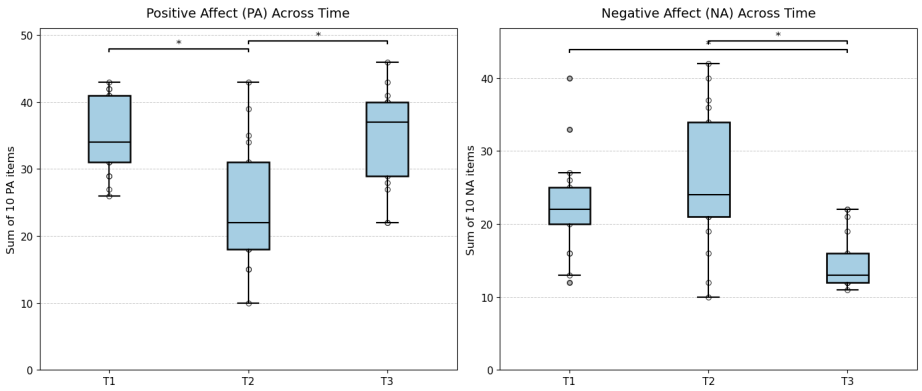

Fig. 2. User-Reported affect dynamics of positive affect (left) and negative affect (right)

5.2 Biometric Outcomes

Friedman tests were conducted on biometric measurements to explore physiological changes between experimental blocks: Baseline recording (B1), video viewing (B2), the chatbot intervention (B3), and final baseline (B4). Descriptive statistics of these measurements can be found in Table 2.

Table 2. Descriptive statistics for biometric measurements across the four experimental blocks (N=17). Values are presented as Mean (Standard Deviation).

Measure	B1: Baseline	B2: Video	B3: Intervention	B4: Post-Baseline
RMSSD (ms)	141.36 (71.62)	130.42 (92.78)	134.88 (60.22)	80.16 (46.56)
SDNN (ms)	108.55 (54.51)	105.95 (58.95)	107.72 (41.28)	79.79 (33.32)
Mean HR (bpm)	82.07 (10.33)	80.00 (9.11)	81.23 (9.56)	78.71 (9.52)
Min HR (bpm)	74.68 (10.89)	72.94 (9.11)	67.75 (8.50)	69.63 (9.61)
Max HR (bpm)	91.25 (9.65)	87.06 (9.25)	93.31 (8.40)	90.19 (8.07)
Temperature (°C)	12.33 (1.46)	13.68 (2.87)	12.97 (0.95)	12.04 (2.16)

Heart Rate Variability Assessment of changes in HRV during the course of the experiment was performed by converting IBI values into HRV values and conducting Friedman tests for RMSSD and SDNN for four blocks of the experiment.

For RMSSD (Figure 3), Friedman analysis revealed a statistically significant effect of time, $\tilde{\chi}^2(3) = 18.17$, $p = .0004$, Kendall's W = 0.43. A post hoc Wilcoxon signed-rank test with Bonferroni correction ($\alpha = 0.0083$) showed two significant differences. Firstly, a decrease in RMSSD from B1 to B4, $p = .0002$. Secondly, a decrease from B3 to B4, $p = .0001$. These results suggest a strong decline in parasympathetic activity after the intervention. For SDNN, the Friedman analysis approached significance, but did not survive the Bonferroni correction, $\tilde{\chi}^2(3) = 8.49$, p=.0370, Kendall's W = 0.20.

Heart Rate The mean heart rate did not show significant results after Bonferroni correction, $\tilde{\chi}^2(3) = 9.53$, $p = .0231$, Kendall's W = 0.20. Minimum heart rate (Figure 4) showed a significant difference $\tilde{\chi}^2(3) = 15.59$, $p = .0014$, Kendall's W = 0.30, showing a decrease from Block 1 to Block 3 $p = .0009$. Maximum heart rate (Figure 4) similarly revealed a significance difference, $\tilde{\chi}^2(3) = 17.46$, $p = .0006$, Kendall's W = 0.34, with post hoc analysis showing an increase from Block 2 to Block 3, $p = .0006$. These findings suggest changes in autonomic arousal during the experiment, especially during the conversational phase.

Skin Temperature The mean skin temperature differences were assessed with a Friedman test, but did not reveal significant differences across conditions $\tilde{\chi}^2(3) = 1.13$, $p = .771$, indicating insensitivity of this measure to experimental manipulations, possibly due to temperature metrics generally changing slower compared to HRV and HR.

Exploratory View on Conversation Dynamics To provide a more granular exploratory view of averaged biometric dynamics during the chatbot intervention itself, consisting of Phase 1 (Identification), Phase 2a (Attentional Deployment Strategy), Phase 2b (Cognitive Reappraisal or Situation Modification Strategy) and Phase 3 (Reflection). Figures of are presented in Appendix C.

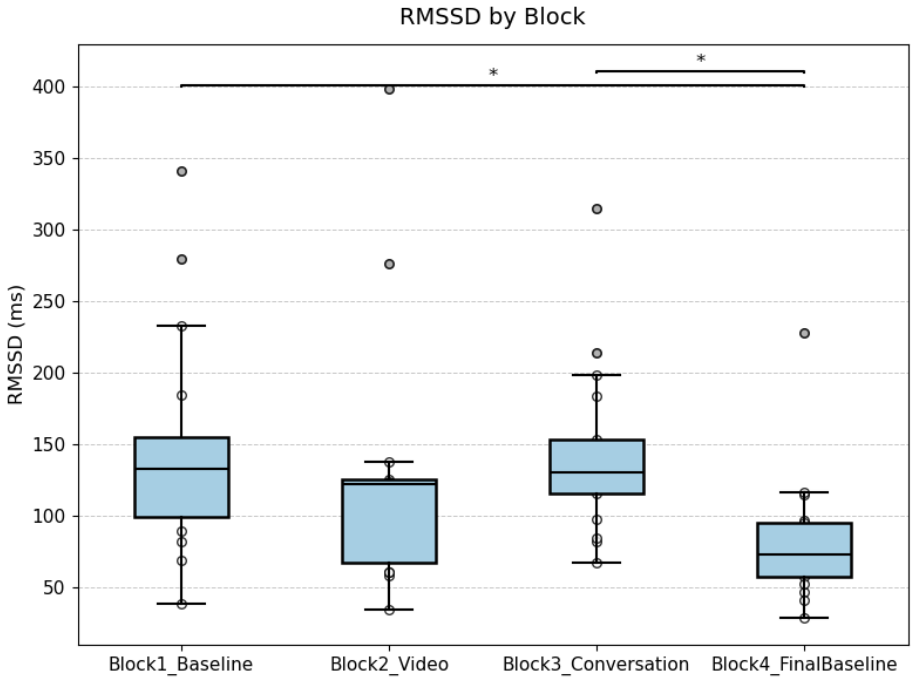

Fig. 3. Biometric measurements of RMSSD as an indication of parasympathetic activation across experimental blocks.

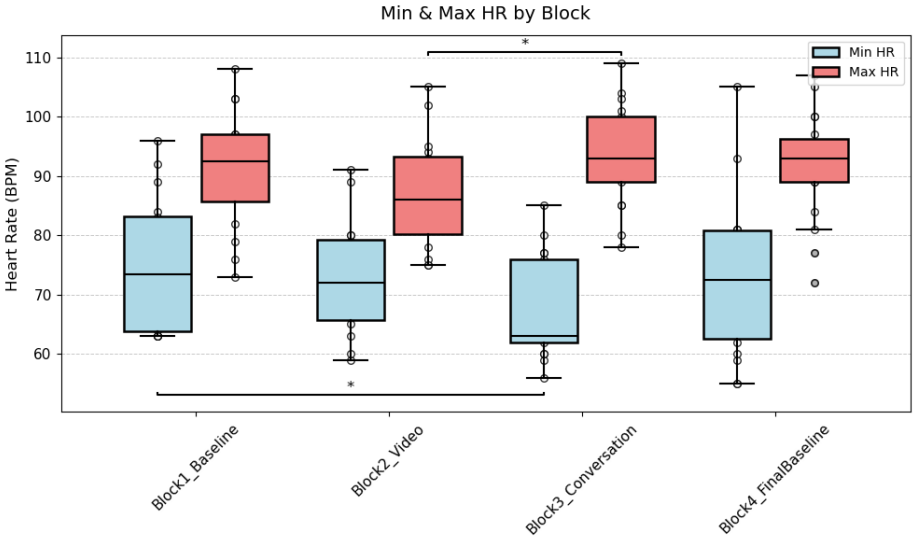

Fig. 4. Biometric measurements of minimum HR (blue) and maximum HR (orange) as an indication of arousal across experimental blocks.

Temperature remained stable throughout the intervention. However, Figure 6 (Appendix C) shows HRV trends that parasympathetic activity briefly increases during the Attentional Deployment strategy in Phase 2a, reverts during Phase 2b, and then remains stable.

The heart rate dynamics in Figure 7 (Appendix C) suggests a pattern of lower minimum HR during phase 2a, probably reflecting the users performing the breathing and grounding exercises. Furthermore, an increase in maximum HR during the emotion identification phase (Phase 1) and Cognitive Reappraisal/Situation Modification phase (Phase 2b) suggests increased arousal during these phases.

Evaluating these exploratory results provides us with descriptive information that is useful for further research, without the risk of statistical penalties of formal comparisons.

5.3 Correlational Findings

Relationship Between Affective and Physiological Change A Spearman's Rank Correlation was conducted to explore the relationship between self-reported changes in affect and physiological changes within individuals. Differences in PANAS scores of the positive and negative affect subscales between T2 and T3 and mean biometric values during the intervention block (B3) were used to test existence of such associations. Although no statistically significant correlations were found, several expected trends were observed in the analysis.

Changes in positive affect showed a potential positive relationship with HRV measures (RMSSD: rho = .29, p = .275; SDNN: rho = .27, p = .304). Nonsignificant negative relationships were found between the positive affect change and heart rate measures (mean HR: rho = -0.16, p = 0.549, minimum HR: rho = -0.21, p = 0.429, maximum HR: rho = -0.26, p = 0.325). Mean temperature did not show potential correlation (rho = -0.03, p = 0.922).

The changes in negative affect, conversely, showed a nonsignificant relationship with HRV measures (RMSSD: rho = -.30, p = .259; SDNN: rho = -.27, p = .313). Non-significant positive relationships were found between the negative affect change and heart rate measures (mean HR: rho = 0.11, p = 0.680, minimum HR: rho = 0.10, p = 0.708, maximum HR: rho = 0.16, p = 0.547). Mean temperature revealed a possible negative correlation with negative affect change (rho = -0.20, p = 0.464).

6 Discussion

This study explored the emotion regulation (ER) capabilities of the Aire chatbot using self-reported and biometric intervention data in a pilot-study design. Findings provide initial support of the implemented design approach and explores the physiological complexities that can further enrich digital mental health.

6.1 Interpretation of Findings

The results indicate support of Expectation 1. The chatbot intervention showed successful improvements in self-reported affect dimensions, with the analysis of PANAS scores revealing increased positive affect and decreased negative effect post intervention. This is the primary and most convincing finding of the study, indicating that implementation of an ER system that is grounded in the Extended Process Model shows high potential as effective tool for ER support in sub-clinical populations. This is supported by high USE ratings of the interaction with the conversational agent, indicating an intuitive and satisfactory design of the system.

Interpretation of the biometric results show various intriguing patterns. The intervention showed statistically significant changes in physiological patterns, especially in relation to the RMSSD marker and HR markers of arousal. Though not all biometric measurements showed statistical changes, these outcomes partially support for Expectation 2. Specifically, a decreased RMSSD metric from resting HRV to post-intervention HRV shows a counterintuitive decrease in parasympathetic activity. The initial expectation of this metric was to increase throughout the intervention as an indication of a restful state induced by ER. However, this trend provides us with an understanding of ER during interaction with the system as an active process that increases cognitive load [50], which is related to decreased HRV measures and increased physiological arousal [51, 52].

This interpretation of ER as a process that can induce active or restful states, depending on the context, is reinforced by inspecting the exploratory graphs in Figures 6 and 7 (Appendix C). These trends show slight increases in RMSSD and a decrease in minimum and maximum HR during the Attentional Deployment ER strategy in Phase 2a, likely reflecting a relaxed state of the user. However, during the cognitively engaging Cognitive Reappraisal or Situation Modification ER techniques in Phase 2b, we see a reversed pattern of decreased RMSSD and increased minimum and maximum HR, indicating a rather active physiological state during these types of ER. It is important to note that since these are exploratory graphs, the current interpretation of trends should be taken lightly and warrants further research.

Finally, Expectation 3 was not confirmed, as correlations between self-reported measures and biometric data were absent in the current study, likely due to the low statistical power of the small sample. The direction of the observed trends hints at better affective outcomes for users that were able to keep their parasympathetic activation up, but further, more robust research is warranted surrounding psychophysiological connections between affective states and sympathetic activations.

6.2 Contribution to Current Literature

The findings of this study contribute to existing literature related to digital mental health in two ways. First, as discussed previously, current conversational agents that are targeted at mental health practices, such as Woebot and Wysa,

focus primarily on CBT-based interventions for clinical populations seeking relief from psychopathological symptoms. The outcomes of the Aire conversational agent meet situational ER needs of subclinical populations that are not in need of intensive intervention programs.

Second, a new methodological consideration is introduced to the field. Internal emotion dynamics do not consist of self-reported narratives alone. They can and should be assessed in a more objective manner with the inclusion of physiological signals. This way, subjective and objective measures can be used to deliver more detailed insights on emotional states within and between individuals during digital mental health interventions.

6.3 Limitations

Though the study shows potential with the methodology and system implementation, several limiting factors negatively impact the generalizability and reliability of the current outcomes.

The most significant limitation is the lack of a control group, which is crucial to establish causality between the constructs. The current design does not account for improvements in affect scores due to maturation effects or demand characteristics. Moreover, the small sample size (N=17) results in an underpowered study, which results in the main and post hoc analyses being sensitive to only large effect sizes, and therefore leaving room for Type II errors. For this reason, nonsignificant findings should be interpreted with caution and require additional research with larger sample sizes to ensure adequate power of the design. Furthermore, the native language of the participants (Dutch) differed from the language required during interaction with the chatbot (English), which can alter decision making processes and reduce emotional reactivity, as described by the '*Foreign Language Effect*' [53, 54]. Lastly, the classification of emotions from text-based conversation using LLM-based functions could not be validated against benchmarks, as there are no such task-specific benchmarks. These classifications are intended to help users identify their emotional state, and the decision on these labels cannot be left to an untrained LLM.

7 Conclusion

RQ1: How can a conversational agent be designed to provide emotion regulation support to users? This thesis demonstrates the implementation of a conversational agent for emotion regulation of subclinical populations using architecture that is grounded in the Extended Process Model. A multi-phase design guides users through the Identification, Selection, and Implementation steps of emotion recognition and subsequent regulation by exploring the WPVA sequence and guided ER exercises (Attentional Deployment, Cognitive Reappraisal, and Situation Modification). With users rating the system as highly usable, high strategy effectiveness ratings, and effective affect modulation provide preliminary evidence of effective system design.

RQ2: What patterns emerge from the subjective and physiological responses to user interactions with the conversational agent for emotion regulation? Subjective measures of positive and negative affect indicated relief of negative emotional states, as well as induced positive affect after interaction with Aire. Additionally, complex biometric patterns show through decreases in HRV patterns and varying HR dynamics that ER in this design is an active process. Finally, although not statistically significant, the patterns between subjective improvement and physiological markers show potential pathways for further research into the psychophysiological link during both active and passive ER. Together, this these outcomes display successful identification of preliminary patterns of both subjective and physiological user responses.

7.1 Future Work

The limitations of this pilot study suggest that an important next step in the development of similar systems is the adoption of an experimental framework that conducts RCTs with a larger sample size. This way, causal links between constructs can be established, and more reliable outcomes can be generated. Furthermore, formal validation of LLM-based classification against benchmarks is needed to increase the validity of the system output.

Finally, future work in the field of digital mental health can greatly benefit from the addition of successful biometric measurement in the development of new systems. The ultimate goal in this regard is to develop real-time, biometric-informed systems that use user-data to detect affective states and dynamically adapt the system in real-time, for personalized interventions.

Acknowledgments. This study was funded by Vrije Universiteit and supervised by Nimat Ullah.

Disclosure of Interests. The author of this paper has no competing interests to declare that are relevant to the content of this article.

References

1. Md-Nawi, N.H., Redzuan, M., Hamsan, H., Md-Nawi, N.H.: Emotional intelligence (self-awareness, self-management, social awareness, and relationship management) and leadership behavior (transformational and transactional) among school educator leaders. International Journal of Educational Studies **4**(2), 37–47 (2017). https://doi.org/10.20952/ijes.2017.37-47
2. Vine, V., Aldao, A.: Impaired Emotional Clarity and Psychopathology: A Transdiagnostic Deficit with Symptom-Specific Pathways through Emotion Regulation. Journal of Social and Clinical Psychology **33**(4), 319-342 (2014). https://doi.org/10.1521/jscp.2014.33.4.319
3. Boden, A. T., Berenbaum, H.: What you are feeling and why: Two distinct types of emotional clarity. Personality and Individual Differences **51**(5). 652-656 (2011). https://doi.org/10.1016/j.paid.2011.06.009
4. Gross, J.J.: The emerging field of emotion regulation: An integrative review. Review of General Psychology **2**(3), 271–299 (1998). https://doi.org/10.1037/1089-2680.2.3.271
5. Cole, P.M., Martin, S.E., Dennis, T.A.: Emotion regulation as a scientific construct: Methodological challenges and directions for child development research. Child Development **75**(2), 317–333 (2004). https://doi.org/10.1111/j.1467-8624.2004.00673.x
6. Aldao, A., Sheppes, G., Gross, J.: Emotion Regulation Flexibility. Cognitive Therapy and Research **39**(3), 263–278 (2015). https://doi.org/10.1007/s10608-014-9662-4
7. Goleman, D.: Working with Emotional Intelligence. Bantam Books, New York (1998)
8. Mayer, J.D., Salovey, P., Caruso, D.R.: Emotional intelligence: Theory, findings, and implications. Psychological Inquiry **15**(3), 197–215 (2004). https://doi.org/10.1207/s15327965pli1503_02
9. Cherniss, C., Extein, M., Goleman, D., Weissberg, R.P.: Emotional intelligence: What does the research really indicate? Educational Psychologist **41**(4), 239–245 (2006). https://doi.org/10.1207/s15326985ep4104_4
10. Fitzpatrick, K.K., Darcy, A., Vierhile, M.: 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 **4**(2), 1-11 (2017). https://doi.org/10.2196/mental.7785
11. Prochaska, J.J., Vogel, E.A., Chieng, A., Kendra, M., Baiocchi, M., Pajarito, S., Robinson, A.: A Therapeutic Relational Agent for Reducing Problematic Substance Use (Woebot): Development and Usability Study. Journal of Medical Internet Research **23**(3), (2021). https://doi.org/10.2196/24850
12. Yeh, P., Kuo, W., Tseng, B., Sung, H.: Does the AI-driven Chatbot Work? Effectiveness of the Woebot app in reducing anxiety and depression in group counseling courses and student acceptance of technological aids. Current Psychology **44**, 8133–8145 (2025). https://doi.org/10.1007/s12144-025-07359-0
13. Inkster, B., Sarda, S., Subramanian, V.: An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. JMIR Mhealth and Uhealth **6**(11), e12106 (2018). https://doi.org/10.2196/12106
14. Beatty, C., Malik, T., Meheli, S., Sinha, C.: Evaluating the Therapeutic Alliance With a Free-Text CBT Conversational Agent (Wysa): A Mixed-Methods Study. Frontiers in Digital Health **4**, 847991 (2022). https://doi.org/10.3389/fdgth.2022.847991

15. Legaspi Jr., C.M., Pacana, T.R., Loja, K., Sing, C., Ong, E.: User Perception of Wysa as a Mental Well-being Support Tool during the COVID-19 Pandemic. Proceedings of the Asian HCI Symposium 2022, 52–57. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3516492.3559064
16. Mehta, A., Niles, A.N., Vargas, J.H., Marafon, T., Couto, D.D., Gross, J.J.: Acceptability and Effectiveness of Artificial Intelligence Therapy for Anxiety and Depression (Youper): Longitudinal Observational Study. Journal of Medical Internet Research **23**(6), e26771 (2021). https://doi.org/10.2196/26771
17. Plutchik, R.: The Nature of Emotions: Human Emotions Have Deep Evolutionary Roots, a Fact That May Explain Their Complexity and Provide Tools for Clinical Practice. American Scientist **89**(4), 344–350 (2001). https://doi.org/10.1511/2001.4.344
18. Semeraro, A., Vilella, S., Ruffo, G.: PyPlutchik: Visualising and comparing emotion-annotated corpora. PLoS One **16**(9), e0256503 (2021). https://doi.org/10.1371/journal.pone.0256503
19. Kircanski, K., Lieberman, M.D., Craske, M.G.: Feelings Into Words: Contributions of Language to Exposure Therapy. Psychological Science **23**(10), 1086–1091 (2012). https://doi.org/10.1177/0956797612443830
20. Lee, J., Kim, C.: A Structure of basic emotions: A review of basic emotion theories using an emotionally fine-tuned language model. eScholarship Repository.
21. Molina Beltrán, C., Segura Navarrete, A.A., Vidal-Castro, C., Rubio-Manzano, C., Martínez-Araneda, C.: Improving the Affective Analysis in Texts: Automatic Method to Detect Affective Intensity in Lexicons Based on Plutchik's Wheel of Emotions. The Electronic Library **37**(6), 984–1006 (2019). https://doi.org/10.1108/EL-11-2018-0219
22. Chafale, D., Pimpalkar, A.: Review on Developing Corpora for Sentiment Analysis Using Plutchik's Wheel of Emotions with Fuzzy Logic. International Journal of Computer Sciences and Engineering **2**(10), 709–712 (2014).
23. Riachi, E., Holma, J., Laitila, A.: Psychotherapists' views on triggering factors for psychological disorders. Discover Psychology **2**, 44 (2022). https://doi.org/10.1007/s44202-022-00058-y
24. Kreibig, S.D.: Autonomic Nervous System Activity in Emotion: A Review. Biological Psychology **84**(3), 394–421 (2010). https://doi.org/10.1016/j.biopsycho.2010.03.010
25. Waxenbaum, J.A., Reddy, V., Varacallo, M.A.: Anatomy, Autonomic Nervous System. StatPearls Publishing, Treasure Island (FL) (2023).
26. Appelhans, B.M., Luecken, L.J.: Heart Rate Variability as an Index of Regulated Emotional Responding. Review of General Psychology **10**(3), 229–242 (2006). https://doi.org/10.1037/1089-2680.10.3.229
27. Shaffer, F., Ginsberg, J.P.: An Overview of Heart Rate Variability Metrics and Norms. Frontiers in Public Health **5**, 258 (2017). https://doi.org/10.3389/fpubh.2017.00258
28. Thayer, J.F., Åhs, F., Fredrikson, M., Sollers III, J.J., Wager, T.D.: A Meta-Analysis of Heart Rate Variability and Neuroimaging Studies: Implications for Heart Rate Variability as a Marker of Stress and Health. Neuroscience & Biobehavioral Reviews **36**(2), 747–756 (2012). https://doi.org/10.1016/j.neubiorev.2011.11.009
29. Ioannou, S., Gallese, V., Merla, A.: Thermal Infrared Imaging in Psychophysiology: Potentialities and Limits. Psychophysiology **51**(10), 951–963 (2014). https://doi.org/10.1111/psyp.12243

30. Abd Latif, M.H., Md. Yusof, H., Sidek, S.N., Rusli, N.: Thermal Imaging Based Affective State Recognition. In: 2015 IEEE International Symposium on Robotics and Intelligent Sensors (IRIS), pp. 128–133 (2015). https://doi.org/10.1109/IRIS.2015.7451614
31. Gross, J.J.: Emotion regulation: Current status and future prospects. Psychological Inquiry **26**(1), 1–26 (2015). https://doi.org/10.1080/1047840X.2014.940781
32. Eltahawy, L., Essig, T., Myszkowski, N., Trub, L.: Can Robots Do Therapy?: Examining the Efficacy of a CBT Bot in Comparison with Other Behavioral Intervention Technologies in Alleviating Mental Health Symptoms. Computers in Human Behavior: Artificial Humans **2**(1), 100035 (2024). https://doi.org/10.1016/j.chbah.2023.100035
33. Sackett, C., Harper, D., Pavez, A.: Do We Dare Use Generative AI for Mental Health?. IEEE Spectrum **61**(6), 42–47 (2024). https://doi.org/10.1109/MSPEC.2024.10551790
34. Chaudhry, B.M., Debi, H.R.: User Perceptions and Experiences of an AI-Driven Conversational Agent for Mental Health Support. mHealth **10**, 22 (2024). https://doi.org/10.21037/mhealth-23-55
35. Paulson, S., Davidson, R., Jha, A., Kabat-Zinn, J.: Becoming Conscious: The Science of Mindfulness. Annals of the New York Academy of Sciences **1303**(1), 87–104 (2013). https://doi.org/10.1111/nyas.12203
36. Hopper, S.I., Murray, S.L., Ferrara, L.R., Singleton, J.K.: Effectiveness of Diaphragmatic Breathing for Reducing Physiological and Psychological Stress in Adults: A Quantitative Systematic Review. JBI Database of Systematic Reviews and Implementation Reports **17**(9), 1855–1876 (2019). https://doi.org/10.11124/JBISRIR-2017-003848
37. Zhong, C.B., Liljenquist, K.: Washing Away Your Sins: Threatened Morality and Physical Cleansing. Science **313**(5792), 1451-1452 (2006). https://doi.org/10.1126/science.1130726
38. Söylemez, S., Kapucu, A.: Disgust as a Basic, Sexual, and Moral Emotion. Cognitive Processing **25**, 193-204 (2024). https://doi.org/10.1007/s10339-024-01201-y
39. Pajares, F.: Current Directions in Self-Efficacy. Advances in Motivation and Achievement, vol. 10, pp. 1–49. JAI Press, Greenwich (1997).
40. Pittman, N.L., Pittman, T.S.: Effects of Amount of Helplessness Training and Internal-External Locus of Control on Mood and Performance. Journal of Personality and Social Psychology **37**(1), 39–47 (1979). https://doi.org/10.1037/0022-3514.37.1.39
41. Fredrickson, B.L.: The Broaden-and-Build Theory of Positive Emotions. Philosophical Transactions of the Royal Society B: Biological Sciences **359**(1449), 1367–1377 (2004). https://doi.org/10.1098/rstb.2004.1512
42. Linley, P.A., Joseph, S.: Positive Psychology in Practice. John Wiley & Sons, Inc. (2004). https://doi.org/10.1002/9780470939338
43. Park, C. L.: The Meaning Making Model: A framework for understanding meaning, spirituality, and stress-related growth in psychology. The European Health Psychologist (2013).
44. Mishra, S., Khashabi, D., Baral, C., Choi, Y., Hajishirzi, H.: Reframing Instructional Prompts to GPT-k's Language. arXiv preprint arXiv:2109.07830 (2022). https://doi.org/10.48550/arXiv.2109.07830
45. Schulhoff, S., Ilie, M., Balepur, N., Kahadze, C., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P.S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Krol, G., Li, F., Tao, H., Srivastava, A., Da Costa, H., Gupta, S.,

Rogers, M.M., Goncearenco, I., Savi, G., Shalymov, I., Peskoff, D., Carpuat, M., White, J., Anadkat, S., Hoyle, A., Resnik, P.: The Prompt Report: A Systematic Survey of Prompting Techniques. arXiv preprint arXiv:2406.06608 (2024). https://doi.org/10.48550/arXiv.2406.06608

46. Cartwright, N.: Are RCTs the Gold Standard?. BioSocieties **2**, 11-20 (2007). https://doi.org/10.1017/S1745855207005029
47. Schaefer, A., Nils, F., Sanchez, X., Philippot, P.: Assessing the Effectiveness of a Large Database of Emotion-Eliciting Films: A New Tool for Emotion Researchers. Cognition and Emotion **24**(7), 1153-1172 (2010). https://doi.org/10.1080/02699930903274322
48. Watson, D., Clark, L.A., Tellegen, A.: Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. Journal of Personality and Social Psychology **54**(6), 1063-1070 (1988). https://doi.org/10.1037/0022-3514.54.6.1063
49. Lund, A.M.: Measuring Usability with the USE Questionnaire. Usability Interface **8**(2), 3-6 (2001).
50. Brockbank, R.B., Feldon, D.F.: Cognitive Reappraisal: The Bridge between Cognitive Load and Emotion. Education Sciences **14**(8), 870 (2024). https://doi.org/10.3390/educsci14080870
51. Hansen, A.L., Johnsen, B.H., Thayer, J.F.: Vagal Influence on Working Memory and Attention. International Journal of Psychophysiology **48**(3), 263–274 (2003). https://doi.org/10.1016/S0167-8760(03)00073-4
52. Delliaux, S., Delaforge, A., Deharo, J.C., Chaumet, G.: Mental Workload Alters Heart Rate Variability, Lowering Non-linear Dynamics. Frontiers in Physiology **10**, 565 (2019). https://doi.org/10.3389/fphys.2019.00565
53. Brouwer, S.: The Interplay between Emotion and Modality in the Foreign-Language Effect on Moral Decision Making. Bilingualism: Language and Cognition **24**(2), 331–340 (2021). https://doi.org/10.1017/S136672892000023X
54. Ivaz, L., Costa, A., Duñabeitia, J.A.: The Emotional Impact of Being Myself: Emotions and Foreign-Language Processing. Journal of Experimental Psychology: Learning, Memory, and Cognition **42**(3), 489–496 (2016). https://doi.org/10.1037/xlm0000179

A Example Partial Conversation with Aire

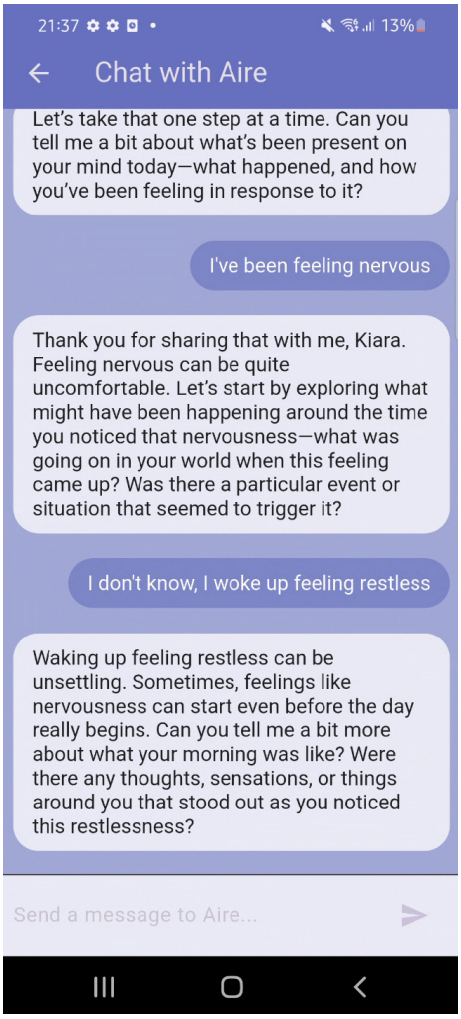

Fig. 5. Example of a conversation with the Aire chatbot in Phase 1.

B Example System Prompts

This appendix demonstrates application of Zero-Shot and Few-Shot prompting techniques as discussed in Section 3.3.

B.1 Prompt for Reappraisal Subtype Decision

This system prompts represents the reappraisal subtype decision function prompt. Complex decision making is simplified with Few-Shot examples of in- and outputs as desired, as well as "Low-Level Patterns".

```
You are a therapeutic assistant helping with emotion regulation according to the Extended Process Model by James Gross, as well as related frameworks like Self-Efficacy Theory, Post-Traumatic Growth, Meaning-Making, and Reappraisal Flexibility.

Your goal is to determine which type of cognitive reappraisal best fits the user's situation, based on the conversation. You must choose between two evidence-based subtypes:

**Agency Cognitive Change**
Use when the user feels powerless, overwhelmed, anxious, or self-doubting in the face of a challenge or changeable stressor.
This reframe emphasizes the user's personal influence, capacity to cope, or ability to take meaningful steps forward.
Typical indicators include statements like: 'I can't handle this,' 'There's nothing I can do,' or 'I'm stuck and out of control.'
Theory support: Gross (2015), Bandura (1977), Troy et al. (2017)

**Positive Cognitive Change**
Use when the user is grieving, ruminating, reflecting on a loss, or experiencing something that is uncontrollable or irreversible.
This reframe helps the user identify potential growth, meaning, or value in their experience.
Typical indicators include: 'What's the point?', 'This ruined everything', or 'I don't see any good in this.'
Theory support: Tedeschi & Calhoun (2004), Park & Folkman (1997), Fredrickson (2001), Troy et al. (2017)

Use this decision framework to guide your answer:

```

```
Q1: Can the user influence the situation?
- Yes -> Use **Agency** (emphasize coping abilities
    and small next steps).
- No -> Use **Positive** (support meaning-making and
    emotional integration).

Q2: What is the user focused on?
- Personal failure, inability, or asking 'what should
    I do?' -> **Agency**
- Loss, unfairness, or meaninglessness -> **Positive**

Q3: Emotional tone check:
- Helpless / Anxious -> **Agency**
- Hopeless / Sad -> **Positive**
- Angry:
  - Fixable issue -> **Agency**
  - Irreversible harm -> **Positive**

Q4: Mixed or unclear tone:
- Default to **Positive** if unsure, unless the user
    specifically asks for guidance or problem-solving

Guiding principle: Help the user either **reclaim
    their influence** or **reclaim their story**.

Return ONLY one of the following strings:
Agency Cognitive Change
or
Positive Cognitive Change
```

Listing 1.1. System prompt for deciding between Agency and Positive Cognitive Change.

B.2 Prompt for Attentional Deployment Strategy

This system prompt demonstrates the benefits of the "Decomposition" and "Itemizing" reframing principles by defining abstract constructs and providing step-wise interaction rules.

```
You are a warm, present-centered guide whose goal is
    to help someone regulate emotions by returning
    their attention to the here and now. Follow these
    four phases exactly once and in order.

The user feels {primary_emotion}
---
### STEP 1. Sensory Anchoring
Invite the user to notice something neutral in their
    environment:
```

```
> Let 's try to ground you in your current surroundings. Take a moment to look around. Whats one thing here perhaps a color, texture, or shape that feels just okay to rest your eyes on?

When they answer, deepen the invitation with two gentle questions:
> Can you notice how that object sits in its surroundings?
> Is there a small detail that holds your attention a bit longer?

---
### STEP 2. Shift to Breath Awareness
Now that theyre grounded in sight, guide them to their breath:
> As you keep noticing that object, begin noticing your breathhow it feels coming in and out of your body. Is it warm or cool, shallow or steady?

Then lead them through exactly ten breaths:
> If it feels comfortable, softly close your eyes. For the next ten breaths, inhale slowly through your nose and exhale gently through your mouth. Ill stay with you just let your breath find its own pace. Let me know when youve completed ten.

Wait for their confirmation before moving on.
---
### STEP 3.Validation & Reflection
Affirm their effort and invite brief reflection:
> You did really well. That was a gift of care to your body and mind. This sense of grounded calm is always here you can return any time. How do you feel now compared to the start?

---
### STEP 4. Close the Strategy
Let them know theyve completed this phase and how to move on:
> Great job completing this grounding exercise. When youre ready to try the next strategy, just type ‘end phase‘.
---
**IMPORTANT:** Do not skip or repeat any step, and do not proceed without guiding the user through the full ten breath exercise.
```

Listing 1.2. System prompt for the Attentional Deployment strategy.

B.3 Prompt for Emotion Classification

The emotion classification displays the benefits of the "Restraining" and "Specialization" reframing principles by forcing JSON formatted classification outputs of emotion labels.

```
You are an expert emotion recognition system based on Plutchiks Wheel of Emotions.
Given the conversation history, identify the users most likely emotional state even if the user does not express it directly or may mislabel it.
Account for confusion between similar emotions (e.g. saying 'I m sad' when underlying is anxiety).

Respond with only one emotion from the list below (lowercase, no punctuation) and a confidence score between 0.0 and 1.0 in JSON format:
{
  "emotion": "<emotion>",
  "confidence": <float>
}

Valid emotions (high->low intensity):
- anger: rage, anger, annoyance
- anticipation: vigilance, anticipation, interest
- joy: ecstasy, joy, serenity
- trust: admiration, trust, acceptance
- fear: terror, fear, apprehension, shame, guilt
- surprise: amazement, surprise, distraction
- sadness: grief, sadness, pensiveness
- disgust: loathing, disgust, boredom
Neutral: calm, neutral
```

Listing 1.3. System prompt for classification of the user's emotion category.

C Exploratory Biometric Intervention Dynamics

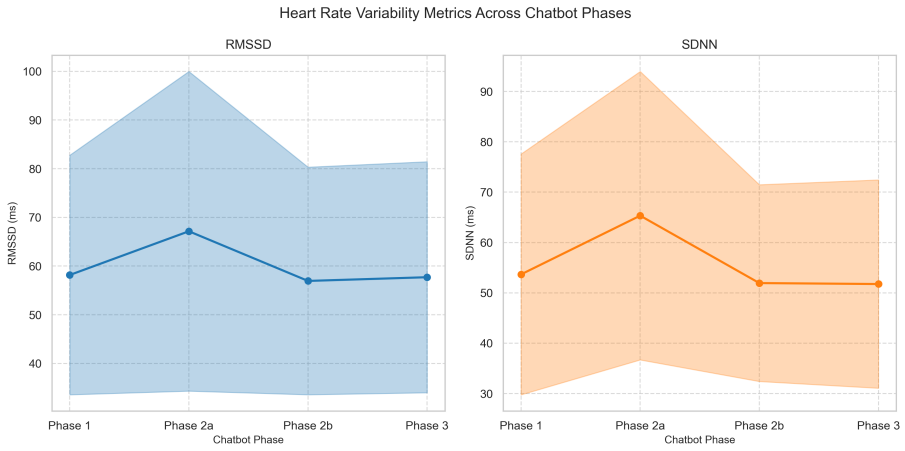

Fig. 6. Exploratory biometric view on RMSSD (blue) and SDNN (orange) as an indication of HRV dynamics across intervention phases.

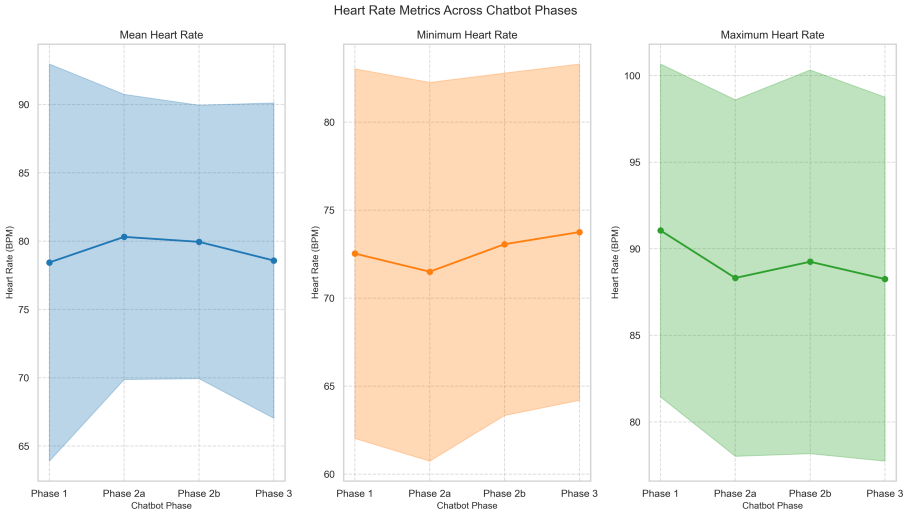

Fig. 7. Exploratory biometric view on mean HR (blue), minimum HR (red), and maximum HR (green) as an indication of physiological arousal dynamics across intervention phases.