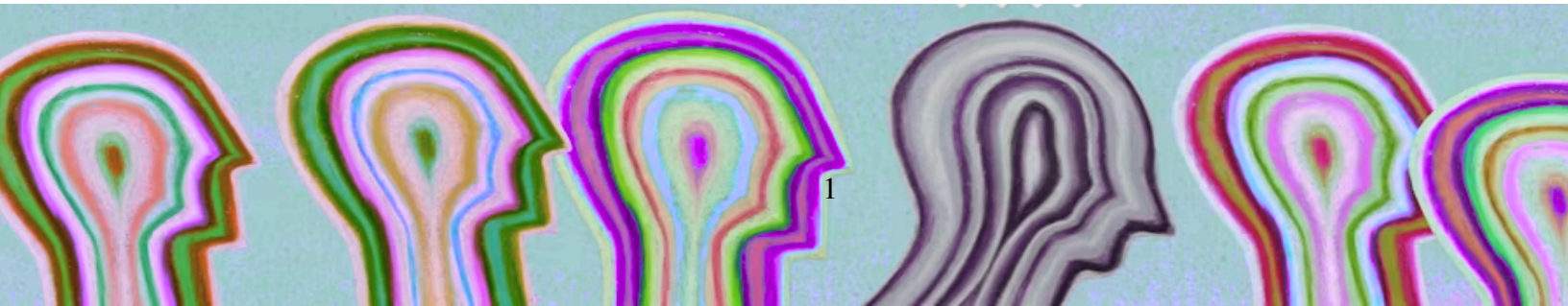
Towards Human-Centered Data-Driven Emotion Recognition

Pragya Singh



Towards Human-Centered Data-Driven Emotion Recognition

by
Pragya Singh (PhD21004)

Jointly advised by
Prof. Pushpendra Singh (IIIT-Delhi, India)
Prof. Mohan Kumar (RIT, NY, USA)

A dissertation
submitted in partial fulfillment
of the requirements for the degree

Doctor of Philosophy

Department of Computer Science and Engineering
Indraprastha Institute of Information Technology Delhi

January 2026

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Copyright © Pragya Singh 2026

Approved By (External Examiners):
Anind K. Dey (University of Washington)
Kristof Van Laerhoven (University of Siegen)
Vassilis Kostakos (University of Melbourne)

Certificate

This is to certify that the thesis titled as "**Towards Human-Centered Data-Driven Emotion Recognition**" submitted by Pragya Singh to the *Indraprastha Institute of Information Technology Delhi* (IIIT-Delhi), for the award of the degree of Doctor of Philosophy, is an original research work carried out by her under the joint supervision of Prof. Pushpendra Singh (IIIT-Delhi, India) and Prof. Mohan Kumar (Professor, RIT, New York, USA). In our opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree. The results presented in this thesis have not been submitted in part or whole to any other university or institute for the award of any degree/diploma.

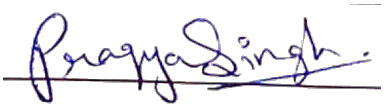
Dr. Pushpendra Singh
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi, India
February 2026

Prof. Mohan Kumar
Department of Computer Science and Engineering
Rochester Institute of Technology, New York, USA
February 2026

Declaration Regarding the Use of Artificial Intelligence (AI) in Thesis

I acknowledge that I am fully responsible for the entire content of this thesis, including any sections that may have been assisted by online tools, including Artificial Intelligence (AI)-based tools. I affirm that all submitted work has been reviewed and validated by me, and I accept full accountability for the accuracy, integrity, and originality of the thesis. Furthermore, I accept responsibility for any violations of ethical, academic, or publication standards arising from the use of such tools.

Signature of PhD Student:



Name of PhD Student: Pragya Singh

Roll Number: PhD21004

To my maa, my shining star ★

Your curiosity, creativity, and compassion inspire me every day ♡♡♡

Abstract

As artificial intelligence-enabled wearable devices become increasingly integrated into everyday life, they offer new opportunities to continuously monitor physiological signals and support emotional well-being outside clinical settings. Such devices are already capable of monitoring basic health indicators, such as activity levels and heart rate, yet their ability to reliably infer emotional states and support real-world monitoring of mental well-being remains limited. Early research in physiological emotion recognition demonstrates technical feasibility within lab settings, but several challenges must be addressed before these algorithms can be effectively deployed at scale in real life. Key challenges include data availability, data consistency, human-centred validity, and real-world usability. Large, diverse, high-quality labelled emotion datasets remain scarce, limiting robust model development. Existing datasets are often heterogeneous in sensing modalities, devices, experimental protocols, and annotation methods, which complicates benchmarking and hinders model generalisation across contexts. Additionally, emotion labelling inherently involves subjective interpretation, yet many data collection practices insufficiently incorporate participants' lived experiences, leading to inconsistent annotations and variable data quality. Finally, even technically capable models must align with user needs, expectations, and deployment realities to deliver meaningful support for mental well-being.

With everyday mental well-being as the primary application, this dissertation addresses these challenges by (1) systematically analyzing and benchmarking heterogeneous physiological emotion datasets, (2) understanding participant's perspectives and developing human-centered approaches for collecting more authentic and reliable emotion-labeled data, (3) introducing new datasets and modeling approaches aimed at improving robustness and generalizability, and (4) examining user perspectives to inform the design of deployable mental well-being technologies. Together, these contributions advance the design of physiological emotion datasets, data-collection methodologies, and modelling strategies aimed

at improving ecological validity, robustness, and generalisability in emotion recognition research. This dissertation, therefore, offers a comprehensive framework to support the development of data-driven interventions for emotional well-being in everyday settings. Collectively, the work moves the field closer to a future in which wearable intelligent systems can more reliably interpret emotional experiences and deliver meaningful, context-sensitive support for mental well-being in everyday life.

Acknowledgements

Long journeys change you over time. And this one was definitely an evolving one. It made me question, rethink, learn, relearn, try, fail, and try again. It made me humble. It made me compassionate. It made me anxious, it made me laugh, and it taught me so much about Emotions in between. It was definitely a personal revelation. I learned a great deal from each of my papers. It all started during the COVID pandemic, in the years of lockdown and uncertainty, when everyone and everything seemed to have come to a halt. And nature slowly found its glory in the absence of humans and their chaos. In these years of ample time, working from home, and vivid imagination, I decided to take a leap of faith and leave my job to pursue a PhD at IIITD, without knowing what it meant or why. I was completely unprepared. I didn't know what research meant, or how to choose what I want to work on, or what I would do. All I knew was that I was an embedded engineer, and I love machines and giving them souls. I never realized that deep down, it was all an emotional journey. Deep down, the little girl in me was trying to understand herself, her emotions, her struggles, and a way out. I began by understanding what it meant to be an HCI researcher, the empathy required, and the ability to not let empathy and compassion drive my findings. But to see beyond and find what data really meant. To see the meaning hidden in words and make sense of it all. Let's not get technical because this chapter is about the creative person in me :) Soon after reading enough, I realized I want to pursue something in between the latest technology and devices, and something that my heart truly wants to pursue without any pressure. Where can I find my flow state, where I can discover my Ikigai - something that I feel is worth doing, while keeping my options open to make an income from it? That's when I realized it is not just me; the world is as confused, and deep down, we all need a machine that can help us figure out what we feel, why we feel it, and how we can all make life a little less overwhelming. I realized the crises we are all in, the changing landscape of information, the availability of choices, the slow impact of COVID, and the isolation it brought, and how

we are all somewhere fighting our own emotional battles. This is exactly what motivated me to pursue this research. Finding meaning beyond words, finding meaning in what is deeply personal to all of us, and what truly impacts our lives daily. Our Emotions. And that's how it all started, emotionally, not going into specific technical details here. But now, looking back, it has made all the difference. And it had made me a better person. And someone who knows how hard it is to show up, to keep trying, to find joy, and to let sadness be there and not consume you. And someone who knows that life is only about choices in the face of what we cannot control - the vast randomness around us. It made me understand that we all pass through ripples of deep happiness, exhilarating thrills, moments of pure bliss, and the depth of melancholy, awkward silence, and an utter desire to cry it all out and give up on the efforts. And yet it taught me to hold it all, giving each and every one of these emotions a space to live, letting them follow their natural cycles and find their way out, bringing forth the next part of our journey. Finally, after I poured my heart out in the most abstract way possible. I would like to acknowledge those who have been part of this incredibly beautiful, challenging, yet deeply spiritual journey of mine. I want to start with my Maa. I really love you, and you already know how much of this is because of you. You are and will always be my source of inspiration, kindness, and ability to keep trying. Papa, you have taught me to be a problem solver. Your ability to solve it all without us even knowing about it motivates me to try even harder. You taught me that it's ok to try and fail and have the confidence to try again, because in the end, we will find a solution somehow. Vasu, Somu, Ritik, and Khushi, you are my inspirations because, in your unique ways, you all make me push harder. Vasu, your ability to question and look beyond and learn, Ritik, your ability to analyze and find the big pictures in what no one else notices, Somu, your empathy and strength to do the impossible with pure will and courage, Khushi, your dedication and commitment to giving your 100% every time, you all are amazing. I don't want to forget my Amma. You are an inspiration to all of us. Your struggles and stories matter. Your inquisitiveness and curiosity to learn are my forever template for keeping learning and finding joy in it. Chacha Chachi,

thank you for your silent support, constant encouragement, and for your kind words and love. Thank you for being there always. Also, Mummy, I know all you want to see is for me to grow and be fierce like you. You are a fighter. And you taught me to fight for what matters, to be bold, and to deliver the best in my abilities. I learn that from you every day. You're all that makes me who I am and who I will be. And it is surreal as I write this on a train journey back from home to Delhi.

Now, I would like to acknowledge both my advisors. Pushpendra sir, thank you for all your advice and for believing in my ideas. Your advice truly made everything I do so much better. But out of all, I will always remember: *"No research idea is good or bad, no area is exceptional or not interesting, it is your true intention, willingness, passion, and effort to truly solve a problem that makes all the difference."* These words made me pursue research from my heart. Mohan sir, thank you for all the feedback and advice, and for always encouraging me to look beyond the paper to make a contribution that truly matters to society. Whether it's health and well-being or care for older adults, you always told me to consider the broader impact of the research we do. You always encouraged me to try more and look deeper. Thank you both for all your support and feedback, and for guiding me through. I also want to thank Pankaj Jalote Sir for encouraging me to always start from the beginning and consider the broader impact before every new research question and idea. Saket Anand Sir and Sujay Deb Sir from IIITD, for their valuable feedback as members of my internal committee. Dr. Koushik, Dr. Swati, and Suruchi from AIIMS Delhi, for sharing their wisdom on mental health research from the perspectives of domain experts. Swarnava Dey, Avik Ghose, and Shalini Mukhopadhyay from TCS Research, for my first published research work outside IIITD, which gave me the initial confidence I needed.

They say we have 5 to 6 people in our inner circle, while a few remain the core of who we are; we also have changing members we meet along our life's journey. These are our people outside our families, whom we choose and who choose us to live our lives beyond our sacred and safe shelters. I have met many such people in my life and during my PhD journey,

and all of them made a difference. Deepak, thank you for being my support system and appreciating me through it all. Without you, I would have felt less confident in trying this out. Thank you for sowing the seeds of research in me. And clapping for all my achievements and efforts. And showing me the positives hidden behind the negatives. Next comes my support system at IIIT-Delhi, Akshit. How would I be thanking you? Your encouragement, empathy, critical feedback, and unbiased opinions were crucial in this journey. I learned so much from you. You made me understand what's beyond emotions and see the other side of it. You have given strength to my ideas, and without you pointing out the grammatical errors, it would surely have been difficult for Grammarly alone. Jokes aside, thank you so much for listening to me and reading through the initial drafts. You saw the messy sides of me and my published versions. Asrar, my buddy, my gym partner, my unofficial life-advice council. Thank you for your appreciation and encouragement. In a world where everyone starts with criticism to make you tougher, you always start with trust, kindness, and positivity. Keep that attitude. It truly lights up the world. Bushra, you met me at the perfect time. I know that you will always be there. Your cheerful disposition, positive attitude, and mischievous angles added the spice to my PhD journey. Not all friends in your PhD journey discuss research; some prefer to talk about photographs, stars, and travels. And they remind you that there is more to life. They make you laugh, enjoy, and remember the journey. They make you live, see, take risks, and enjoy without overthinking too much! Rajan, you are a gem of a person. Your silence says a thousand words. And so do your photographs. Thank you for all your kind words and advice. And for giving me the IIITD experience that many people miss out on. Now, whenever I see birds, owls, cats, dogs, and stars (especially Orion), I feel a sense of wonder. I will remember you. Tanya, Depanshu, Karan, Jatin, Mayuri, Ekta, Ramyanee, Aayushi, Shivaksh, thank you all for being a part of my journey. Research can happen at any time and is not bound by space. But labs exist for a reason, and so do lab mates; they make you feel the community. They help you bear the reviewer 2. They multiply the joy of acceptance and divide the sadness of a rejection. Anupriya,

Asra, Jasmeet, Shyama, Kirti, Gautami, Sara, Manshul, and Rhea. Thank you for being there. You are all amazing researchers, curious thinkers, and, above all, believers. I know that, with your research, you will all continue to work towards making this world a better place. I would also like to acknowledge my co-authors and student mentees: Ritvik, Anshul, Ankush, Somay, Prashasti, Hitesh, Krishnam, Kanishk, Sakshat, Jayant, Samridh, Jogith, Vaibhav, Abhijay, Mohit, Hardikya, Hardeekh, Aniket, and Sahil. Thank you all for being a part of my journey. It has been a pleasure working with each of you. I wish you all every success in your future endeavors, and I look forward to seeing the impactful work you will continue to do. Lastly, I extend my sincere gratitude to all my study participants; without your time, openness, and willingness to share your experiences, this research would not have been possible. Finally, I would like to thank the many people, both known and unknown, whom I encountered throughout my PhD journey and who, in small but meaningful ways, have inspired me. I am also deeply grateful to the researchers whose prior work inspired me to pursue this field and laid the foundations for this dissertation.

Pragya Singh

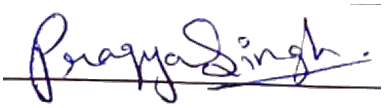


Table of Contents

List of Tables

List of Figures

Publications

Publications Relevant to this Thesis

- **Pragya Singh**, Ritvik Budhiraja, Pankaj Jalote, Mohan Kumar, and Pushpendra Singh. 2025. Translating Emotions to Annotations: A Participant's Perspective of Physiological Emotion Data Collection. Proc. ACM Hum.-Comput. Interact. 9, 2, Article CSCW195 (May 2025), 30 pages. https://doi.org/10.1145/3711093 [1]

- **Pragya Singh**, Ankush Gupta, Mohan Kumar, and Pushpendra Singh. 2025. AnnoSense: A Framework for Physiological Emotion Data Collection in Everyday Settings for AI. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 9, 3, Article 131 (September 2025), 47 pages. https://doi.org/10.1145/3749519 [2]

- **Pragya Singh**, Ritvik Budhiraja, Ankush Gupta, Anshul Goswami, Mohan Kumar, and Pushpendra Singh. 2024. EEVR: a dataset of paired physiological signals and textual descriptions for joint emotion representation learning. In Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24), Vol. 37. Curran Associates Inc., Red Hook, NY, USA, Article 503, 15765–15778. [3]

- **Pragya Singh**, Gupta, A., Jalan, S., Kumar, M., and Singh, P. (2025). FEEL: Quantifying Heterogeneity in Physiological Signals for Generalizable Emotion Recognition. In The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track. [4]

- **Pragya Singh**, Mohan Kumar, and Pushpendra Singh. "Can we say a cat is a cat? Understanding the challenges in annotating physiological signal-based emotion data." arXiv preprint arXiv:2406.14908 (2024). [5]

Other Works

- Sara Moin, Manshul Belani, **Pragya Singh**, Nishtha Phutela, and Pushpendra Singh. 2025. "But I Won't Say That It Was Bad Seeing a Real Vagina": Understanding Perspectives toward Learning Sensitive-Critical Health Topic. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 697, 1–15. https://doi.org/10.1145/3706598.3713807

- S. Mukhopadhyay, S. Dey, A. Ghose, **Pragya Singh** and P. Dasgupta, "Generating Tiny Deep Neural Networks for ECG Classification on Micro-Controllers," 2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), Atlanta, GA, USA, 2023, pp. 392-397, doi: 10.1109/PerComWorkshops56833.2023.10150311.

Chapter 1

Introduction

"I put my heart and soul into my work, and have lost my mind in the process."

— Vincent Van Gogh

As Vincent Van Gogh observed several decades ago, these words remain strikingly relevant in today's fast-paced society, where maintaining good mental well-being has become increasingly challenging. Escalating pressures to remain productive and competitive across work, education, and personal life have contributed to rising stress levels, reduced quality of life, and a growing mental health burden affecting a substantial portion of the global population [6, 7, 8]. At the core of these challenges lies an individual's emotional state, which profoundly influences daily functioning, decision-making, and overall mental well-being. In response, researchers in human–computer interaction (HCI), affective computing, and ubiquitous computing have increasingly turned toward computational approaches for modeling and recognizing human emotions [9, 10, 11]. Beyond the recognition of momentary emotional states, a growing emphasis in this research is on understanding emotional trajectories over time and leveraging these insights to support more effective emotion regulation. Such efforts aim to enable technological interventions that not only detect emotional experiences as they occur but also help individuals manage them in ways that contribute to sustained mental well-being. In the existing literature, multiple approaches for predicting and recognizing human emotions have been explored. These include visual modalities such as facial images and videos, auditory cues derived from speech and vocal characteristics, and a range of behavioral signals, including body movements, interaction patterns, and smartphone use [12, 11]. More recently, physiological signals captured by wearable sensors, such as heart rate, electrodermal activity, and skin temperature, have attracted increasing attention [13], as they provide continuous, relatively objective measures

that are less susceptible to conscious control or social masking than externally observable affective cues, making them promising for capturing emotional dynamics longitudinally in everyday contexts. Despite the potential of physiological signal–based emotion recognition to support mental well-being in everyday settings, critical gaps persist, slowing progress in the field. The following sections discuss two key gaps that this thesis aims to address.

1.1 Data Challenges in Physiological Emotion Recognition

One of the primary factors contributing to this slow progress is the persistent scarcity of high-quality emotion data needed to develop robust, physiology–based emotion recognition systems for mental well-being applications. Although numerous emotion and affect datasets have been created over the years to support model development, the overall demand for reliable labeled and unlabeled data remains largely unmet. A key factor underlying this persistent data gap is the diversity of contexts in which emotion and affect data are collected, each involving trade-offs between experimental control, ecological validity, and data quality. At one end of the spectrum, laboratory studies use structured emotion induction protocols, such as visual, auditory, or cognitive stimuli, to elicit specific emotional states under controlled conditions [14, 15, 16]. Semi-controlled settings introduce greater realism by engaging participants in constrained real-world activities (e.g., interviews, debates, presentations, or games) designed to evoke emotions while preserving some experimental structure [17, 18]. At the other end, in-the-wild approaches capture emotional experiences during everyday life, often through wearable sensors that enable continuous physiological monitoring. In these contexts, emotional states are typically recorded via self-reports delivered through companion mobile applications or smartphone-based survey prompts [19, 20]. While this diversity of data collection paradigms broadens the field's empirical scope, it also introduces substantial heterogeneity in the types of data collected, directly impacting data availability, consistency, and reuse across the community. This heterogeneity arises from the selection of biosignals, sensing devices, and labeling strategies used

across studies. Different research efforts prioritize different physiological signals depending on their objectives, constraints, and experimental settings, often using non-overlapping sensor configurations. In laboratory environments, where participant mobility is less constrained, researchers can collect high-fidelity signals such as electroencephalography (EEG), electromyography (EMG), electrocardiography (ECG), photoplethysmography (PPG), and electrodermal activity (EDA). In contrast, in-the-wild studies typically rely exclusively on signals readily available from consumer-grade wearable devices, such as smartwatches or rings, thereby limiting data collection to a narrower set of physiological measures [14]. Beyond sensing modalities, substantial variability also exists in how emotional states are labeled. Annotation strategies range from extensive emotion scales to standardized mental health questionnaires administered at various intervals across lab-based experiments [21], to lightweight approaches that rely on single-item self-reports collected periodically in daily life settings [22]. These labeling practices further diverge in their theoretical grounding, with studies adopting different emotion frameworks, including discrete or basic emotion models [23], dimensional representations such as valence–arousal, or hybrid formulations [24, 25]. Collectively, this variability in biosignal selection, sensing technology, and emotion annotation frameworks contributes to fragmented datasets that are difficult to align, compare, or integrate, thereby constraining the effective reuse of data and limiting the development of generalizable emotion recognition models.

1.2 The Missing Human-Centeredness in Data Collection and Intervention Design

Another key challenge constraining the development of physiological signal–based interventions for supporting mental well-being is the strong reliance on human participants throughout the emotion data pipeline, including both the acquisition of physiological measurements and the provision of emotion annotations via self-reports. Consequently, data quality is closely tied to participants' willingness, sustained motivation, and ability to interpret and report their emotional experiences. This reliance introduces multiple sources

of inconsistency in emotion datasets, including missing or irregular annotations, context-dependent variability in self-reports, environmental noise, and signal degradation during physiological data acquisition. Within HCI, substantial attention has been devoted to understanding stakeholder perspectives to design systems that are more participant-friendly and supportive, both in the context of emotion data collection and the development of interventions for mental well-being [26, 27]. Prior work has explored strategies to sustain participant motivation and engagement during data collection, particularly in real-world settings where long-term participation can be challenging, leading to improved interfaces, prompting mechanisms, and wearable interaction designs aimed at reducing burden and enhancing compliance with study protocols [28, 29]. However, comparatively limited research systematically examines how human participation itself shapes the quality, consistency, and long-term reliability of emotion data [30]. While usability and engagement are often considered, factors such as individual differences in emotional interpretation, contextual influences on reporting, and psychosocial variability and their implications for both data quality and downstream intervention design remain underexplored. Moreover, participants' perspectives are still insufficiently integrated into the design of interventions intended to support emotion regulation and sustained mental well-being [9].

This limitation reflects a broader and long-standing conceptual tension in psychology and affective science between *nomothetic* and *idiographic* approaches to understanding human emotions. Nomothetic approaches aim to derive generalizable laws of emotion by aggregating data across individuals, implicitly assuming that emotional states and their physiological correlates can be meaningfully aligned across populations. In contrast, idiographic approaches emphasize within-person structure, variability, and meaning, treating emotional experience as fundamentally shaped by individual interpretation, context, and temporal dynamics rather than universal mappings. Classic work in psychology has long highlighted the challenges of generalizing from group-level statistics to individual-level processes, arguing that inter-individual averages may obscure stable intra-individual dynamics

[31, 32]. In physiological emotion recognition, the prevailing paradigm has largely followed a nomothetic orientation, in which models are typically trained on pooled datasets under the assumption that physiological signals such as heart rate variability, skin conductance, or electrodermal activity carry consistent emotional meaning across individuals [13]. However, relatively limited attention has been given to the design of idiographic methods for both data collection and modeling in emotion recognition systems. With recent advances in natural language processing and large language models, there is a renewed opportunity to revisit and operationalize idiographic approaches in this domain, enabling more personalized and context-sensitive representations of emotional experience. This thesis builds on this methodological tension by adopting an idiographic-oriented perspective on emotion data collection and interpretation. Rather than treating inter-individual variability as noise to be minimized, it is treated as a contextual participant-specific property of emotion data that must be explicitly modeled. In doing so, the work aligns with recent trends in personalized sensing and computational modeling that reconsider the limitations of purely population-level inference in favor of individualized or hybrid modeling approaches. Overall, there remains a need for a deeper understanding of how the subjectivity, behaviors, and contextual constraints of human participants shape both emotion datasets and the design of deployable interventions. Addressing these gaps is essential for developing signal-based, data-driven mental well-being interventions that are technically robust, grounded in real human experiences, and capable of supporting sustained emotional well-being across diverse contexts.

1.3 Research Questions

Together, these challenges in emotion recognition research underscore the need to examine emotion data through both data-centric and human-centered lenses. This dissertation addresses these interrelated challenges to generate insights that inform more robust data collection practices, improved modeling approaches, and the design of human-centered

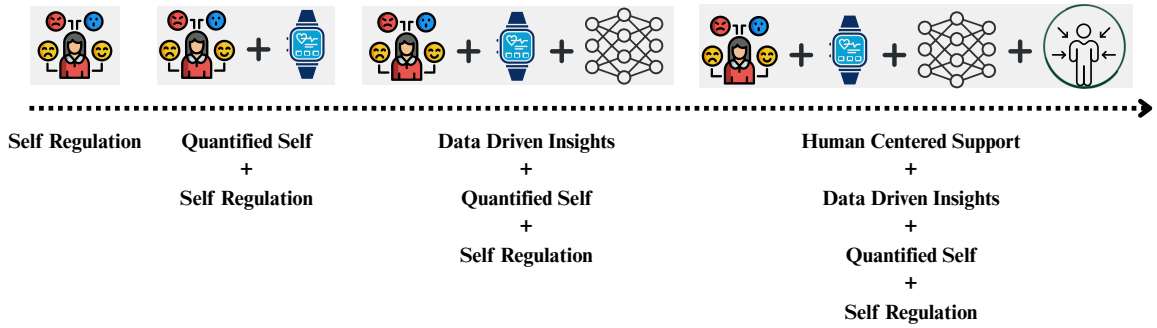


Figure 1.1: The main contributions of this thesis are towards adding a human-centric perspective to data-driven emotion recognition research.

interventions to support mental well-being. The work is guided by the following research questions:

- **RQ1:** How are emotions currently captured, labeled, and represented in physiological emotion recognition research, and what opportunities and limitations arise from existing data collection practices?
- **RQ2:** How can we design newer data collection approaches and AI models that are both data-driven and human-centered, leveraging physiological signals while accounting for participant variability and context?
- **RQ3:** How do participants experience and interpret emotion data collection and labeling processes in both controlled and in-the-wild settings, and how do these experiences influence the quality and reliability of the collected data?
- **RQ4:** How can data-driven emotion self-tracking systems be designed to provide meaningful, sustainable, and participant-aligned support for emotional awareness, regulation, and well-being in everyday life?

1.4 Research Contributions

My dissertation makes the following key contributions to the field of physiological emotion recognition:

- **Large-Scale Benchmarking and Analysis of Data Heterogeneity**: I conducted the first large-scale benchmarking and data quality analysis of 19 physiological emotion datasets to examine how heterogeneity in demographics, experimental settings, sensing devices, and labeling strategies influences model performance [4]. By systematically studying these variations, I provide insights into the challenges of learning across diverse datasets and inform the design of modeling approaches that are more robust to dataset differences, enabling the training of more generalizable emotion recognition models.

- **New Datasets and Modeling Methods for Capturing Subjective Emotional Responses**: I introduce a novel physiological emotion dataset (EEVR) [3] that extends beyond typical datasets by explicitly capturing the subjectivity inherent in participants' emotional responses alongside physiological signals. This dataset incorporates richer contextual and self-report information, enabling a more nuanced understanding of emotional experience. In parallel, I develop and evaluate a contrastive modeling approach designed to leverage this dataset effectively, supporting the development of more robust emotion recognition models that better account for individual variability.

- **Human-Centered Insights into Emotion Data Collection**: I contribute insights into emotion data collection by critically examining the current practices. Through mixed-method investigations in both controlled (laboratory) [1] and everyday (in-the-wild) [2] settings, including interviews, focus groups, surveys, and laboratory studies, I show how limited consideration of participants' subjectivity, context, and interpretive processes contributes to data quality issues, inconsistencies, and reliability challenges. Building on these findings, I propose human-centered strategies to address these challenges and offer design and methodological guidance to better align emotion data collection with how emotions are experienced and expressed in everyday life.

- **Human-Centered Intervention Design and Application Prototypes**: Finally, I

contributed to designing participant-centered approaches for collecting emotion data in everyday contexts and developing emotional well-being interventions through in-the-wild deployment studies. Building on the findings, I have provided insights into participants' needs and intrinsic motivations, informing the design of future human-centered interventions for emotional well-being.

Overall, this dissertation advances the field of physiological emotion recognition by integrating human-centered insights into AI data and modeling pipelines. Collectively, these contributions provide a holistic framework for designing more robust, interpretable, and human-aligned emotion recognition systems that are applicable in everyday contexts and supportive of mental health. Finally, I would like to emphasize that, although the dissertation includes work conducted in collaboration with others, I served as the primary researcher responsible for conceptualizing, designing, executing, and analyzing all the studies presented. I was the first author on all core publications included in this dissertation and led all major research activities, including identifying research gaps, formulating research questions, designing methodologies and experiments, conducting data collection and analysis, running computational experiments, and writing and revising the manuscripts. The use of "we" throughout the dissertation reflects collaborative contributions, while the central research activities were independently led and carried out by me.

1.5 Organization of Thesis

This dissertation is organized as follows. Chapter 2 presents the background of emotion recognition research, beginning with an overview of foundational emotion theories and their influence on how emotions are conceptualized and studied. It then examines human-centered perspectives and their role in shaping prior work in mental health research. The chapter further reviews existing approaches to physiological sensing for emotion recognition, including relevant datasets, data collection practices, and modeling techniques that have informed the field to date. Lastly, I discussed past research on developing methods to support

the collection of emotion data. In Chapter 3 (RQ1), I examined existing physiological emotion datasets in detail, focusing on how heterogeneity manifests across data collection settings, sensing modalities, and labeling practices. I then present large-scale benchmarking results, including both within-dataset and cross-dataset evaluations, to demonstrate how these sources of heterogeneity affect emotion recognition performance and model generalization. In Chapter 4 (RQ2), I demonstrate how human-centered data, grounded in participants' perspectives, can be effectively utilized in practice. The chapter introduces the EEVR dataset and a contrastive modeling approach to design a new emotion recognition architecture. In Chapter 5 (RQ3), I examine stakeholders' perspectives on current emotion data collection practices in both laboratory and real-world settings and analyze how these practices affect data quality. The chapter presents findings on how to design data collection approaches with participants' challenges, attitudes, and behaviors in mind, leveraging rather than constraining human behavior to support more effective, human-centered emotion data collection. In Chapter 6 (RQ2), I present a multimodal, participant-centric emotion annotation system designed to capture emotional experiences in ways that reflect users' context, availability, and expressive preferences, demonstrating how such approaches can improve both data richness and participant engagement. In Chapter 7 (RQ4), I build on this work by examining how emotion self-tracking systems can provide meaningful and sustainable support for emotional well-being through a multi-phase user study exploring engagement, perceived value, and long-term adoption. Finally, in Chapter 8, I conclude by discussing the integrated summary of thesis across the chapters, followed by outlining key directions for future research informed by this dissertation's contributions. An overview of the dissertation structure is provided in Table 1.1.

Chapter (Publication)	RQ	Methods	Contribution
Chapter 3 [4]	RQ1	Benchmarked (n = 19 Datasets) using Machine Learning and Deep Learning Models	Examines heterogeneity across physiological emotion datasets and demonstrates how differences in sensing modalities, labeling practices, and collection settings affect emotion recognition performance and cross-dataset generalization.
Chapter 4 [3]	RQ2	Dataset Collection (n = 37), and Benchmarking	Introduces the EEVR dataset and a contrastive emotion recognition architecture integrating physiological and textual descriptions for richer emotion modeling.
Chapter 5 [1, 2]	RQ3	**Part 1**: Lab study (n = 37), **Part 2**: Surveys (n = 75), Interviews (n = 32), Focus Groups (n = 3, involving 12 mental health professionals), and Expert Evaluation (n = 25)	Investigates stakeholder perspectives on emotion data collection and identifies principles for designing more human-centered and ecologically valid datasets.
Chapter 6	RQ2	Field study (n = 33), and Interviews (n = 33)	Introduces a participant-centric multimodal emotion annotation system.
Chapter 7	RQ4	Surveys (n = 145), Field study (n = 42), and Interviews (n = 10)	Explores long-term emotion self-tracking and demonstrates the challenges of sustained engagement in emotion well-being interventions.

Table 1.1: Overview of dissertation chapters, publications, methods, and contributions.

Chapter 2

Background and Related Work

This chapter provides a detailed overview of (1) current practices in physiological emotion data collection, and (2) the role of participants in everyday emotion data collection.

2.1 Physiological Emotion Data Collection Practices

Emotion recognition research initially focused on identifying behavioral cues, such as facial expressions, vocal cues, text data, or movement [12, 33, 11]. With the advent of wearable sensing technologies, researchers increasingly leveraged physiological signals to capture internal states underlying emotional experiences [34, 35]. Physiological signals commonly used for emotion recognition include heart rate (HR) and heart rate variability (HRV) measured via electrocardiography (ECG) or photoplethysmography (PPG), skin conductance measured as electrodermal activity (EDA), muscle activity recorded through electromyography (EMG), brain activity captured via electroencephalography (EEG) or functional magnetic resonance imaging (fMRI), physical activity measured using acceleration (ACC) or gyroscope and skin temperature (SKT) [36]. These signals provide direct insight into internal bodily processes and are less susceptible to conscious control or intentional masking, complementing behavioral and expressive cues while enabling continuous, objective monitoring. To study these signals, researchers have collected data across a spectrum of experimental settings, ranging from controlled laboratory studies to semi-naturalistic environments and real-world contexts. Laboratory studies provide high experimental control, allowing researchers to elicit specific emotions through images, videos, audio, virtual reality (VR), games, or cognitive tasks [37, 38, 39, 40, 3]. Semi-naturalistic studies introduce more realistic tasks, such as interviews, driving simulations, or phone-based activities, enabling the capture of contextually rich emotional responses while maintaining some

control over stimuli and timing [21, 41]. Real-world, in-the-wild data collection captures participants' emotions as they naturally unfold, often using wearable devices alongside ecological momentary assessments (EMA) or end-of-the-day prompts [11, 42, 19]. Tables 2.1 and 2.2 present the list of datasets collected in the past. Across these three settings, the choice of physiological modalities is closely tied to the experiment constraints and research objectives. Laboratory-based studies typically prioritize high-fidelity signals, such as multi-lead ECG, high-resolution EEG, and precise EMG recordings, which require stable, motion-restricted environments to minimize artifacts and ensure accurate measurement. The controlled nature of the lab allows researchers to systematically manipulate stimuli, elicit targeted emotional responses, and capture detailed physiological data. In contrast, semi-naturalistic or real-world studies, which aim to observe emotions as they occur in daily life, often rely on portable, wearable, or unobtrusive sensors that can be comfortably worn over extended periods. These modalities include signals often available in wearable devices, such as PPG, EDA, and skin temperature, etc. As a result, researchers often balance the fidelity and richness of the physiological signals with the practicality and realism of data collection, a trade-off that directly shapes the type of datasets available and the subsequent design of emotion recognition models.

Furthermore, emotion annotation strategies used in physiological datasets are also diverse, reflecting both theoretical frameworks and practical considerations. Self-reports via standardized questionnaires such as the Self-Assessment Manikin (SAM) [25] or Positive and Negative Affect Schedule (PANAS) [37] are commonly used, alongside experience sampling methods, continuous annotation tools [64, 65, 66], expert or external raters [67], and reflective self-reporting techniques [68]. Hybrid approaches combining multiple labeling methods have also been explored to capture nuanced emotional responses. Theoretical grounding of emotion labeling varies widely. Early evolutionary theories, such as Plutchik's psycho-evolutionary model [69] and Ekman's Basic Emotion Theory [23], focus on biologically hardwired, universal emotions. Appraisal theories, pioneered by Lazarus [70],

Dataset	Elicitation Method	Annotation Approach
WESAD [37]	Video Clips, Public speaking, mental arithmetic, and Meditation	PANAS, SAM, State-Trait Anxiety Inventory (STAI), Short Stress State Questionnaire (SSSQ), physiological signals
ASCERTAIN [43]	Video Clips	Valence-Arousal, Engagement, Liking, Familiarity, Personality Traits
CASE [44]	Video Clips	Continuous Valence-Arousal Annotations
Neurological Status [45]	Physical Activities	Task-based Labels
CLAS [46]	Video Clips, Images, Math, Stroop, Logic tasks	Arousal-Valence, Task-based Labels
VREED [47]	VR Video Clips	SAM, PANAS
POPANE [48]	Speech preparation, Anticipation task, Interpersonal communication, Affective Images, and Video Clips	Discrete Emotion Categories, SAM, Avoidance Approach Motivation
EMOGNITION [49]	Audio-visual stimuli	Discrete Emotion Categories, SAM, Avoidance Approach Motivation
StressID [50]	Cognitive load tasks	SAM, Custom Perceived Stress Assessment
BIRAFFE2 [51]	Games, Affective Music, and Images	SAM, Game Experience Questionnaire (GEQ)
EEVR [3]	VR Video Clips	Textual Descriptions, SAM, PANAS, Familiarity, Liking, Personality Traits
KEMOCON [52]	10-minute-long debate on social issues	Self-report Valence-Arousal, Discreet Emotion Category, Partner Annotations, and Expert Annotation
AMIGOS [53]	Video clips (Long and Short)	SAM, PANAS, Personality traits
RECOLA [54]	Collaborative task (video chat)	SAM, PANAS

Table 2.1: Lab-based emotion datasets: Elicitation methods, annotation strategies, and labeling approaches

Dataset	Context/Task	Annotation Method
ForDigitStress [55]	Job interview tasks simulating time pressure	Custom Stress Scale and Saliva Cortisol
NURSE [56]	Healthcare workers during COVID-19	Custom Stress Questionnaire
G-REx [57]	Long movie viewing sessions	Post-Hoc SAM Scale Based Tool
Laureate [58]	University setting with student academic routines	Custom EMA (PANAVA-KS, physical activity, breakfast ingestion, caffeine intake, study-time and sleep quality)
StudentLife [11]	University campus life over multiple weeks	Photographic Affect Meter (PAM) EMA, Single-item Stress EMA
GLOBEM [59]	Naturalistic daily experiences across diverse locations	EMA Survey (PHQ-4, PSS-4, PANAS), and Pre-Post Survey
TILES [60, 60]	Workplace monitoring in hospital environment	Single-item Stress EMA, Survey on daily stressors, work behaviors, and sleep
DAPPER [61]	Daily life across varied settings (field study)	20-Item ESM (Information about daily events, Participants' openness to sharing emotion, TIPI-C, PANAS), DRM with Open-ended Question
K-EmoPhone [62]	Daily life across varied settings (field study)	Custom Questionnaire (Valence, Arousal, Attention, Stress, Emotion Duration, Task Disturbance, Emotion Change)
SWEET Study [63]	Office workers' daily routines in real-life settings	EMA (Stress, Activity, Food and Beverage Consumption, Sleep Quality, and Gastro-intestinal Symptoms)

Table 2.2: Tasks and Annotation Methods in Semi-Controlled Emotion Datasets.

emphasize the cognitive evaluation of events and the subjectivity of emotional responses. More recent constructivist perspectives, such as Barrett's Theory of Constructed Emotion [71], view emotions as actively constructed from sensory input, prior experiences, and cultural context. These theoretical frameworks have informed the development of diverse annotation tools, including dimensional models like the circumplex model [72], operationalized through scales such as the SAM, Geneva Emotion Wheel (GEW) [24], and EmojiGrid [73], as well as discrete emotion measures including PANAS [74], Emotions Twenty Questions [75], and Ekman's six basic emotions [23]. Custom questionnaires and standard mental health instruments are also commonly employed to supplement these approaches.

Despite the proliferation of emotion datasets collected across laboratory, semi-naturalistic, and real-world settings, each dataset is constrained by its collection protocols. Many datasets are small, lack participant diversity, and focus on limited tasks or scenarios, limiting their generalizability and representativeness. Moreover, differences in experimental setups, annotation strategies, and the types of physiological data collected make it challenging to systematically compare datasets or combine them to build larger, more generalizable models. Most datasets still rely on relatively simple approaches for collecting emotion labels, such as standardized scales, discrete categories, or researcher-driven methods. While several interactive or context-aware annotation approaches exist, their usability is often limited, and they are rarely adopted in actual dataset collection, as evident in Tables 2.1 and 2.2. As a result, most datasets fail to fully capture the richness of participants' subjective experiences or the nuances of situational context. Cognitive, motivational, and perceptual biases further affect annotation reliability [76, 77, 78]. Together, these factors contribute to the lack of large-scale, high-quality, human-centered emotion datasets, constraining model generalization, reproducibility, and progress in the emotion recognition field. My research aims to address these data challenges by understanding and designing participant-centered data collection approaches.

2.2 Participants' Role in Everyday Emotion Data Collection

Emotion self-tracking has its roots in psychological and behavioral research, where individuals manually recorded and reflected on moods, feelings, and daily experiences through diaries, questionnaires, or other reflection exercises [79]. While these manual approaches provided valuable introspective insight, they were limited by memory biases, inconsistent logging practices, privacy concerns, and the cognitive effort required for regular documentation. With the proliferation of mobile and wearable technologies, emotion self-tracking has become increasingly digital and accessible. Active emotion self-tracking systems typically rely on notifications or reminders to encourage users to reflect and self-report their emotional states at regular intervals [80, 81, 82]. These systems are primarily user-driven, requiring individuals to actively log their moods and emotions, enabling them to quantify their emotional landscapes, recognize patterns over time, and gain insights into their personal affective dynamics. While passive self-tracking leverages digital behavioral and contextual data as proxies for affective states, including physiological signals, voice, facial expressions, mobile phone usage, location, and physical activity [83, 84, 85, 86]. These passive approaches aim to reduce user effort while providing objective, quantified insights that can support emotional awareness and regulation [87, 88, 89, 13]. While early research demonstrated high emotion recognition accuracy in controlled laboratory environments [37, 53, 90], these emotion recognition algorithms often underperform in real-world contexts [91, 92]. Emotional experiences in naturalistic settings are inherently subjective, context-dependent, and dynamic, challenging the reliability of passive monitoring approaches [93, 94].

To enhance the performance of passive sensing systems, researchers have increasingly explored methods for collecting emotion data in real-world settings. These approaches often combine passive monitoring with active self-reporting, which are essential for training and validating physiological and behavioral models [91, 95]. By capturing subjective emotional experiences alongside continuous physiological signals, these hybrid pipelines

provide ground truth labels that contextualize and improve the interpretability of sensor data. However, the effectiveness of such systems depends heavily on user engagement and sustained participation. Active self-tracking requires participants to consistently log their emotions, and research shows that engagement is influenced by multiple factors, including personal interest, emotional comfort, privacy concerns, cognitive effort, and the perceived stigma of sharing emotions [29, 82, 96, 97, 2, 98]. Users often express skepticism regarding the utility and reliability of digital self-tracking tools and interventions compared to human-provided care or social support, and engagement can vary with symptom severity, personal context, and individual needs [99, 94, 100, 101]. From a physiological emotion recognition perspective, these engagement challenges directly impact data quality, coverage, and model performance. Inconsistent or sparse self-reports can reduce the reliability of supervised learning approaches, while missed contextual information may lead to models misinterpreting physiological signals. Consequently, integrating participants' subjective experiences into the data collection pipeline is critical for capturing meaningful, context-aware emotion labels.

To address these challenges, human-centered design approaches are increasingly employed in emotion data collection. These strategies prioritize participants' needs and preferences, incorporating features such as lightweight prompts [28, 102], context-aware prompts [103], or personalized prompts [104, 105, 106] to reduce cognitive load and emotional burden. In recent years, interactive, technology-assisted, and human-centric emotion self-reporting methods have also emerged to address some limitations of static scales and support user engagement. Examples include web- and mobile-based tools such as Find the Bot [107] and Reconexp [108], reflective journaling platforms like mirrorU [109], and image-based self-report tools such as the Affect Grid [110], Premo [111], and Photographic Affect Meter (PAM) [112]. Other innovative approaches include opportunistic probing frameworks (PResUP [113]), AI-assisted reflective systems (Mirror Ritual [114], Mirror Hearts [115]), Technology-Assisted Reconstruction (TAR) [116], and methods leveraging circadian pat-

terns in emotions [117]. More recently, LLM-based self-reporting and in-context journaling systems, such as DiaryHelper [118], Mindshift [119], Diarymate [120], and Mindscape [121], have been explored to support behavioral monitoring in daily life. Research has further emphasized participatory and stakeholder-informed design methods, considering the perspectives of clinicians, individuals with or without mental health conditions, and at-risk populations [102, 122, 123, 120, 124, 2, 125, 27].

Despite these advances, studies show that engagement with both emotion data collection studies and mental health interventions often declines over time, particularly when systems require frequent manual input, present emotionally burdensome prompts, fail to protect privacy adequately, or provide feedback perceived as unhelpful or misaligned with user needs [29, 126, 97, 80, 127, 128]. Further research has shown that participants often struggle to accurately self-report their emotions, as reflecting on and labeling feelings in real time can be difficult or cognitively demanding [129, 130, 29]. This can create a gap between experienced emotions and reported labels, introducing noise and variability into emotion datasets [76, 77, 26]. These findings underscore the critical role of participants not only in designing participant-centered interventions but also in collecting reliable, high-quality emotion data in everyday settings. Moreover, they also highlight the importance of studying participants' experiences in emotion data collection to better understand how subjective interpretations, situational context, lived realities, and engagement influence emotion labels and overall data quality. In this dissertation, my research examines these challenges from participants' perspectives, focusing on how their experiences influence the quality of emotion datasets and how data collection methods and deployable interventions can better reflect their lived realities.

Chapter 3

How We *FEEL*? Quantifying Heterogeneity in Emotion Data

Despite the potential of physiological emotion recognition to support everyday emotional well-being, these models face a major barrier: heterogeneity across datasets. Differences in experimental settings (e.g., lab versus in-the-wild), sensor types and configurations, emotion elicitation methods, and labeling strategies create domain shifts that make it difficult for models to generalize across datasets or real-world contexts [131, 132]. Most models are therefore trained and tested on isolated, homogeneous datasets, limiting their scalability, reproducibility, and practical impact [133, 132]. Unlike fields such as computer vision or natural language processing, where standardized, large-scale datasets enable robust, cross-domain models, physiological emotion recognition is constrained by reliance on human participants, specialized sensors, and ethical considerations [1, 2, 13]. Most publicly available datasets are small, structurally diverse, and collected under varying protocols and labeling schemes, making them difficult to combine or leverage effectively. This fragmentation hinders data harmonization, reduces model generalizability, and slows progress toward large-scale, reliable emotion recognition systems [131, 132, 133]. To move toward scalable, high-impact physiological signal-based emotion recognition systems, we need to treat data as a **"shared resource"** and harmonize datasets across key dimensions, such as signal representations and labeling strategies. This is vital not only for enabling large-scale training and effective domain adaptation but also for establishing fair and reproducible benchmarks. In the absence of such coordination, research remains fragmented, and findings from one dataset may fail to generalize to others. Furthermore, the physiological emotion recognition research also lacks a systematic, dataset-level benchmark for evaluating emotion recognition models across widely used physiological signals. Such a benchmark would enable standardized model evaluation, facilitate signal-specific insights, and support

assessment of generalization across datasets. By providing consistent pre-processing and labeling protocols, it promotes fairness, reproducibility, and meaningful comparison. This foundation is essential for accelerating progress toward deployable emotion recognition systems in real-world, wearable contexts.

To address the lack of standardized evaluation for heterogeneity and its impact on model performance, in this chapter, we curated a diverse collection of **19 publicly available datasets** covering a wide range of experimental conditions and labeling strategies (as detailed in appendix A.1. We performed a **meta-analysis** of data quality and benchmarked this dataset suite using **four representative modeling approaches** commonly employed in prior studies [133, 131, 3]: (i) traditional machine learning using handcrafted features, (ii) deep learning applied to handcrafted features, (iii) deep learning directly on segments of raw physiological signals, and (iv) pre-trained representation learning methods that leverage signal embeddings learned from external tasks or domains, and presented **comprehensive performance comparisons** to highlight the challenges and opportunities posed by heterogeneous data. In addition to performance evaluation, we also present a comprehensive **cross-dataset analysis** to examine key dimensions of dataset heterogeneity that impact model generalization, with the goal of addressing fundamental questions about which modeling paradigms are effective under varying conditions. Specifically, we examined three harmonization dimensions: Experimental Setting, Device Type, and Labeling Method. In addition, we conducted transferability experiments focusing on participants' demographic characteristics. By systematically analyzing these dimensions, we uncovered how design choices across datasets contribute to performance variability in cross-data models.

We present *FEEL*, the first unified cross-dataset evaluation framework for emotion recognition from physiological signals, enabling systematic analysis of model generalizability and transferability across diverse data-collection scenarios. By moving beyond isolated dataset evaluations, FEEL facilitates a holistic assessment of model performance under varying experimental conditions. This work aims towards the following contribu-

tions: (1) a comprehensive benchmark of 19 publicly available emotion recognition datasets based on physiological signals; (2) a unified binning strategy for data harmonization. (3) a novel fine-tuning strategy for contrastive language-signal pretraining (CLSP) applied to datasets lacking textual modalities; and (4) extensive cross-dataset analyses to evaluate model transferability across variations in labeling strategies, devices, and settings, as well as transferability across demographic groups. Together, FEEL lays the groundwork for developing scalable, robust emotion recognition models for real-world affective computing applications. Code implementation is available here. More information about FEEL can be found on our website.

3.1 Methods

3.1.1 Data Curation

To enable a comprehensive evaluation of heterogeneity in emotion recognition, we curated a collection of 19 datasets comprising PPG and EDA. All datasets are either publicly accessible through research repositories or available upon request from the authors. These datasets collectively represent diverse demographic populations, recording environments, and experiment protocols, providing a basis for evaluating heterogeneity and its influences. PPG and EDA signals were specifically chosen due to their non-invasive nature, widespread implementation in commercial wearable devices, and demonstrated utility for detecting emotional states in ecological settings relevant to real-world applications. The detailed list of our selected datasets, participant counts, devices used, experimental settings, task descriptions, and labeling methods is provided in Table 3.1 and Appendix A.1.

3.1.2 Data Preprocessing and Standardization

To ensure consistency across the heterogeneous formats of the 19 datasets, we developed a unified preprocessing pipeline. For each dataset, we generated standardized CSV files containing (i) extracted features along with participant ID (PID), arousal, and valence labels,

Dataset	#Subjects	Devices	Settings	Task Descriptions	Labeling
WESAD	15	E4	Lab	Neutral Reading, Funny Video Clips, Trier Social Stress Test (TSST), Meditation	Stimulus-Label
NURSE	15	E4	Real	Stress in a Work Environment (Hospital)	Self-report
EMOGNITION	43	E4	Lab	Short Film Clips	Stimulus-Label
UBFC_PHYS	56	E4	Lab	Speech Task - Interview/Holiday description, Arithmetic Task - Countdown	Stimulus-Label
VERBIO	49	E4	Lab	Public speaking anxiety in real and virtual environments	Self-report
PhyMER	30	E4	Lab	Video Stimuli	Self-report
EmoWear	48	E4	Lab	Video Stimuli	Self-report
MAUS	22	Procomp Infinit	Lab	N-Back Task	Stimulus-Label
CLAS	62	Shimmer3 GSR+	Lab	Video Stimuli, Math Problems, Logic Problems, and Stroop Test	Stimulus-Label
CASE	30	ThoughtTech SA9309M, SA9308M	Lab	Video Clips	Self-report
Unobtrusive	24	E4	Lab+Real	Lab: Mental Arithmetic, Sudoku, N-back, Stroop, Eye-Closing, Relaxation; Real Life: Work from Home	Stimulus-Label
CEAP-360VR	32	E4	Lab	VR Video Clips	Self-report
ScientISST MOVE	15	E4	Constraint	Lift a Chair, Greetings, Gesticulate, Jumps, Walk, Run	Stimulus-Label
LAUREATE	44	E4	Real	13-Week Study in University Settings	Self-report
ForDigitStress	38	IOM biofeed-back	Constraint	Digital Job Interviews	Expert-Annotation
Dapper	88	Custom Wristband	Real	Emotional Experiences in Daily Life Over Five Days	Self-reports
ADARP	11	E4	Real	Daily Diary Study (4 Times/14 Days) – Individuals with Alcohol Use Disorders	Self-report
MOCAS	21	E4	Lab	CCTV Monitoring Task Scenario	Self-report
Exercise	36, 31, 30	E4	Constraint	Stroop, Trier Mental Challenge, Debate, Counting, Anaerobic/Aerobic Exercise, Rest	Stimulus-Label

Table 3.1: Overview of our 19 Emotion Datasets: Participant Count, Devices Used, Experimental Settings, Task Descriptions, and Labeling Methods. More information added in the appendix A.1.

and (ii) raw physiological signal data with corresponding metadata. Separate files were created for EDA, PPG, and their combined modalities. We first segmented the data on a per-participant, task-wise basis for the dataset collected in lab or constraint settings with fixed stimuli or tasks. In datasets with no task-specific segments (real-world datasets), we subdivided the signals into hourly segments [133] as per self-reports. Each segment was labeled using a **unified binary scheme** where all data was mapped to arousal and valence dimensions [84]. For datasets where self-reported arousal-valence labels were available, we used them directly; otherwise, we inferred arousal and valence levels based on stimulus type, task metadata, or other self-report data available. This approach harmonized the labeling schemes across datasets and enabled consistent categorization of our data into binary categories. Additional dataset-specific information and binning procedures are documented in the appendix A.1. Following binning, the signal segments were preprocessed to remove artifacts. We then extracted features (see section 3.1.3 separately for EDA and PPG, followed by their concatenation to form a combined feature set. To account for inter-individual variability and prevent dominance of any single feature due to scale differences, participant-wise min-max normalization was applied to the extracted features [3]. Feature selection was guided by prior work [133, 3]. Additionally, on raw signal segments, we applied z-score normalization on a per-participant basis to standardize signal distributions, ensuring comparability across datasets while preserving temporal structure and within-participant variability [132]. To visualize the data-wise distribution of our features after unification, see figure 3.1. Additionally, to enable more nuanced emotion classification experiments, we performed a four-class binning based on the widely accepted circumplex model of affect [134]. Each segment was further labeled into the following four classes: High Arousal Positive Valence (HAPV), High Arousal Negative Valence (HANV), Low Arousal Positive Valence (LAPV), and Low Arousal Negative Valence (LANV), based on the available arousal and valence labels. This four-class approach was chosen for the following reasons: **i) Theoretical grounding and granularity** – it is widely used in the

emotion recognition and affective computing community and captures more fine-grained emotional distinctions than binary labels. **ii) Practical feasibility** - arousal and valence annotations are the most commonly available labels across our benchmark datasets, enabling consistent application of this scheme. **iii) Alignment with cross-dataset harmonization –** it maintains our experimental setting and label unification strategy while better reflecting the complexity of human emotions.

3.1.3 Feature Extraction

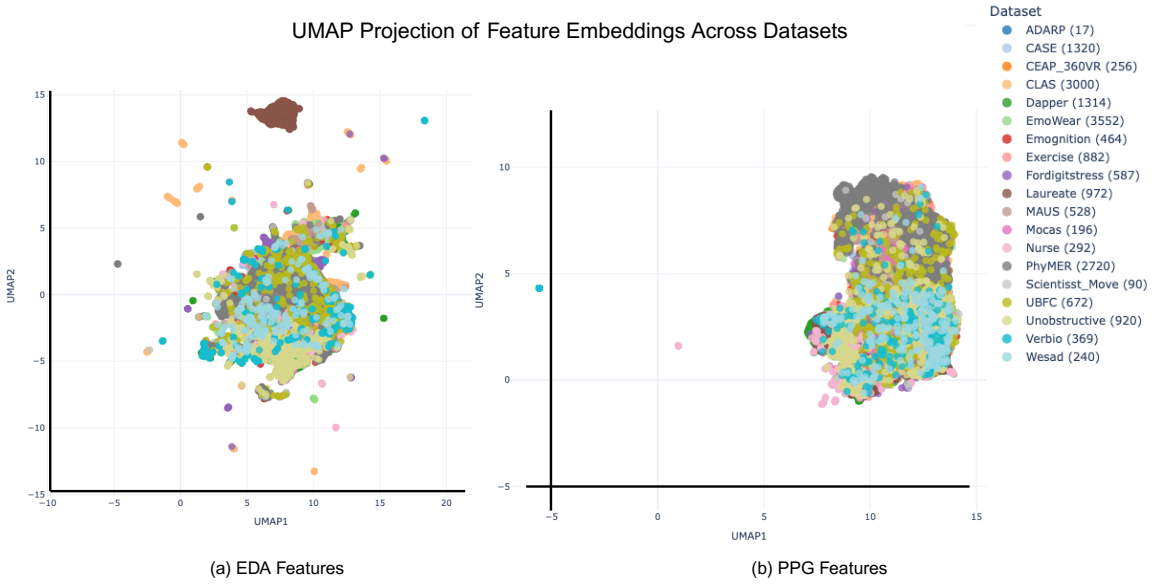


Figure 3.1: UMAP projection of Electrodermal Activity (EDA) and Photoplethysmography (PPG) signals across 19 physiological datasets. Each point represents EDA and PPG features from the dataset, color-coded by dataset source.

EDA: To extract meaningful statistical and physiological features from the EDA signal, we first decompose the raw signal using NeuroKit2's eda_phasic function, which separates it into tonic (slow-changing baseline) and phasic (fast, event-related) components. The phasic signal captures skin conductance responses (SCRs), while the tonic signal represents the baseline skin conductance level (SCL). Following this, the basic statistical features are computed from the raw EDA signal.

PPG: To extract physiological features from the PPG signal, we first iterate over each windowed segment in the raw_ppg_window dataset. Each segment signal is then cleaned using NeuroKit2's ppg_clean function to remove noise and artifacts. The cleaned PPG signal is processed using ppg_process, which extracts key features such as heart rate and waveform characteristics. Then, ppg_analyze is applied to compute PPG-specific features. Table 3.2 outlines the handcrafted statistical features chosen for training our models.

Signal	**Features Selected**
EDA	ku_eda, sk_eda, dynrange, slope, variance, entropy, insc, first_derivative_mean, max_scr, min_scr, nSCR, meanAmpSCR, meanRespSCR, sumAmpSCR, sumRespSCR
PPG	BPM, PPG_Rate_Mean, HRV_MedianNN, HRV_Prc20NN, HRV_MinNN, HRV_HTI, HRV_TINN, HRV_LF, HRV_VHF, HRV_LFn, HRV_HFn, HRV_LnHF, HRV_SD1SD2, HRV_CVI, HRV_PSS, HRV_PAS, HRV_PI, HRV_C1d, HRV_C1a, HRV_DFA_alpha1, HRV_MFDFA_alpha1_Width, HRV_MFDFA_alpha1_Peak, HRV_MFDFA_alpha1_Mean, HRV_MFDFA_alpha1_Max, HRV_MFDFA_alpha1_Delta, HRV_MFDFA_alpha1_Asymmetry, HRV_ApEn, HRV_ShanEn, HRV_FuzzyEn, HRV_MSEn, HRV_CMSEn, HRV_RCMSEn, HRV_CD, HRV_HFD, HRV_KFD, HRV_LZC
EDA+PPG	ku_eda, sk_eda, dynrange, slope, variance, entropy, insc, first_derivative_mean, max_scr, min_scr, nSCR, meanAmpSCR, meanRespSCR, sumAmpSCR, sumRespSCR, BPM, PPG_Rate_Mean, HRV_MedianNN, HRV_Prc20NN, HRV_MinNN, HRV_HTI, HRV_TINN, HRV_LF, HRV_VHF, HRV_LFn, HRV_HFn, HRV_LnHF, HRV_SD1SD2, HRV_CVI, HRV_PSS, HRV_PAS, HRV_PI, HRV_C1d, HRV_C1a, HRV_DFA_alpha1, HRV_MFDFA_alpha1_Width, HRV_MFDFA_alpha1_Peak, HRV_MFDFA_alpha1_Mean, HRV_MFDFA_alpha1_Max, HRV_MFDFA_alpha1_Delta, HRV_MFDFA_alpha1_Asymmetry, HRV_ApEn, HRV_ShanEn, HRV_FuzzyEn, HRV_MSEn, HRV_CMSEn, HRV_RCMSEn, HRV_CD, HRV_HFD, HRV_KFD, HRV_LZC

Table 3.2: Handcrafted Features Selected for EDA, PPG, and Combined (EDA+PPG) Signals

3.1.4 Datasets Benchmarking

We benchmarked our 19 physiological emotion datasets (Table 3.1 across four representative modeling paradigms to evaluate the performance of different signal modalities.

1) Traditional Machine Learning (ML): In this paradigm, we used our extracted handcrafted statistical features $\mathbf{f}_{\mathrm{HC}}$ directly to train classical ML classifiers - RF and LDA.

This paradigm was chosen because it serves as a strong baseline and remains prevalent in prior literature [37, 39, 135], especially for small-sized datasets. Implementation details are provided in the appendix A.3.1.

2) Deep Learning with Handcrafted Features: In this paradigm we used our extracted handcrafted features $\mathbf{f}_{\text{HC}}$ as input to deep learning architectures - MLP ($\mathbf{f}_{HC}$+*MLP*), ResNet ($\mathbf{f}_{HC}$+*ResNet*, LSTM ($\mathbf{f}_{HC}$+*LSTM+MLP*), and Attention-based model ($\mathbf{f}_{HC}$+*Attention+MLP*). This paradigm was chosen because it combines domain knowledge embedded in engineered features with non-linear learning capabilities, representing practical scenarios where interpretability and complex pattern recognition are both required [136, 137, 138, 139]. Implementation details are provided in appendix A.3.2.

3) Deep Learning on Raw Signals: In this paradigm we used raw time-series signals x(t) directly as input to deep learning architectures - *ResNet, LSTM+MLP*, and *CNN + Transformer Encoder Block*. This paradigm was chosen because it enables end-to-end learning without manual feature extraction, allowing models to autonomously discover novel representations and potentially capture subtle signal characteristics overlooked in traditional feature engineering [140]. Implementation details are provided in Appendix A.3.3.

4) Pretrained Representation Learning: In this paradigm, we evaluated the zero-shot and fine-tuned performance of each of our datasets using models based on *Contrastive Language-Signal Pretraining* (CLSP) [3]. This paradigm was chosen to evaluate how each of our datasets performs when utilizing pre-trained models with or without dataset-specific training. We selected CLSP models because, to the best of our knowledge, they are the only available pre-trained models specifically developed for physiological emotion recognition, trained on the EEVR dataset (described in detail in Chapter 5), and incorporating both PPG and EDA modalities. Moreover, CLSP models have been shown to exhibit strong cross-dataset generalization, which motivated our selection. We first performed Zero-shot inference without any dataset-specific adaptation to assess the direct transferability of

pretrained representations. Then we performed Dataset-specific fine-tuning to examine how well these representations can be adapted to each dataset's characteristics and influence their classification performance. For fine-tuning, we split the dataset into participant-wise 50-50 train and test sets, and then we employed three progressively increasing data efficiency regimes: **few-shot (5%), low-resource (25%)**, and **partial-participant (50%)** of samples per class from the training set. This systematic approach enables us to comprehensively benchmark not only the baseline zero-shot performance but also the adaptation potential of pre-trained physiological representations across our diverse dataset collection.

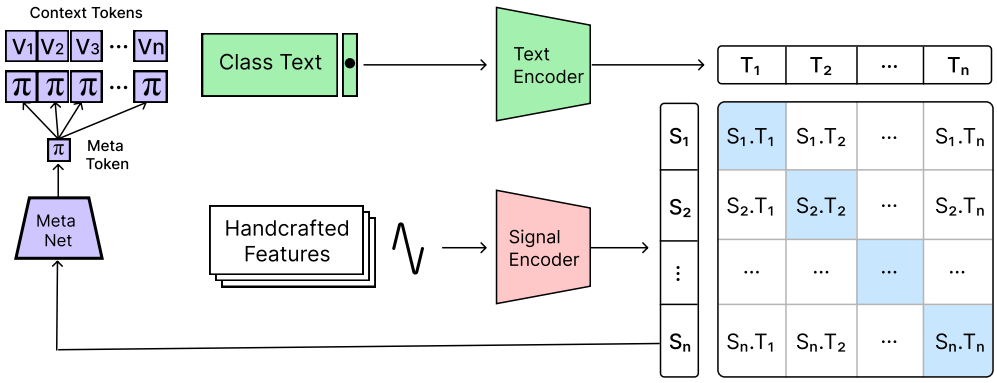


Figure 3.2: Our CLSP Fine-Tuning Approach, consists of lightweight neural network (Meta-Net) that generates for each signal segment an input-conditional context token.

For fine-tuning, we adopted *Conditional Context Optimization* (CoCoOp) [141] (figure 3.2 with two CoCoOp MetaNet-inspired modulation networks to condition textual prompts based on physiological input signals. This approach was chosen because it enables adaptation of pretrained representations without requiring ground-truth text annotations in the target datasets. The modulation networks were designed in two architectural variants: one using **two linear layers**, and another employing **two stacked 1D convolutional layers** applied to the signal embedding sequence. In our implementation, we expanded beyond basic *class text* labels ("high arousal," "low arousal," "positive valence," or "negative valence") to incorporate more nuanced emotion-specific textual descriptions (detailed in appendix A.3.4 processed through the CLSP text-encoder. These enriched descriptions, along with context

tokens, function as context-adaptive textual prompts tailored to each physiological input, facilitating more precise mapping between signal patterns and emotional states. Complete implementation specifications are in Appendix A.3.4.

3.1.5 Cross-Dataset Generalization Analysis

We conducted a systematic cross-dataset analysis to quantify the impacts of contextual heterogeneity. We grouped our datasets (see A.4.1 based on three dimensions chosen for their potential to introduce systematic variability in data and modeling outcomes. The dimensions include: **1) Experimental Setting:** to capture the influence of differences in ecological validity on model performance. Our dataset collection included three settings: the lab, a semi-realistic constraint setting, and a real-life setting. **2) Device Type:** to account for variation in sensor hardware. Our collection included wrist-worn research-grade wearables (Empatica E4), custom-designed wearables, and lab-based devices. **3) Labeling Method:** to reflect the influence of emotion labeling technique. Our collection includes three annotation strategies: stimulus-based labels (assigned based on predefined stimulus-response assumptions), self-reported labels (reflecting participants' subjective emotional states), and expert annotations (provided by trained observers based on behavioral cues and self-reports). Together, these harmonization dimensions offer a structured approach to assessing the contextual factors. To examine the influence of demographic factors on model transferability, we conducted supplementary experiments across two demographic attributes. For gender-based analysis, we utilized binary male/female labels available across 8 datasets. For age-based analysis, we partitioned subjects into two groups: Young (18–25 years) and Old (25+ years) across 7 datasets, based on data availability. More details are provided in Section A.4.1.

3.1.6 Meta Analysis

In this section, we present our meta-analysis, which included computing the numbers of high- and low-arousal and positive- and negative-valence samples in each dataset to assess class balance and artifact percentages. This was done as part of our qualitative analysis of benchmarking performance to determine whether these factors could have impacted overall performance. We begin by defining artifacts per prior literature, discussing their impact on data quality and performance, and then explain our artifact detection methods. Finally, we presented our comprehensive dataset-wise analysis in Table 3.3.

Artifact Detection:

Artifacts in biosignals (EDA, PPG) are "non-avoidable distortions which get superimposed on the signals representing emotional changes that originate from external (e.g., movement, ambient light) or internal (e.g., electrode-skin impedance changes) sources" [142]. Artifacts in EDA are "transient, non-sweat-gland–related perturbations of the skin conductance signal that do not reflect sympathetic nervous system activity. Common sources of these perturbations include abrupt motions, unstable electrode contact, and environmental factors such as humidity or temperature fluctuations [143]. Similarly, PPG artifacts are distortions in the optical pulse waveform that are not caused by pulsatile blood-volume changes, thereby degrading the actual cardiac-related component. Limb motion and sensor contact pressure are significant sources of these distortions [142]."

Impact of Artifacts:

Artifacts in EDA signals impact the quality of emotion-representation in the data by inflating noise relative to true sympathetic responses. In-the-wild studies show that including artifact-contaminated segments can reduce arousal–valence classification accuracy by up to 20 percentage points compared to manually cleaned data, directly undermining the reliability of affective state estimates. Moreover, unfiltered motion and contact artifacts bias tonic–phasic

decomposition methods, misestimating key features such as mean SCR amplitude, thereby impacting machine-learning emotion classifiers that rely on these features (e.g., [144]. In PPG signals, motion artifacts disrupt the detection of valid inter-beat intervals, leading to heart-rate variability (HRV) metrics with mean absolute errors exceeding 30 ms and affecting arousal prediction models (e.g., [145].

Artifact Detection Methods

EDA Artifacts: For our artifact detection, we have used the EDA artifact detection pipeline proposed in EDArtifact [143], which involves a structured sequence of preprocessing, feature extraction, and classification. The raw EDA signals are initially sampled at 4 Hz and segmented into non-overlapping windows of 60 samples (equivalent to 15 seconds). Each segment undergoes Haar wavelet decomposition up to level 3 to capture multi-resolution signal characteristics. A comprehensive set of 36 features is extracted from each segment, encompassing statistical properties (e.g., mean, variance), first- and second-order derivative statistics, wavelet coefficients, and characteristics of skin conductance response (SCR) peaks. The SCR peaks are identified using a minimum amplitude threshold of 0.01 μS, with an onset validation offset of 1 sample, a pre-apex search window of 3 seconds, and a post-apex half-amplitude decay window of 10 seconds. The extracted features are then normalised and input into a pre-trained XGBoost classifier, which has been trained to distinguish between clean and artifact-contaminated segments.

PPG Artifacts: We used the Tiny-PPG motion artifact detection pipeline for PPG signals, leveraging a lightweight pre-trained 1d convolutional neural network designed for real-time deployment on edge devices [145]. Raw PPG time series data is loaded and, for each subject (PID), segmented into non-overlapping windows of 60 samples. Each window is reshaped to the model's expected input tensor shape [batch, channel, length] = [1, 1, 60]. Inference is performed window-wise, with each segment passed through the model to obtain a segmentation mask indicating the likelihood of motion artifact. A sigmoid activation is

applied to convert logits into probabilities, followed by binarisation using a threshold of 0.5. The prediction for the entire signal is aggregated using a mean-based criterion. If the average probability of artifact presence exceeds 0.5, the segment is labelled as containing a motion artifact.

Dataset Name	**High Arousal**	**Low Arousal**	**Positive Valence**	**Negative Valence**	**EDA Artifacts (%)**	**PPG Artifacts (%)**
WESAD	**120**	**120**	**120**	**120**	11.34 ± 21.22	**91.84 ± 7.56**
NURSE	130	162	248	44	18.76 ± 14.99	**100 ± 0**
EMOGNITION	252	212	295	169	33.75 ± 36.92	**94.09 ± 5.08**
UBFC_PHYS	208	464	**328**	**344**	8.3 ± 19.18	0 ± 0
VERBIO	254	115	288	81	24.7 ± 33.37	53.45 ± 10.35
PhyMER	1036	1684	1128	1592	2.89 ± 10.09	0 ± 0
EmoWear	1950	1602	1216	2336	7.70 ± 14.03	0.03 ± 1.68
MAUS	352	176	352	176	0.64 ± 1.52	**100 ± 0**
CLAS	**1560**	**1440**	**1560**	**1440**	70.82 ± 27.5	0 ± 0
CASE	332	988	816	504	**98.67 ± 1.9**	**100 ± 0**
Unobtrusive	707	213	359	561	12.6 ± 4.5	3.80 ± 19.13
CEAP-360VR	98	158	**135**	**121**	13.41 ± 9.67	6.67 ± 3.5
ScientISST MOVE	30	60	31	59	59.81 ± 10.9	**96.67 ± 17.95**
LAUREATE	718	254	703	269	18.92 ± 18.42	**92.01 ± 4.67**
ForDigitStress	505	82	154	433	0 ± 0	11.83 ± 10.02
Dapper	2244	757	307	2694	7.28 ± 2.94	37.6 ± 2.48
ADARP	13	4	14	3	68.7 ± 14.47	**100 ± 0**
MOCAS	**98**	**98**	36	160	33.67 ± 10.3	31.9 ± 14.03
Exercise	**427**	**455**	347	535	50.55 ± 7.47	**95.08 ± 1.17**

Table 3.3: **Data quality statistics** across datasets. Bold values in the Arousal and Valence columns indicate **near-balanced class distributions** (min/max ≥ 0.9). Bold values in the artifact columns denote cases where over **90% of the data** across participants is affected by artifacts.

3.2 Experiments

We conducted separate experiments for arousal, valence, and four-class classification using three input configurations: EDA only, PPG only, and EDA+PPG. For four-class classification, CLSP fine-tuning benchmarking could not be performed on several datasets: NURSE, UBFC_PHYS, MAUS, Unobtrusive, VERBIO, and ADARP lacked samples from one or more classes, while MOCAS and ScientISST MOVE had insufficient samples for 5% fine-

tuning. Additionally, we did not benchmark DL-based models for four-class classification (except HC+MLP) due to their poor performance in binary classification. Overall, this set of experiments enabled a systematic comparison of the standalone utility of each modality and its combined contribution across various modeling paradigms.

ML and DL Paradigm: We evaluated the performance of our ML and DL models using Leave-One-Subject-Out Cross-Validation (LOSO-CV), to ensure subject-independent evaluation and reflect real-world deployment conditions. Performance was measured using average accuracy and F1 scores, standard metrics in physiological emotion recognition [37, 3] computed across all LOSO folds. For raw signal-based DL models, we used a sliding window approach with a window size of 60 samples and 50% overlap between consecutive windows. To mitigate class imbalance in both arousal and valence classification, we applied random oversampling to all datasets and SMOTE (Synthetic Minority Oversampling Technique) to datasets with significant imbalance (defined as a class size difference exceeding one-third). We did not apply oversampling during CLSP fine-tuning, except in cases of significant imbalance. Additional architectural and training details are provided in Appendix A.3.

Pretraining Paradigm: We evaluated pretrained models under subject-independent conditions using three data efficiency regimes: 5%, 25%, and 50% of training data per class. Each dataset was split participant-wise into 50% training and 50% testing sets. To ensure robustness, we repeated the experiment with the train and test splits swapped and computed the average accuracy and F1 score across both folds. See additional details in Appendix A.3.4.

Benchmarking Analysis: To evaluate performance variations across datasets, we conducted a comprehensive meta-analysis, examining the impact of data quality factors (see 3.1.6, sensor modality (EDA, PPG, EDA+PPG), and task (Arousal and Valence) on model effectiveness. We further ranked overall dataset performances and performed a qualitative analysis to interpret the observed trends. This involved assessing model behavior across

different modalities and tasks, and attributing performance differences to underlying factors in the data collection pipeline, specifically, the recording environment, labeling methodology, elicitation task, and sensing devices used.

Cross-Dataset Analysis: To evaluate generalization performance, we first identified the top three performing models for each dataset. Through majority voting across these top models, we observed that classical machine learning approaches (LDA, RF), the hybrid handcrafted feature-based HC+MLP architecture, and CLSP models with MLP and CNN meta-learners consistently outperformed more complex signal-based deep learning models. Based on this analysis, we selected LDA, RF, HC+MLP, and all CLSP variants for cross-dataset evaluation. To specifically assess cross-domain transferability, each selected model was retrained on its corresponding training partition and evaluated on non-overlapping dataset groups. These results were then compared against two key baselines: **1)** the leave-one-dataset-out (LODO) performance, representing in-domain transferability within the training cohort, and **2)** the CLSP zero-shot performance, which served as a pretrained, no-adaptation baseline for out-of-domain generalization. Where the LODO evaluation used RF and HC+MLP models to examine within-cohort generalization, i.e., how well models trained on all but one dataset performed when applied to the held-out dataset from the same domain. This comparison enabled a comprehensive analysis of the generalization capabilities of both traditional and pre-trained models when applied to unseen datasets (see more details in Appendix A.4.

Computation: For our benchmarking experiment, the computational cost varied by dataset complexity and size. The five largest datasets, including CLAS, CASE, Unobtrusive, DAPPER, and LAUREATE, required approximately 600–720 GPU hours (5–6 days) each. While the remaining 14 datasets required 24–30 hours each, totaling approximately 336–420 GPU hours. Raw signal-based deep learning models were the most computationally intensive, whereas traditional ML models were significantly faster. Fine-tuning CLSP models was highly efficient, requiring only 1–30 minutes of GPU training time, depending on dataset

size. All experiments were conducted on 4 NVIDIA A100, 2 H100, and 2 H200 GPUs; more information about the compute resources is provided in Appendix A.2.

DataSet	EDA		PPG		EDA+PPG	
	Best Model	**F1**	**Best Model**	**F1**	**Best Model**	**F1**
WESAD	Signal + Resnet	0.83	HC+MLP	0.8	HC+MLP	**0.91**
NURSE	CLSP+CNN (5%)	**0.62**	CLSP+MLP (50%)	0.52	CLSP+CNN (5%)	**0.62**
EMOGNITION	CLSP+MLP (5%)	**0.68**	CLSP+MLP (50%)	0.62	CLSP+CNN (5%)	0.57
UBFC_PHYS	CLSP+MLP (5%)	**0.45**	CLSP+CNN (25%)	0.34	CLSP+MLP (5%)	0.41
PhyMER	CLSP - Zero Shot	**0.51**	LDA	0.42	LDA	0.42
EmoWear	RF	0.64	RF	0.64	CLSP+MLP (50%)	**0.67**
MAUS	Signal + Resnet	**0.83**	RF	0.82	RF	0.82
CLAS	RF	0.69	RF	0.66	RF	**0.70**
CASE	Signal+CNN+Transformer	**0.47**	HC+MLP	0.30	CLSP+MLP (5%)*	0.40
Unobtrusive	RF	**0.88**	RF	0.87	CLSP+MLP (50%)	0.86
CEAP-360VR	CLSP+CNN (5%)	**0.56**	CLSP+CNN (5%)	0.43	CLSP+MLP (5%)	0.45
ScientISST MOVE	HC+Attention+MLP	0.77	RF †	0.81	HC+MLP	**0.88**
Dapper	CLSP+CNN (50%)	0.77	CLSP+CNN (5%)	0.70	CLSP+MLP (5%)	**0.81**
ForDigitStress	CLSP+MLP (25%)	0.94	RF †	**0.99**	RF †	**0.99**
ADARP	CLSP+MLP (25%)	**0.83**	CLSP+CNN (50%)	0.80	CLSP+MLP (25%)	0.62
Exercise	CLSP - Zero Shot	**0.63**	CLSP+CNN (25%)	0.57	CLSP+MLP (5%)	0.54
MOCAS	CLSP+CNN (5%)	**0.65**	CLSP+MLP (5%)	0.62	CLSP+MLP (25%)	0.63
LAUREATE	RF†	0.69	CLSP+MLP (50%)	0.77	CLSP Zero-Shot	**0.82**
VERBIO	CLSP+CNN (50%)	**0.83**	CLSP+CNN (50%)	0.77	CLSP+CNN (50%)	0.72

Table 3.4: Best-performing model and corresponding F1 score for **arousal classification** across all datasets and modalities (EDA, PPG, EDA+PPG). The table lists, for each dataset and modality, the model that achieved the highest F1 score. * : reflects results that are achieved after applying random sampling before CLSP fine-tuning. †: reflects results achieved after applying SMOTE.

3.3 Results

We summarize the main results in this section, with additional details provided in Appendix A.5. We begin by outlining the performance trends observed across the individual datasets, followed by a detailed explanation of benchmarking performance and cross-dataset evaluation.

3.3.1 Benchmarking Performance across Modeling Paradigms

Overall Comparision: The overall benchmarking results are summarized in Tables 3.4, 3.5, and 3.6. On average, pretrained models, particularly various CLSP variants, consistently outperformed other approaches across datasets for binary classification, contributing to

DataSet	EDA		PPG		EDA+PPG	
	Best Model	**F1**	**Best Model**	**F1**	**Best Model**	**F1**
WESAD	CLSP+CNN (50%)	0.83	CLSP+CNN (50%)	0.83	HC+MLP	**0.98**
NURSE	CLSP - Zero Shot	**0.62**	CLSP+CNN (5%) †	0.39	CLSP Zero-Shot	0.38
EMOGNITION	CLSP - Zero Shot	**0.53**	CLSP+MLP (5%)	0.50	CLSP+CNN (5%)	0.39
UBFC_PHYS	RF	**0.76**	LDA	0.68	RF	0.72
PhyMER	CLSP - Zero Shot	**0.72**	CLSP+CNN (50%)	0.69	CLSP+MLP (50%)	0.70
EmoWear	CLSP+CNN (50%)	**0.78**	CLSP+CNN (50%)	0.77	RF	0.77
MAUS	HC+MLP	0.58	LDA	0.56	LDA	**0.59**
CLAS	CLSP - Zero Shot	**0.64**	CLSP+CNN (25%)	0.61	HC+Attention+MLP	0.63
CASE	CLSP+MLP (5%)	**0.54**	LDA	0.48	LDA	0.49
Unobtrusive	CLSP - Zero Shot	**0.71**	RF	**0.71**	CLSP+CNN (25%)	0.70
CEAP-360VR	CLSP+CNN (5%)	**0.62**	CLSP+CNN (5%)	0.61	LDA	0.50
ScientISST MOVE	CLSP+MLP (50%)	**0.82**	CLSP+CNN (50%)	0.80	CLSP+CNN (50%)	**0.82**
Dapper	CLSP+CNN (50%)	0.87	CLSP+CNN (50%)	0.85	CLSP+CNN (50%)	**0.94**
ForDigitStress	CLSP+CNN (5%)	0.87	RF†	**0.92**	RF†	**0.92**
ADARP	CLSP - Zero Shot	0.30	CLSP Zero-Shot	0.40	HC+MLP	**0.47**
Exercise	CLSP - Zero Shot	**0.75**	CLSP+CNN (50%)	0.72	CLSP+MLP (50%)	0.71
MOCAS	CLSP - Zero Shot	**0.89**	CLSP+CNN (50%)	0.87	CLSP+CNN (25%)	0.82
LAUREATE	HC+MLP	0.36	HC+MLP	**0.41**	CLSP+MLP (50%)*	0.40
VERBIO	HC+MLP	**0.40**	HC+MLP	0.38	CLSP+MLP (5%)	0.34

Table 3.5: Best-performing model and corresponding F1 score for **valence classification** across all datasets and modalities (EDA, PPG, EDA+PPG). The table lists, for each dataset and modality, the model that achieved the highest F1 score. * : reflects results that are achieved after applying random sampling before CLSP fine-tuning. †: reflects results achieved after applying SMOTE.

Dataset	EDA		PPG		EDA+PPG	
	Best Model	**F1**	**Best Model**	**F1**	**Best Model**	**F1**
WESAD	RF	**0.987**	RF	0.794	LDA	**0.987**
NURSE	CLSP Zero Shot	0.433	CLSP Zero Shot	**0.667**	CLSP Zero Shot	0.52
EMOGNITION	RF	0.572	CLSP+CNN (50%)	**0.601**	RF	0.513
UBFC_PHYS	CLSP Zero Shot	**0.705**	LDA	0.551	LDA	0.622
PhyMER	CLSP+CNN (50%)	**0.723**	RF	0.3	RF	0.342
EmoWear	CLSP+CNN (50%)	**0.293**	HC+MLP	0.27	HC+MLP	0.282
MAUS	HC+MLP	0.7	RF	0.705	RF	**0.728**
CLAS	RF	0.43	HC+MLP	0.408	RF	**0.459**
CASE	RF	0.476	RF	0.397	RF	**0.498**
Unobtrusive	RF	0.402	CLSP Zero Shot	**0.409**	HC+MLP	0.393
CEAP-360VR	CLSP+MLP (25%)	0.285	RF	0.307	RF	**0.314**
ScientISST MOVE	CLSP+MLP (25%)	0.701	CLSP+CNN (50%)	0.74	CLSP+CNN (50%)	**0.8**
Dapper	RF	0.434	RF	0.426	RF	**0.555**
ForDigitStress	LDA	0.682	RF	0.821	RF	**0.826**
ADARP	CLSP Zero Shot	0.269	CLSP Zero Shot	**0.433**	CLSP Zero Shot	0.354
Exercise	CLSP+CNN (25%)	**0.552**	HC+MLP	0.438	RF	0.48
MOCAS	RF	**0.412**	RF	0.357	RF	0.366
LAUREATE	CLSP+MLP (5%)	**0.527**	RF	0.46	RF	0.461
VERBIO	CLSP Zero Shot	0.48	CLSP Zero Shot	**0.582**	CLSP Zero Shot	0.436

Table 3.6: Best-performing model and corresponding F1 score for **four class classification** across all datasets and modalities (EDA, PPG, EDA+PPG). The table lists, for each dataset and modality, the model that achieved the highest F1 score.

Dataset	Data Collection Context	Observed Performance and Interpretation
ForDigitStress	Laboratory-based, semi-controlled digital interview tasks with expert annotations.	Among the strongest performers for arousal and consistently high for valence, likely due to realistic stress elicitation in constraint setting, combined with precise expert labeling.
WESAD	Controlled laboratory study incorporating multiple elicitation modalities (videos, stress induction, meditation).	High performance for both arousal and valence, particularly when combining EDA and PPG, reflecting a well-thought-out experiment design and high signal quality.
MOCAS	Surveillance-style monitoring task with self-reported SAM ratings.	Strong valence classification but weaker arousal performance, possibly due to relatively mild or less immersive emotional stimuli, which doesn't impact arousal.
ScientISST MOVE	Physically engaging activities (e.g., handshake, jumping) with self-reported emotion labels.	Very strong arousal classification driven by physical activation, with reasonably good valence performance despite motion-related signal noise, suggesting advantages of a controlled setting.
Unobtrusive	Office-like cognitive tasks conducted in both lab and naturalistic settings with Likert-scale emotion reports.	High performance in arousal classification, driven by realistic work scenarios; valence classification was moderate, possibly due to the complexity of labeling subtle emotional changes.
Dapper	Real-world experience sampling study using custom wearable sensors.	Strong valence classification, as ESM possibly allowed timely and accurate self-reporting of naturally occurring emotions.
MAUS	Controlled cognitive workload paradigm (N-Back task).	Strong arousal classification, as task reliably induced measurable physiological changes linked to cognitive load.

Table 3.7: Qualitative overview of high-performing datasets from our individual benchmarking, illustrating how elicitation context, annotation approach, and task design shape performance. See Table 3.8 for stats.

71 of the 114 top-performing model instances for binary classification. Among classical machine learning techniques, RF and LDA followed, with 17 and 8 top-performing entries, respectively. Within the deep learning category, the handcrafted feature-based MLP achieved 11 top results, while signal-based deep models accounted for 3, and handcrafted features combined with attention mechanisms contributed 2 best-performing instances. Within the CLSP model family, we observed that fine-tuning played a crucial role in achieving strong cross-dataset performance. In 29 out of 73 top-performing instances, models required fine-tuning on up to 50% of the target dataset, indicating that while CLSP models offer transferability, moderate domain adaptation is often necessary. Notably, 21 instances achieved competitive results with only 5% of the data used for fine-tuning, suggesting that CLSP models can exhibit effective few-shot generalization. A smaller subset (9 instances) performed best with 25% fine-tuning, reinforcing the spectrum of adaptation needs across datasets. Among CLSP variants, CLSP+CNN demonstrated the highest overall performance, contributing to 37 top-performing cases, followed by CLSP+MLP with 22. Zero-shot variants of CLSP, which require no fine-tuning, were top performers in 14 cases, highlighting the generalizability of the CLSP baseline. Collectively, these findings suggest that while zero-shot CLSP offers a useful starting point, performance can be significantly improved through lightweight dataset-specific fine-tuning, particularly with a CNN metanet that better captures transferable patterns in physiological signals. Detailed comparison of all modeling paradigms and their performances across our datasuite for both recognition tasks and all three modality variations are shown in Figures A.3, A.4, A.5, A.6, A.7, and A.8. For four-class classification, machine learning models (RF and LDA) dominate with 54% of best-performing cases. CLSP-based models appear in 35% of instances, while deep learning models (HC+MLP) account for only 11% of the best-performing outcomes (see Table 3.6).

Dataset Specific Performance Variations: Model performance for binary classification exhibited significant variability across datasets, ranging from a minimum F1 score of 0.30 (e.g., in CASE and ADARP) to a maximum of 0.98 (WESAD), as detailed in Table 3.8. Our

	Arousal			Valence		
Statistic	**EDA**	**PPG**	**EDA+PPG**	**EDA**	**PPG**	**EDA+PPG**
MIN	0.33	0.30	0.36	0.30	0.33	0.34
MAX	0.94	0.99	0.99	0.89	0.92	0.98
	Arousal			**Valence**		
Rank	**EDA**	**PPG**	**EDA+PPG**	**EDA**	**PPG**	**EDA+PPG**
1	ForDigitStress	ForDigitStress	ForDigitStress	MOCAS	ForDigitStress	WESAD
2	Unobtrusive	Unobtrusive	WESAD	Dapper	MOCAS	Dapper
3	ADARP	MAUS	ScientISST MOVE	ForDigitStress	Dapper	ForDigitStress
4	MAUS	ADARP	Unobtrusive	WESAD	WESAD	MOCAS
5	VERBIO	WESAD	LAUREATE	ScientISST MOVE	ScientISST MOVE	ScientISST MOVE
6	WESAD	LAUREATE	MAUS	EmoWear	EmoWear	EmoWear
7	Dapper	ScientISST MOVE	Dapper	UBFC_PHYS	Exercise	UBFC_PHYS
8	ScientISST MOVE	VERBIO	VERBIO	Exercise	Unobtrusive	Exercise
9	CLAS	Dapper	CLAS	PhyMER	PhyMER	Unobtrusive
10	EMOGNITION	CLAS	EmoWear	Unobtrusive	UBFC_PHYS	PhyMER
11	MOCAS	EmoWear	MOCAS	CLAS	CEAP-360VR	CLAS
12	EmoWear	MOCAS	ADARP	CEAP-360VR	CLAS	MAUS
13	Exercise	EMOGNITION	NURSE	NURSE	MAUS	CEAP-360VR
14	NURSE	Exercise	EMOGNITION	MAUS	EMOGNITION	CASE
15	CEAP-360VR	NURSE	Exercise	CASE	CASE	ADARP
16	PhyMER	CEAP-360VR	CEAP-360VR	EMOGNITION	LAUREATE	EMOGNITION
17	CASE	PhyMER	PhyMER	VERBIO	ADARP	NURSE
18	UBFC_PHYS	UBFC_PHYS	UBFC_PHYS	LAUREATE	VERBIO	LAUREATE
19	LAUREATE	CASE	CASE	ADARP	NURSE	VERBIO

Table 3.8: F1 score statistics (MIN, MAX, AVG, STD) and rankings of our 19 datasets according to their performance for arousal and valence prediction using EDA, PPG, and their combination (EDA+PPG).

qualitative analysis of dataset collection methodologies, as discussed in Tables 3.7, 3.10, and 3.9, revealed that high-performing datasets generally shared key characteristics: well-balanced experimental setups, ecologically valid elicitation protocols, and robust labeling techniques that accounted for participants' labeling subjectivity [1]. In contrast, suboptimal performance was often linked to weak elicitation strategies, misaligned labels, the absence of stimuli covering the full emotional spectrum, and the presence of signal artifacts (see Figures A.9, A.10, and A.11). Overall, models utilizing EDA consistently outperformed those trained on PPG only or EDA+PPG data, as shown in Figure A.1. EDA-based models achieved top performance on 12 of 19 datasets for arousal classification and 13 of 19 datasets for valence classification, highlighting the robustness of EDA signals in emotion recognition tasks. Notably, EDA+PPG also showed strong performance, particularly in real-life and constrained task settings, where multimodal input helped mitigate signal noise. While models based solely on PPG performed comparably in some cases, they generally yielded lower performance than EDA-based approaches, suggesting that PPG may be less sensitive to subtle

emotional variations. Overall, our results highlight the critical importance of thoughtfully designing data collection protocols to effectively capture meaningful emotional variations and of aligning labeling strategies that accurately reflect these underlying physiological changes across settings to overcome the impact of heterogeneity across datasets.

Dataset	Data Collection Context	Observed Performance and Interpretation
CLAS	Laboratory tasks including logic problems, mathematics, and videos.	Near-random classification performance, likely due to weak elicitation intensity and inconsistent labeling.
CASE	Video-based lab stimuli with continuous joystick self-report annotation.	Poor performance across both dimensions, possibly because of imposed cognitive load on participants due to continuous self-reporting.
LAUREATE	Real-world classroom recordings with engagement-focused self-reports.	Stronger arousal detection using EDA/PPG signals, but weak valence classification due to abstract or indirect labeling.
ADARP	Real-life setting involving participants with alcohol-use disorder using self-reports.	Reasonable arousal classification for individual modalities but poor valence performance, likely reflecting skewed emotional labels and participants' selection bias.

Table 3.9: Qualitative overview of datasets showing lower or inconsistent performance in individual benchmarking. See Table 3.8 for performance statistics.

The four-class classification results varied considerably across datasets, with F1 scores ranging from 0.987 (WESAD) to 0.269 (ADARP). Among the lowest-performing datasets were EmoWear and CLAS, which, despite being lab-based, relied solely on video stimuli, suggesting that limited or low-arousal stimuli can constrain physiological differentiation. ADARP, collected in real-life daily settings, exhibited low performance primarily due to an imbalance in sample sizes across four classes (as shown in Table 3.3). Meanwhile, CEAP-360VR, CASE, and MOCAS, which combined lab and semi-realistic stimuli, also led

to overall poorer performance than other datasets. These patterns suggest that dataset quality, stimulus richness, and ecological validity collectively influence classification outcomes, beyond simple distinctions between laboratory and real-world environments. Overall, for four-class classification, this suggests that modality combination did not yield a consistent strong advantage across the 15 datasets, while EDA+PPG and EDA-only models generally achieved relatively high performance, PPG alone performed comparably in many cases (see Figure A.2).

3.3.2 Performances Across Harmonizing Dimensions

Experiment Setting: Our experiments provide important insights into how training environments shape cross-domain generalization in physiological emotion recognition. Models trained on real-world datasets demonstrated strong transferability to both laboratory and constraint-based settings, achieving F1 scores up to 0.79 (CLSP-MLP at 5%) and 0.78 (RF), respectively. However, their performance within the same real-world domain remained relatively low (maximum F1 0.49), suggesting substantial heterogeneity within datasets collected in naturalistic settings. Models trained on constraint-based datasets showed particularly strong transfer to real-world data, achieving the highest overall performance (F1 = 0.88 with RF), while demonstrating only moderate success within their own domain and in laboratory settings, especially for arousal prediction. This pattern indicates that semi-structured elicitation protocols may provide a useful intermediate balance between ecological validity and experimental control, producing data that generalize well without being overly constrained. Similarly, laboratory-trained models transferred reasonably well to both real-world and constraint-based datasets (maximum F1 0.76) but underperformed within their own datasets (F1 0.5), potentially reflecting cohort-specific differences. Detailed results are added in Table A.5, A.6, and figure A.14. Building on these observations, the overall results suggest that the experimental setting plays a significant role in determining the generalizability of physiological emotion recognition models. Constraint-based datasets appear to

Dataset	Data Collection Context	Observed Performance and Interpretation
NURSE	Real-world recordings from nurses with stress-focused self-reports.	Low performance for both arousal and valence, likely driven by class imbalance and limited representation of positive emotional states.
Emowear	Laboratory audiovisual stimuli with participant self-report annotations.	Generally weak performance, possibly due to mild elicitation intensity and controlled lab conditions limiting ecological validity.
UBFC_PHYS	Speech and arithmetic stress tasks labeled primarily using stimulus-based annotations.	Slightly better valence than arousal, but overall constrained by limited correspondence between stimulus labels and experienced emotions.
VERBIO	Public speaking tasks conducted in both VR and real-world environments.	Arousal classification moderately strong; valence weaker due to relatively narrow emotional variability.
EMOGNITION	Short film clips presented in laboratory settings.	Overall low performance, likely reflecting the poor emotion elicitation potential of short film clips.
CEAP-360VR	Immersive VR video stimuli with self-reported emotion labels.	Slightly improved valence classification but overall limited performance, suggesting insufficient emotional elicitation strength.
Exercise	Laboratory cognitive and physical stress tasks.	Valence captured somewhat better than arousal, potentially reflecting the lack of alignment between physiological changes and labels.
PhyMER	Lab-based video stimuli accompanied by self-reports.	Marginally stronger valence performance, but overall constrained by relatively weak emotional elicitation.

Table 3.10: Qualitative overview of datasets showing lower or inconsistent performance in individual benchmarking. See Table 3.8 for performance statistics.

offer the most consistent cross-domain performance, likely because they balance ecological realism with structured elicitation and reliable annotation. Real-world datasets, while highly valuable for capturing natural emotional variability, introduce substantial heterogeneity that can reduce intra-domain consistency. Laboratory datasets, despite high signal quality and controlled conditions, may inadvertently encode context- or cohort-specific biases that limit broader applicability.

Device: Our analysis also revealed substantial variability in cross-cohort transferability across sensing devices. Models trained on wearable data collected using the Empatica E4 generalized well to the custom wearable cohort (best F1 = 0.82 with CLSP-MLP at 50%), but transferred poorly to laboratory-based datasets (minimum F1 0.45), suggesting limited alignment between wearable and lab-grade sensing conditions. In contrast, models trained on lab-based devices demonstrated strong and more consistent generalization to both custom wearable data (best F1 = 0.81 with LDA) and E4 data (best F1 = 0.73 with CLSP-CNN at 50%), indicating that high-quality lab recordings may provide robust feature representations that transfer across hardware types. Models trained on custom wearable datasets showed generally weaker transferability overall, although slightly better performance was observed for arousal detection, particularly when evaluated on lab-based datasets (best F1 = 0.64). Notably, zero-shot CLSP models using EDA signals showed promising device-agnostic behavior, achieving F1 scores up to 0.83 on the custom wearable cohort while maintaining moderate generalization across other datasets. Detailed results are presented in Tables A.7, A.8, and Figure A.13.

Labeling: Labeling strategy emerged as a critical factor shaping model generalization across datasets. Models trained on expert-annotated data demonstrated relatively strong transferability to both stimulus-labeled datasets (best F1 = 0.72) and self-reported datasets (best F1 = 0.76), suggesting that expert-generated labels provide stable and temporally consistent references for learning physiological emotion patterns. Interestingly, models trained on stimulus-derived labels also generalized well to expert-annotated datasets (best

F1 = 0.87 with LDA), indicating some alignment between controlled elicitation protocols and expert interpretation. However, stimulus-labeled models showed weaker transfer to self-reported datasets (best F1 = 0.63), likely reflecting the limited ability of externally induced stimuli to fully capture individuals' subjective emotional experiences. Similarly, models trained on self-reported labels exhibited inconsistent transferability; they performed strongly when tested on expert-labeled data (best F1 = 0.87 with CLSP CNN at 5% and RF) but struggled with stimulus-labeled datasets (F1 = 0.62%), highlighting the variability inherent in subjective emotion reporting. Notably, zero-shot CLSP models trained with EDA achieved the highest overall performance (F1 = 0.91) on expert-labeled data, further underscoring the value of high-quality annotations combined with robust pretraining. Taken together, these findings suggest that expert and stimulus-based labels tend to offer greater consistency because they reflect a more external or standardized perspective on emotion, whereas self-reported labels capture subjective emotional experience but introduce greater variability. The results point toward the potential value of hybrid labeling approaches that integrate subjective self-reports with expert interpretation and contextual information to balance ecological validity with consistency, ultimately supporting more reliable physiological emotion recognition systems.

Age and Gender: For arousal classification, results indicate only moderate generalization both within and across gender groups (F1 0.50–0.56). In zero-shot CLSP transfer settings, EDA consistently outperformed both PPG alone and the combined EDA+PPG modality, often by a noticeable margin, suggesting that electrodermal activity may capture arousal-related physiological dynamics more robustly across demographic variability. In contrast, valence classification shows a different trend. Cross-gender transfer achieved substantially higher performance (F1 0.69–0.71 with fine-tuned CLSP models) than within-gender evaluations, where performance dropped to approximately 0.47–0.55. A similar pattern emerged in age-based transfer experiments: cross-age generalization for valence (F1 0.67–0.73) exceeded within-age-group performance, indicating stronger transferability

across demographic cohorts than within them. Overall, these cross-demographic findings suggest that physiological markers of emotional valence may reflect broadly shared emotional processes that generalize relatively well across gender and age groups. In contrast, arousal-related physiological responses appear more individualized and sensitive to participant-specific factors. Consequently, while valence recognition models may benefit from demographic diversity during training, arousal modeling may require more personalized or context-aware approaches. Detailed gender-based results are reported in Tables A.11 and A.12, and age-based analyses are presented in Tables A.13 and A.14.

3.3.3 Meta Analysis Observations

Artifact analysis was conducted prior to the pre-processing or standardization step. The results are presented in Table 3.3, where artifact presence is reported as the mean artifact percentage ± standard deviation across participants. Our artifact detection results showed that datasets collected in real-life settings (such as Nurse, ADARP, and Laurate) or using physical stressors as emotion-elicitation tasks (such as Exercise and ScientISST_MOVE) tend to exhibit more EDA and PPG motion artifacts. We further observed that datasets such as WESAD and CLAS have balanced class distributions across arousal and valence, whereas datasets such as ForDigitStress, Nurse, ADARP, and VERBIO exhibit high imbalance, skewing toward negative samples. Moreover, we identified that data collected "in the wild" settings generally featured more negative-valence instances, while video-based elicitation datasets have a higher proportion of high-arousal samples.

3.4 Discussion

Overall, the benchmarking results reveal several methodologically important findings for physiological signal–based emotion recognition. Below, we will discuss these findings in detail:

1) Limitations of Isolated Data Benchmarks: We observe performance saturation in

several controlled laboratory datasets, where relatively simple models achieve F1 scores above 0.95 in binary emotion classification tasks. This pattern is not consistent across datasets and depends on the interaction between dataset design, elicitation protocol, and physiological modality (e.g., EDA, PPG). Datasets such as WESAD and ForDigitStress, with well-defined, structured elicitation protocols, yield physiological responses that are more separable in feature space, enabling simple models to achieve high performance. Under these conditions, high F1 scores appear to depend more on the extent to which the experimental design increases signal separability than on model robustness alone. At the same time, performance differences across datasets indicated a more complex relationship between model behavior and dataset characteristics. Some datasets yield near-ceiling performance, while others produce substantially lower results despite using the same models and feature representations. This variation suggests that performance depends strongly on dataset-specific factors, including elicitation strength, alignment between physiological responses and emotion labels, and the suitability of the modality for capturing the target states. In weakly elicited datasets such as EMOGNITION, UBFC-Phys, PhyMER, and CASE, physiological changes are often subtle or inconsistent. Passive stimuli or low-intensity tasks, such as mental arithmetic, may not elicit sufficiently distinct emotional states, leading to weaker physiological differentiation and lower correspondence between labels and biosignals. In these cases, reduced model performance may reflect limited discriminative structure in the data rather than limitations of the models themselves. We also observe that real-world datasets with specific participant populations, such as NURSE and ADARP, generally yield lower performance than datasets collected from healthier, more heterogeneous cohorts in everyday settings, such as DAPPER and Unobtrusive. In these datasets, participant conditions and recording contexts are often dominated by a limited set of emotion states, such as prolonged stress or alcohol-related impairment, which can reduce class balance and limit variability across emotional categories. This creates difficulties for binary or multi-class classification, particularly when discrete labels do not adequately

reflect gradual or overlapping emotional states. Under these conditions, continuous or ordinal formulations may better capture the underlying structure of the data than discrete categorical representations. Across datasets, modality choice also interacts with elicitation protocol and labeling strategy in non-trivial ways. Certain elicitation procedures produce clearer responses in specific modalities, while others fail to generate stable physiological patterns regardless of sensor type. This further indicates that benchmark performance cannot be attributed solely to model architecture, and that better evaluation metrics are needed for comparing emotion recognition models.

Overall, these findings suggest that performance emerges from the interaction between dataset construction, elicitation protocol, sensing modality, and model complexity. High performance score is typically observed when these components are aligned and when physiological responses exhibit strong separability. Lower performance also reflects weak alignment between experimental design and physiological expression, or limited signal structure, rather than inherent model limitations. This changes the interpretation of high benchmark scores, which may primarily reflect favorable experimental conditions and strong signal separability rather than robust emotion recognition under naturalistic conditions. These observations have implications for future evaluation protocols. Rather than focusing primarily on separability within a small number of datasets, future work should evaluate robustness under varying conditions, including lower signal quality, increased noise, and ecologically valid recording settings. Evaluation protocols should span multiple elicitation intensities, participant populations, and acquisition environments within a unified framework. Evaluations should also include performance evaluations under both strong and weak induction conditions to capture variability in physiological expression. Cross-dataset validation should be treated as a core requirement for assessing generalisability across experimental contexts. Together, these changes would support more realistic evaluation criteria and improve the extent to which benchmark performance reflects model capability in real-world emotion recognition settings.

2) Label construction and semantic instability: Our results further indicate that emotion label semantics are not fixed and can vary substantially across experimental protocols and participant interpretations. In our benchmarking experiments, performance differences between arousal and valence prediction were highly dataset-dependent, with certain datasets yielding strong results on one dimension while performing poorly on the other. In cross-dataset evaluations, we also observe limited generalization between datasets with labels derived from self-reports and those derived from stimulus-based annotation schemes. Qualitative analysis suggests that these discrepancies are strongly influenced by the labeling strategy. In self-report–based datasets (see Table A.1), labels reflect subjective interpretation, individual reporting strategies, and contextual framing effects rather than a consistent or shared mapping to an underlying emotion taxonomy. Participants may report recalled or inferred emotional states rather than directly mapping experience onto predefined theoretical dimensions [71]. Consequently, similar physiological patterns can be assigned different labels depending on the annotation protocol, introducing inconsistency that directly impacts model training and evaluation. In contrast, stimulus-derived labels tend to be more consistent across participants due to their standardized elicitation structure; however, they may not accurately reflect internally experienced emotional states, limiting their ecological validity. Finally, expert annotations depend on skill and knowledge and are infeasible in real-world settings. We also observed that mapping composite questionnaire scores to a dimensional affect model, as in the Laureate dataset, can introduce label ambiguity and reduce alignment with physiological signals. In this case, self-reported measures of enthusiasm, stress, tiredness, happiness, and calmness were aggregated into arousal and valence scores, which were then used to assign labels to physiological segments. However, collapsing multiple affective constructs into composite dimensions can weaken the correspondence between labels and biosignals, potentially reducing model performance. This suggests that future datasets should use consistent labeling schemes and measurement scales, as heterogeneous or non-standard scaling can hinder cross-dataset comparability and reliable benchmarking,

and overall highlights the need for annotation frameworks that improve cross-study label consistency, explicitly model ambiguity in self-reported emotions, and better account for variability in how participants interpret, construct, and express emotional experience. Moreover, future work can also explore signal-specific attributes as pseudo-labels for pre-training to mitigate issues arising from labeling bias.

3) Modality effects and representation learning limitations: We observe that model performance is strongly conditioned by both signal modality and learning strategy, with consistent patterns across datasets. In most settings, models based on EDA data achieve the strongest performance, indicating that EDA provides comparatively robust and discriminative information for emotion recognition under both controlled and real-world conditions. In contrast, PPG-only models tend to be less stable, with more variable performance across datasets. Multimodal fusion using EDA+PPG yields mixed outcomes: it improves performance in some datasets, particularly in noisy real-world scenarios, but provides limited or no benefit in others. This suggests that naive fusion strategies do not reliably translate into improved generalization and that the utility of multimodal signals is highly dataset-dependent. A second consistent finding is the advantage of pretrained representations, particularly CLSP-based models, especially under fine-tuning. Across datasets and evaluation regimes, pretrained models frequently outperform training-from-scratch baselines. This pattern indicates that physiological representation learning benefits substantially from large-scale pretraining and cross-dataset transfer, rather than relying exclusively on dataset-specific optimization. It also underscores the potential value of leveraging natural-language supervision signals for learning more transferable emotion representations. Collectively, these results highlight that both modality selection and training strategy are critical determinants of performance, alongside dataset design. They further suggest that progress in physiological emotion recognition is likely to depend less on isolated dataset-specific modeling and more on learning unified, generalizable representations through large-scale pretraining and principled multimodal integration across heterogeneous datasets and real-world conditions.

4) Shared Data Collection and Labeling Protocols for Transferable Emotion Models: Finally, our cross-dataset results further indicate that progress in physiological emotion recognition depends not only on model design, but also on improving standardization in dataset construction. In particular, aligning sensing configurations, modality sets, and annotation schemes across studies would enable more consistent representation spaces and improve the ability of models trained on one dataset to generalize to others. The substantial variability observed across existing datasets suggests that differences in acquisition hardware, elicitation protocols, and labeling strategies remain key barriers to scalable pretraining and robust cross-domain transfer. We also observe that datasets collected under semi-structured or constraint-based settings tend to exhibit stronger transferability across domains, suggesting that hybrid designs combining controlled elicitation with elements of ecological variability may provide a more effective balance between experimental control and real-world realism. Overall, these findings support a shift away from isolated laboratory-specific dataset construction toward more harmonized data collection frameworks to facilitate the development of higher-quality physiological datasets.

3.5 Limitations and Social Impact

This work benchmarks a representative set of modeling paradigms to establish a broad baseline, primarily using traditional, general-purpose architectures. However, more advanced or domain-specific models tailored to physiological signals were not included and remain an important direction for future benchmarking. Additionally, our current harmonization strategy is limited by the lack of a common labeling technique. Moreover, our analysis does not account for key sources of heterogeneity, including cultural context and health status; we plan to incorporate them in the future. By systematically studying these limitations, we hope to contribute to the development of responsible, generalizable, and impactful emotion recognition technologies for applications in HCI, affective computing, and mental health support.

3.6 Summary

Building on the observations from this large-scale benchmarking study, the results point to a clear need for more intentional dataset design practices that explicitly address ecological validity, annotation reliability, sensing heterogeneity, and participant diversity, rather than treating these factors as secondary concerns. The findings further emphasize the importance of strengthening labeling methodologies in future work, along with developing more rigorous and transparent evaluation protocols. Collectively, this motivates the need for standardized experimental procedures, sensor configurations, preprocessing pipelines, and annotation strategies to enable more meaningful and comparable benchmarking. Without such harmonization, the proliferation of small, heterogeneous datasets risks fragmenting the field and limiting cumulative scientific progress, despite increasing data availability. The results also reinforce the importance of participant-centered data collection approaches. Emotional experience is inherently subjective and context-dependent, and models trained exclusively on controlled environments or externally imposed labels risk overlooking important nuances of lived emotional states. More broadly, these findings support a shift from isolated, dataset-driven modeling toward integrated data ecosystems supported by standardized evaluation frameworks. Such a shift would improve reproducibility, facilitate more effective use of existing resources, and support the development of models that are not only have high benchmarking performances but also meaningful for applications in emotional well-being, human–computer interaction, and mental health support.

Motivated by these insights, Chapter 4 introduces the development of the EEVR dataset and CLSP-based modeling framework, designed to explicitly account for label variability in semi-realistic laboratory data collection. Chapter 5 then examines participant perspectives on emotion data collection and existing annotation practices, with a focus on participant-centered approaches to dataset design that can improve both model performance and cross-domain generalization. Finally, Chapter 6 builds on these findings to explore the development

of more participant-aware and usability-oriented tools for emotion data collection in real-world settings.

Chapter 4

Emotions in *Context*: Turning Data into Insights

Current approaches to collecting physiological emotion data largely depend on self-reported annotations derived from standardized emotion questionnaires, theory-driven scales, or stimulus-based labeling conventions (e.g., assuming relaxation when a calming video is presented). Commonly used instruments include objective scales such as the Visual Analogue Scale (VAS), Positive and Negative Affect Schedule (PANAS), Self-Assessment Manikin (SAM), Likert-type emotion ratings, and standardized psychological inventories such as the State–Trait Anxiety Inventory (STAI). While these tools provide structured and comparable measurements, they do not always capture the complexity of lived emotional experience. Stimulus-based labeling, in particular, assumes that exposure to a specific stimulus reliably induces a corresponding emotional state, which may not reflect participants' actual subjective experiences. Taken together, reliance solely on self-reports based on standardized scales or stimulus labels can overlook subtle emotional nuances, mixed or rapidly changing emotional states, and even periods where no clear emotion is experienced. Such approaches are also susceptible to reporting biases, interpretation differences, and human error as discussed in Chapter 4. As a result, these conventional methods may limit the accuracy and ecological validity of physiological emotion datasets.

To address these limitations, in this chapter, we introduce *EEVR*, a physiological signal-based emotion dataset collected in laboratory settings using 360° VR audiovisual stimuli. EEVR includes data from the two most commonly available physiological sensors in commercial wearable devices, Photoplethysmography (PPG) and Electrodermal Activity (EDA), which have been widely collected in previous datasets. Emotional annotations were obtained through subjective evaluations using the PANAS and SAM emotion scales, along with self-reported raw textual descriptions of emotions felt by subjects during stimulus exposure.

These descriptions were gathered through semi-structured qualitative interviews, providing a more contextualized understanding of emotions and allowing participants to elaborate on their emotional experiences in detail. EEVR is the first dataset to collect raw textual data for broader supervision, capturing the presence or absence of emotions experienced during the stimulus. This approach to collecting subjective textual responses to emotions has not been explored before. EEVR includes data from 37 participants who experienced emotions across all four quadrants of Russell's circumplex model. Additionally, it contains personality scores for each subject, collected using the Big Five Inventory 10 Item Scale (BFI-10) ([146]), and the psychological well-being details of each subject using the General Health Questionnaire-12 (GHQ-12) ([147]). In Table 4.1 we have compared our dataset with previous datasets that have collected physiological signals for emotion recognition. Further details and access to the dataset are available at https://melangelabiiitd.github.io/EEVR/. Through this work, we make the following contributions:

- A novel multimodal physiological signal dataset collected in an immersive lab setting with aligned raw textual descriptions of emotions felt and self-reported valence and arousal scores.
- A readily replicable experimental procedure for capturing physiological response and textual descriptions within lab settings.
- We provide guidance on utilizing the dataset, along with open-source access to baseline models and the Contrastive Language-Signal Pre-training (CLSP) models, which leverage text supervision to learn more contextualized representations of emotions by combining physiological signals with text data.

4.1 Motivation for Paired Textual Descriptions

Supervision through language or text has become a focal point in computer vision after the introduction of CLIP ([148]). The emergence of large language models has also spurred

Dataset	#Subjects	Stimuli	Data Modalities	Annotations
MANHOB HCI	27	Audiovisual	ECG, GSR/EDA, RESP, TEMP, EYE GAZING, EEG, Facial Expressions and Audio.	Emotions, Arousal, Valence, Dominance.
DEAP	32	Music Video clip	EEG, ECG, PPG, GSR/EDA, EMG (Trapezius, Zygomaticus Muscle).	Arousal, Valence, Liking, Dominance and Familiarity.
WESAD	15	TSST, Audiovisual	ECG, EDA, EMG, BVP, Respiration, Temperature, Acceleration.	Stressor-based, PANAS, STAI, SAM.
CLAS	62	Cognitive load, Audiovisual	ECG, PPG, EDA, Acceleration.	SAM.
SWELL-KW	25	Office work with interruptions and time pressure	ECG, EDA, Face and upper body video, Posture, Computer logging.	NASA task load, SAM, Stress.
EMOGNITION	43	Audiovisual	PPG, EDA, BVP, Temperature, Acceleration, cardiac output measurement, Facial Expression.	Arousal, Valence, Avoidance Approach Motivation, Emotions.
AMIGOS	40	Audiovisual	EEG, ECG, EDA/GSR.	PANAS, SAM, Liking, Familiarity, Personality, Emotions.
ASCERTAIN	58	Audiovisual	GSR, Frontal EEG, ECG, Facial Landmarks.	SAM, Familiarity, Personality, Emotions.
DREAMER	23	Audiovisual	EEG, ECG.	SAM.
KEMOCON	32	10 Minute long paired debate on social issues.	PPG, EDA, BVP, Temperature, Acceleration, EEG, ECG.	Arousal, Valence, Emotional Labels.
BIRAFFE2	103	Music, Images, Games	ECG, EDA, Gamepad Acceleration, Gyroscope.	SAM, Personality, Game experience.
CASE	30	Audiovisual	ECG, BVP, EMG, EDA/GSR, Respiration, Temperature.	SAM.
StressID	65	Cognitive load Audio-Visual Public Speaking	ECG, EDA, Respiration, Speech, face video.	SAM, Stress.
VREED	34	360 degree VR	ECG, EDA, Eye Tracking.	SAM, Emotions.
EEVR (Ours)	**37**	**360 degree VR**	**PPG, EDA.**	**Emotions, Arousal, Valence, Dominance Familiarity, Liking, Personality, GHQ-12, Textual Description.**

Table 4.1: EEVR in comparison with other related datasets

an increase in research exploring language-guided supervision. This trend extends beyond vision and language, with modalities like audio ([149]) and video ([150]]) leveraging language supervision for pre-training to enhance generalization and usability. Concurrently, text-based pre-training methods have revolutionized the NLP domain in recent years. Despite the prominence of such approaches in various fields, there remains a notable gap in utilizing language-guided supervision for emotional recognition through physiological signals. While previous research has looked into leveraging text data (such as social media posts, text messages, and suicide notes) for emotion recognition ([151]), none of these efforts involved recording self-reported emotional descriptions from subjects alongside the collection of physiological signal data. Therefore, the EEVR dataset is an initiative in this regard, prompting new avenues for collecting emotional data based on physiological signals.

4.2 EEVR Dataset

4.2.1 Experimental Protocol

Our experiment protocol to collect the EEVR dataset is illustrated in Figure 5.1. The experiment was conducted using VR 360° audiovisual stimuli. The stimuli consist of *N=8* short videos (two videos from each quadrant of the Russell circumplex model ([72])) covering all emotions. Next, we explain the data collection procedure as follows:

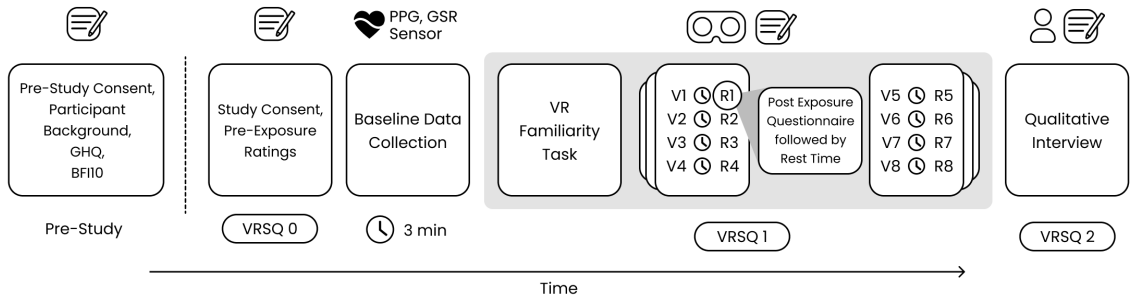


Figure 4.1: Illustration of our Experiment protocol for data collection.

1. **Pre-Study Survey:** The data collection is initiated by collecting participants' consent. Then, we collected participants' background details (gender, age, educational

background, and prior VR exposure), personality scores using BFI-10 questionnaires, and information on prior psychological well-being using GHQ-12, a popular screening questionnaire designed to identify common psychiatric conditions within non-psychiatric clinical and research settings. We have used *over the past week* version of the GHQ-12 questionnaire to collect participants' psychological well-being before participating in our study. GHQ-12 helps mitigate the bias that might be introduced if the participant experienced some psychological lows or highs prior to data collection. The GHQ-12 and personality scores were added as additional contextual details alongside other participant characteristics in our dataset.

2. **Pre-VR exposure:** Following the pre-study survey, participants were introduced to the experiment protocol and sensor setup along with instructions for data collection and the study's possible risks or discomfort (more details in section 4.1 of supplementary). Following instructions, participants **pre-exposure ratings** were gathered. This included PANAS, SAM, and Virtual Reality Sickness Questionnaire (VRSQ) ([152]) scales to collect participants' baseline emotions and pre-VR sickness and fatigue symptoms, if any. The PANAS scale was used to collect positive and negative affect readings on a 5-point scale with ten positive (Interested, Strong, Enthusiastic, Proud, Inspired, Determined, Alert, Attentive, Active) and ten negative (Distressed, Irritable, Guilty, Scared, Upset, Hostile, Jittery, Ashamed, Nervous, Afraid) emotions. The SAM scale was used to collect scores for emotions' valence, arousal, and dominance dimensions. Following baseline ratings, PPG and EDA sensors were attached to the participant's non-dominant hand fingers. Subsequently, participants were asked to relax for 3 minutes of **baseline data collection**. More details on scales are provided in supplementary section 4.4.

3. **VR Familiarity:** Following baseline, participants were familiarized with the VR environment to mitigate any bias that may arise due to VR familiarity by making

them initially sit in a VR waiting room to acclimate to the technology, exploring the surroundings by looking around for approx 4 minutes. Following this, participants transitioned to familiarizing themselves with the VR controller by engaging in a simple game where they used the VR controller to pick up and throw a ball into a box within the VR room. More details on the VR module are provided in supplementary section 4.3.

4. **VR Stimulus Exposure**: Participants were instructed to choose the assigned playlist after the VR familiarity task. Each playlist contained eight videos retrieved from a public database of annotated 360° Videos ([153]). These videos were selected to elicit emotions from all four emotional quadrants of the Russell circumplex model ([72]). The circumplex model organizes emotions based on two dimensions: valence and arousal. Thus, videos from four categories were shown to participants: **High Valence-High Arousal (HVHA)**: elicits high energy positive emotions (such as excitement, Joy) **High Valence-Low Arousal (HVLA)**: elicits low energy positive emotions (such as calmness, relaxation) **Low Valence-High Arousal (LVHA)**: elicits high energy negative emotions (such as stress, anger) **Low Valence-Low Arousal (LVLA)**: elicits low energy negative emotions (such as boredom, depression)

   Following each VR video, *post-exposure ratings* were collected from subjects to annotate their emotions during the VR exposure. The post-exposure questionnaire was the same as the pre-exposure, with additional questions about familiarity and liking ([154]) of content. The familiarity score was collected on a 1-5 scale. Between subsequent VR videos and self-reporting, participants were given rest periods to avoid VR sickness or fatigue. Additionally, participants were asked to fill *VRSQ* after completing the fourth and eighth VR videos. More details on stimulus selection, stimulus order, and playlist creation are provided in section 4.2 of the supplementary document.

5. **Qualitative Interview:** At the end of physiological data collection, the participant's sensors were removed following the qualitative interview. The semi-structured interviews allowed us to adapt the questions based on the participants' feedback. The objective of the interview was to prompt the participants to articulate the emotions that they experienced while watching the VR stimulus and the reason behind those emotions. We used a monitor to show the VR videos from the assigned playlist in order to support the participants in recalling the stimulus while explaining the emotions. The questions like, *"What was the major emotion felt in this video (referring to the video)?"* and *Were there any mixed emotions that you (participant) felt while being exposed to stimulus* were asked to capture the subjective experiences. The interview was audio-recorded after obtaining consent from participants. Later, the audio recording is converted into text during dataset preparation using Google speech-to-text API [1]. The data is then manually cleaned to extract each subject's response to the interviewer's questions.

4.2.2 Experimental Setup

EEVR consists of two physiological signals: Electrodermal Activity (EDA) and Photoplethysmography (PPG). The physiological signals are recorded using the *4-channel Biopac MP36* [2] system. The Biopac MP36 consists of 4 channels to collect a maximum of four synchronized signals simultaneously. The MP36 system was connected to BSL4 data acquisition software to visualize and store the physiological signal data and to the peripheral PPG and EDA sensors. The EDA sensor module (SS57LA Hardware module [3]) was attached to the index and middle fingers ([155]) of the participant's non-dominant hand, utilizing EL507 Electrodes for collecting users' skin electrical conductance. The PPG sensor module (SS4LA Hardware module [4]) was attached to the participant's non-dominant hand ring finger.

[1] https://cloud.google.com/speech-to-text
[2] https://www.biopac.com/product/mp36r-systems/
[3] https://www.biopac.com/product/eda-lead-bsl/
[4] https://www.biopac.com/product/photoplethysmogram-for-pulse-waveform-bsl/

Non-dominant was used to attach sensors for minimizing noise due to motion artefacts. Before attaching the sensors, Isotonic Gel was applied to EDA electrodes to ensure minimal noise in the collected data. The biopac MP36 has been used in prior research for collecting physiological signal data ([156, 157, 158, 159]). It has a resolution of 2000 Hz for all acquired physiological signals. For 360° video stimulus, a *Meta Quest Pro* headset was used. This headset has 2 x LCD panels with 1800 x 1920 pixels per eye, a refresh rate of 90Hz, and a 106º Horizontal × 96º Vertical Field of view. It incorporates eye relief adjustment, lens spacing, and spatial audio support. We have used the OpenXR plugin to integrate the Meta Quest pro headset with the Unity application. OpenXR plugin also helped us with hand gestures and controls for interacting with the application's user interface. Pre and post-exposure ratings were collected using iPad Pro Tablet.

4.2.3 Participants and Experiment Details

EEVR comprised 37 healthy participants (21 males, 16 females) aged 18-33 (M=23.1, SD=4.02). Participants were from varying educational backgrounds - Bachelor (24), Master (8), Senior High School (4) and Doctorate (3). Our exclusion criteria exclude individuals with experience or a history of heart issues, heart arrhythmia, high blood pressure, medical conditions affecting equilibrium, visual or auditory impairments, neurological ailments, cognitive challenges, psychological issues, or diagnosed depression, as per the guidelines laid out in ([155]). Additionally, participants with low proficiency in the English language were not included in the study to avoid any impact of language understanding on the participants. All participants included in the study were requested to sign the consent form. Participants were also instructed to forbid any caffeine intake and refrain from exercising 3 hours before the experiment. The study was conducted in an institute research lab with minimal disturbance. The experiment setup (room temperature and sitting arrangements) remained the same for all the participants. The experiment was conducted with the experimenters present in the lab.

4.2.4 Dataset Description

The EEVR dataset comprises 296 emotion tasks plus 37 baselines, with each of the 37 participants experiencing eight VR 360° videos. These tasks aim to gather physiological data, totaling approximately 797 minutes and 83 seconds, including each participant contributing 3 minutes of baseline data. Along with physiological data, 296 textual descriptions were collected from 37 participants through interviews. The physiological data segment was identified with video_ID and subject_ID. The EDA and PPG data were originally collected at a sampling frequency 2000Hz but were downsampled to 15.625Hz for EDA and 125Hz for PPG data. The downsampling was done to reduce computation costs while maintaining the data quality. Prior work has utilized EDA data at a sampling frequency of 4Hz and PPG of 64Hz minimum ([37]). More details about dataset preparation, cleaning, and analysis are provided in the supplementary section 5.

4.2.5 Annotation

All physiological data segments (Participant_ID - Video_ID) are annotated with self-reported ratings of arousal, valence, dominance (using SAM scale), discrete emotional ratings using PANAS (further used to calculate positive and negative affect scores), and additionally, we have a qualitative textual description for each data segment. Moreover, liking and familiarity scores on a scale of 1-5 are also present for each segment. Further, we have personality scores and GHQ-12 ratings for each participant. More details on the affect score and GHQ score calculation and annotation analysis have been added to the supplementary sections 4.4 and 5.1.

Labels for Supervised Learning: For supervised learning, we propose three 2-class labels based on both participants' responses to arousal, valence questionnaire, and based on stimulus annotations. The arousal data was collected on a scale of 1-5, which was further divided into binary classes by considering data with 0-3 ratings as low arousal and 4-5 as high arousal; we followed a similar process for categorizing valence data. The physiological

data collected during video stimulus from LVHA, LVLA, and baseline are annotated as negative emotions, while the videos from HVHA and HVLA are annotated as positive, creating binary classes. The baseline was annotated as negative valence, considering the stress that participants may undergo due to the sensor attachment procedure and activities before the experiment. Upon analysis, we found our arousal labels were skewed compared to other labels. To overcome this skewness, we have used oversampling for the arousal labels.

4.3 Experiments

4.3.1 Baseline

We conducted baseline classification for three tasks: Arousal Classification, Valence Classification, and Stimulus-label-based Emotion Classification. Each task involved binary labels. We have performed baseline classification separately for each data modality: EDA and PPG for all three labels. Followed by multimodal classification of physiological signals combining EDA and PPG data. All physiological signal-based baseline experiments were conducted using Leave-one-subject-out (LOSO) cross-validation ([36]). All the results are presented in Table 4.2 as the average performance across all LOSO subjects, calculated over three different seed values. To validate our text data, we also performed baseline classification tasks for only text data, and the results are presented in Table 4.2. Next, we conducted contrastive training ([148]) to present our pre-training method using the paired physiological signals and textual data. The baseline results with or without contrastive training on 296 (excluding baseline samples) text-physiological signal pairs are presented in Table 4.3. The code for all baseline implementations is present here `https://github.com/alchemy18/EEVR/`. Next, we present more details on physiological signals, text, and contrastive baseline.

Modality	Models	Stimulus-label		Valence		Arousal	
		Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
EDA	**Logistic Regression**	86.78 ± 0	0.82 ± 0	61.56 ± 0	0.71 ± 0	47.41 ± 0	0.36 ± 0
	Decision Tree	85.09 ± 0.17	0.83 ± 0	58.46 ± 1.06	0.64 ± 0.01	54.05 ± 1.12	0.35 ± 0.02
	Random Forest	**90.79 ± 0.46**	**0.89 ± 0.01**	60.26 ± 1.81	0.66 ± 0.01	57.23 ± 1.19	0.28 ± 0.04
	LDA	87.69 ± 0	0.85 ± 0	**61.86 ± 0**	**0.69 ± 0**	48.97 ± 0	0.37 ± 0
	XGBoost	90.69 ± 0.52	0.89 ± 0.01	59.76 ± 0.52	0.66 ± 0.01	56.61 ± 0.34	0.37 ± 0.01
	SVM	85.29 ± 0	0.81 ± 0	59.16 ± 0	0.71 ± 0	51.66 ± 0	**0.44 ± 0**
	MLP	87.39 ± 0	0.85 ± 0	61.86 ± 0	0.68 ± 0	**57.27 ± 0**	0.39 ± 0
PPG	**Logistic Regression**	81.08 ± 0	0.77 ± 0	61.26 ± 0	0.70 ± 0	**56.29 ± 0**	**0.42 ± 0**
	Decision Tree	68.87 ± 0.35	0.65 ± 0	54.35 ± 0.30	0.59 ± 0.01	49.43 ± 0.32	0.26 ± 0.01
	Random Forest	75.88 ± 0.35	0.69 ± 0.01	**61.66 ± 1.93**	**0.70 ± 0**	49.27 ± 0.42	0.18 ± 0.01
	LDA	**81.08 ± 0**	**0.78 ± 0**	58.96 ± 1.73	0.67 ± 0.06	54.47 ± 3.72	0.40 ± 0.02
	XGBoost	49.44 ± 0	0.68 ± 0	57.26 ± 0.76	0.64 ± 0.01	47.89 ± 7.57	0.26 ± 0.13
	SVM	80.48 ± 0	0.75 ± 0	59.86 ± 1.91	0.70 ± 0.05	47.99 ± 3.78	0.32 ± 0.10
	MLP	78.68 ± 0	0.75 ± 0	56.76 ± 0	0.66 ± 0	54.16 ± 0	0.38 ± 0
PPG + EDA	**Logistic Regression**	85.89 ± 0	0.82 ± 0	60.06 ± 0	0.69 ± 0	55.23 ± 0	0.41 ± 0
	Decision Tree	83.78 ± 0.80	0.83 ± 0.01	**62.77 ± 0.30**	0.66 ± 0	**58.13 ± 0.70**	0.40 ± 0.01
	Random Forest	**90.69 ± 0**	**0.89 ± 0**	61.06 ± 1.35	0.70 ± 0.01	56.78 ± 1.56	0.26 ± 0.01
	LDA	84.89 ± 1.39	0.82 ± 0.01	57.56 ± 2.95	0.66 ± 0.06	55.48 ± 1.04	**0.42 ± 0.01**
	XGBoost	87.19 ± 2.73	0.85 ± 0.03	61.36 ± 4.79	0.67 ± 0.04	58.0 ± 1.66	0.36 ± 0.06
	SVM	87.29 ± 1.39	0.84 ± 0.02	62.16 ± 2.08	**0.72 ± 0.02**	55.97 ± 3.44	0.38 ± 0.04
	MLP	83.48 ± 0	0.81 ± 0	58.86 ± 0	0.63 ± 0	56.89 ± 1.47	0.36 ± 0.03

Table 4.2: Results for Arousal Classification, Valence Classification, and Stimulus-label-based Emotion Classification on EDA and PPG Data

Physiological Signal Baseline

We present our baseline model using hand-crafted features from our physiological signal data, similar to prior work on emotion recognition ([37, 160, 161, 16]). To extract our features, we performed the following steps: First, each EDA and PPG data segment is filtered. EDA data is filtered using a low-pass filter with a 5Hz cutoff frequency and a 4th-order Butterworth filter, while PPG data is cleaned using a bandpass filter. Next, the EDA data is decomposed into tonic and phasic components, referred to as skin conductance level (SCL) and skin conductance response (SCR) using cvxEDA, a convex optimization-based approach ([162]). We then extracted statistical features, including dynamic range and slope, from both SCR and SCL components and time-domain features from SCR, such as the number of peaks, average amplitude, and duration. For PPG data, feature extraction is performed using the Neurokit Library ([163]) to extract HRV-related time domain, frequency domain, and non-linear features. After feature extraction, the features are normalized participant-wise using min-max scaling. Classical machine learning algorithms

are then applied to the extracted features for all three classification tasks. We combine the handcrafted features from both modalities to train our multimodal machine-learning models using both EDA and PPG data. All machine learning models are trained using default hyperparameters from sklearn. For the training MLP, we used two hidden layers with 50 and 100 dimensions. All classification models are trained using the following seeds: 42, 43, and 111. Results are presented in table 4.2 and additional details about experiments are provided in supplementary section 5.5.

Textual Data Baseline

Next, we performed the three classification tasks on our textual data (for 296 samples excluding baseline). We followed the standard text classification pipeline, starting with data preprocessing, which includes data cleaning (removing stopwords and punctuation, converting to lowercase) and lemmatization. Following this, we applied tokenization, and then we fine-tuned two pre-trained models (DistilBERT ([164]) and XLM-RoBERTa Base ([165])) for the classification tasks. For our experiments, we have generated five random splits, with 80% of the data used for training and 20% for testing. Results are presented in table 4.2.

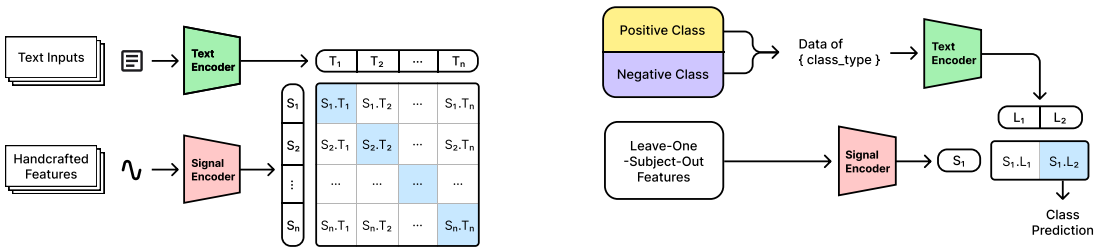


Figure 4.2: The Architecture for Contrastive-Language Singal Pre-Training (CLSP).

4.3.2 Contrastive Language-Signal Pre-training

To underscore the importance of integrating textual descriptions in emotion recognition, we introduce the Contrastive Language-Signal Pre-training (CLSP) method for extracting

more contextualized representations. The model was trained on physiological signals and text pairs to learn a joint embedding space, where both modalities are closely aligned using a contrastive loss function [148]. Following pre-training, we evaluated the model's performance on test subject data using the leave-one-subject-out cross-validation approach, leveraging minimal labels generated in the format "Data of {class_Type}" (e.g., "Data of positive emotion class"). CLSP employs separate neural networks to process the handcrafted features of physiological signals (PPG and EDA signal data) and text data. For signal data, linear layers with hidden dimensions of 50 and 100 are utilized, while the text data is processed using a pre-trained DistilBERT (transformer-based language model). These extracted feature representations are then used to optimize a contrastive objective, maximizing the similarity between positive pairs and minimizing it for negative pairs. The detailed architecture is depicted in Figure 4.2, and our results for CLSP are summarized in table 4.3. We found that the emotion recognition for arousal and valence tasks using the CLSP method led to significant improvement in classification results compared to without-text-supervision (Hand-crafted features + Neural Network (two linear layers of dimensions 50, 100)) training. This highlights the effectiveness of incorporating qualitative textual descriptions into physiological signal-based emotion representation learning. Further experimental details and comprehensive discussions are provided in Supplementary Section 5.5.

Modality	**Model**	**Stimulus-label**		**Valence**		**Arousal**	
		Accuracy	**F1 Score**	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**
EDA	**HC+NN**	**87.39**	**0.85**	61.86	0.68	57.27	0.39
PPG	**HC+NN**	78.68	0.75	56.76	0.66	54.16	0.38
EDA+PPG	**HC+NN**	83.48	0.81	58.86	0.63	58.58	0.40
EDA+Text	**CLSP**	64.19	0.68	**70.38**	**0.73**	**77.25**	**0.81**
PPG+Text	**CLSP**	56.95	0.53	64.74	0.64	69.91	0.62
EDA+PPG+Text	**CLSP**	53.50	0.48	64.87	0.60	69.64	0.64

Table 4.3: Results for Physiological Baseline without text using Hand-crafted features + NN and with text using CLSP on 296 text-signal pairs for seed=43 and epoch=15.

4.3.3 Zero-Shot Transfer

To assess the generalization capabilities of our pre-trained CLSP models across datasets collected in varied environments, we conducted a zero-shot transferability evaluation on our pre-trained model. For these experiments, we utilized three datasets representing distinct data collection settings: Emognition, acquired using 2D video stimuli in laboratory settings ([166]), WESAD, gathered using the TSST psychological task and video stimuli within controlled lab conditions ([37]), and NURSE, recorded in real-life hospital environments during the COVID-19 pandemic ([167]). As detailed in Table 4.4, our pre-trained model demonstrated the ability to predict emotions in these new domains with accuracy comparable to the baseline models, and in several instances, it even surpassed the performance of supervised baselines. These findings show the effectiveness of integrating text-based emotion descriptions for learning representations that transfer robustly across diverse data domains, irrespective of the environment, device, or participant demographics. To ensure fair comparisons, we employed a standardized pipeline encompassing data cleaning, participant-wise normalization, feature extraction, and classification across all experiments.

4.4 Discussion

4.4.1 Discussion on Physiological Baseline

Our baseline results for the Valence and Stimulus_Label classification task across all classical machine-learning models were better than random, suggesting that our models are able to separate features for these labels. We observed that stimulus labels are easy to predict using physiological signal-based features as compared to subjective labels like valence and arousal. The PPG+EDA features gave the best performance for valence classification. And EDA features provided the best performance for Stimulus_Label classification. The results were nearly random for Arousal classification even after performing duplicate upsampling, suggesting arousal classification is a difficult label to predict purely based on physiological

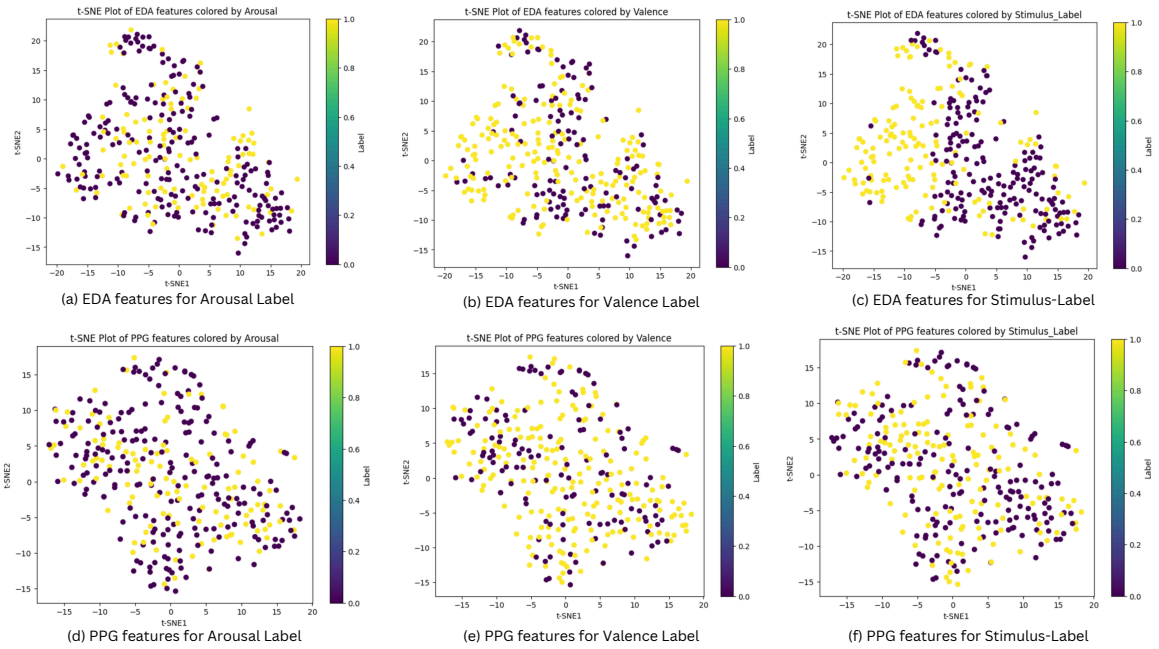
Dataset (Signal Type)	Method	Arousal		Valence	
		Accuracy	F1 Score	Accuracy	F1 Score
Emognition (EDA)	**MLP**	52.80	0.57	**61.89**	0.36
	Zero-shot CLSP	**53.23**	**0.59**	50.32	**0.49**
Emognition (PPG)	**MLP**	49.94	0.53	50.63	0.28
	Zero-shot CLSP	48.19	0.47	51.88	0.41
Emognition (EDA + PPG)	**MLP**	51.53	0.54	55.12	0.34
	Zero-shot CLSP	50.94	0.52	53.58	0.41
WESAD (EDA)	**MLP**	85.00	0.84	96.67	0.97
	Zero-shot CLSP	53.33	0.67	51.67	0.67
WESAD (PPG)	**MLP**	80.00	0.80	75.00	0.75
	Zero-shot CLSP	70.00	0.68	66.67	0.72
WESAD (EDA + PPG)	**MLP**	**91.67**	**0.91**	**98.33**	**0.98**
	Zero-shot CLSP	75.00	0.71	86.67	0.86
Nurse (EDA)	**MLP**	39.88	0.32	71.83	0.03
	Zero-shot CLSP	**55.48**	**0.58**	**84.93**	0.20
Nurse (PPG)	**MLP**	45.10	0.38	72.08	0.05
	Zero-shot CLSP	53.08	0.48	75.34	0.23
Nurse (EDA + PPG)	**MLP**	48.35	0.43	76.04	0.23
	Zero-shot CLSP	53.08	0.45	84.59	**0.42**

Table 4.4: Zero-shot transferability results of our pre-trained model (CLSP) compared to supervised baseline model trained on existing datasets (Emognition, WESAD, and Nurse)

signals-based features. We have visualized our EDA and PPG handcrafted features for all three tasks using t-SNE algorithms as shown in Figure 4.3. We observed that t-SNE features are separable for Stimulus_Label in the case of EDA data. While for other labels, the features are overlapping. This suggests the need for more complex models and better representation learning for valence and arousal prediction.

4.4.2 Discussion on CLSP

To assess the significance of aligning textual descriptions to physiological signal data, we performed CLSP training for all three labels using EDA only, PPG only, and EDA+PPG handcrafted features (for 296 tasks excluding baseline) along with text data. To compare our CLSP results, we first trained a hand-crafted features-based neural network (HC+NN) model with two hidden linear layers of dimensions 50 and 100. These models were trained for a batch size of 32 with 200 epochs, using a learning rate of 0.001. For optimization, we



(a) EDA features for Arousal Label (b) EDA features for Valence Label (c) EDA features for Stimulus-Label

(d) PPG features for Arousal Label (e) PPG features for Valence Label (f) PPG features for Stimulus-Label

Figure 4.3: t-SNE plot depicting feature distribution of physiological signals according to various labels (Arousal, Valence, and Stimulus-Label).

utilized the Adam optimizer with beta values of 0.9 for beta1 and 0.999 for beta2 and an epsilon value of 1e-8. The CLSP model was then trained for contrastive objectives using the HC+NN model for physiological signal embeddings and the DistillBert model with a projection head of a single linear layer with dimension 100 for text embeddings. The project head was added to match the dimensionality of the text embeddings (originally 768) with signal embedding. The training was conducted for 15 epochs with a learning rate of 0.001 and batch size of 32. We found that our results were significantly better for arousal and valence labels, suggesting the importance of augmenting text data for training subjective labels like Arousal and Valence. The results were not as good for Stimulus_Labels, indicating that the features extracted from signals do not complement the features extracted from text. This mismatch could have happened due to the subjective nature of textual descriptions that Stimulus_Labels cannot capture. This misalignment might have led to ineffective contrastive training. The EDA+Text outperformed PPG+Text and PPG+EDA+Text, suggesting that EDA features might align more with textual descriptions for arousal and valence labels.

We also observed that the early fusion of EDA and PPG features for CLSP has led to poor performance compared to EDA-only and PPG-only features, with Text indicating a need for designing a more complex fusion technique. Overall, our results suggest that aligning text data with physiological signals can improve learning for subjective labels, achieving results that cannot be attained with objective labels alone.

4.5 Limitations

EEVR dataset is collected using pre-annotated virtual reality videos within controlled laboratory settings. The experiment design does not consider the influence of external factors that may impact participants' emotional responses to immersive stimuli and assumes isolated responses to stimuli. Factors such as VR sickness, familiarity with VR technology, and attitudes toward this new technology may also affect participants' emotional responses. Furthermore, the placement of sensors and the VR headset can cause discomfort. Therefore, the signals recorded in this setting may not necessarily replicate real-life responses from all participants. Moreover, training based on participants' ratings is susceptible to participant bias, which may affect subsequent results. To address the subjectivity of emotional responses, we have collected qualitative responses in the form of textual descriptions, providing rich contextual annotations alongside objective ratings. Additionally, *EEVR* contains data from a privileged set (upper middle class, educated) of the audience and does not represent other sections of society and thus is biased towards a specific society group.

4.6 Ethical Considerations and Dataset Accessibility

EEVR study is approved by the Institution review board [5] of IIIT-Delhi registered with the National Ethics Committee Registry for Biomedical and Health Research (NECRBHR). All participants in this study provided explicit consent for recording their physiological signals and audio data during qualitative interviews and for releasing this data for research purposes.

[5] `https://irb.iiitd.edu.in/`

To protect their identities, participants were pseudonymized using numerical identifiers. The audio data was transcribed, manually checked for any identifying information, and included as textual descriptions devoid of sensitive content. Participants received merchandise goodies worth 5.39 USD for their participation. The dataset is available for download under a CC BY-NC-SA license for non-commercial research purposes on our website. The codes for data cleaning, feature extraction, and classification are open-source and can be accessed. The open-source code can be accessed through the following repository. Our dataset does not have any direct negative impact on society and is designed and made open source, keeping users' privacy in mind.

4.7 Summary

Overall, the results from the EEVR dataset highlight both the limitations of current approaches in collecting emotional self-reports. Although self-reports remain one of the most reliable methods for capturing emotional experience, they are often constrained by objective rating scales that oversimplify emotions, which are inherently continuous, dynamic, and context-dependent. Such simplifications can lead to the loss of important experiential details. In this chapter, I emphasized the value of richer self-report modalities, particularly text- or audio-based reflections, which provide deeper contextual information and align more closely with physiological signals than conventional scale-based annotations. The proposed experimental protocol is readily reproducible in laboratory settings and can be extended to everyday data collection via conversational interfaces such as chatbots or audio-based companion applications (a prototype discussed in Chapter 6). The EEVR dataset thus offers a valuable resource for researchers working on physiological emotion recognition, enabling the development and benchmarking of new machine learning models alongside existing baselines. By pairing physiological signals with descriptive emotional narratives, the dataset supports more nuanced modeling of emotional experience. It can also enable investigation of how emotional responses relate to broader participant characteristics, in-

cluding psychological well-being, personality traits, and physiological variability. Finally, the availability of open-source baseline code and accessible documentation for datasets promotes reproducibility and encourages further research building on this resource.

Chapter 5

Emotions Aren't Just Numbers: Humanizing Emotion Data Pipelines

In this chapter, building on Chapter 1, which highlighted the importance of participant-centered design in emotion data collection protocols, I present two complementary studies organized into Part 1 and Part 2. Part 1 examines participant perspectives within laboratory-based data collection settings. Participants were exposed to a structured lab emotion-elicitation protocol, followed by qualitative interviews aimed at understanding their experiences, perceptions of the process, challenges encountered during data collection, and the impact of these challenges on data quality. Part 2 shifts the focus to everyday contexts, exploring participants' perspectives on sharing emotion-related data in naturalistic settings. This part primarily adopts a qualitative approach and includes members of the public and mental health professionals to better understand the practical, ethical, and experiential challenges associated with real-world emotion data collection. Part 2 further outlines practical guidelines for designing participant-centered approaches to emotion data collection in real-world settings.

Part I: Participants' Perspectives on Emotion Data Collection in the Lab Settings

Emotion recognition using physiological signal data relies heavily on human participants as significant stakeholders for data collection and labeling. Thus, the quality of physiological signal-based emotion data is also dependent on human participants. Moreover, other essential data quality parameters include using high-resolution wearable sensors with minimal noise or artifacts for physiological signal data collection, employing reliable methods for labeling emotional states, collecting data from diverse groups of participants, using realistic elicitation methods to collect emotional data that accurately reflect genuine emotional responses, and

ensuring the data is complete with information on participants and their characteristics (such as age, gender, personality and health history) [168, 36, 5, 169, 170]. As discussed in Chapter 3, prior datasets have been collected across three settings: i) *Laboratory Settings*, ii) *Field with Constraints Settings*, and iii) *Field Settings*. However, each experimental protocol has its own set of limitations [36], which impact overall performance.

Moreover, prior research has largely emphasized design factors such as stimulus type, stimulus content, and data-collection settings, while comparatively little attention has been paid to how participant characteristics themselves influence emotion elicitation and, consequently, the quality and reliability of the resulting data [53, 39, 155, 37]. Building on this, the emotion recognition community has recently begun to acknowledge participants as active contributors rather than passive data sources; however, there remains a need to bridge the research gap between data needs and participants' needs. Part one of this chapter focuses on exploring the following research questions:

- **RQ1:** What participant-specific factors can impact emotion elicitation, annotation, and data quality?
- **RQ2:** What are the participants' perspectives on the physiological emotion data collection process, including the labeling method, stimuli, and the experimental setup (VR)?

We conducted a lab-based physiological emotion data collection experiment to address our research questions, followed by semi-structured qualitative interviews with 37 participants (21 males, 15 females, 1 undisclosed). Our choice of a lab setting aligns with the prevalent use of such settings in prior research on physiological emotion data collection [53, 39, 155, 37]. We used Virtual Reality (VR) short 360° videos as stimuli in the lab. The selection of VR-based stimuli stems from their demonstrated ability to evoke a higher emotional response compared to 2D stimuli [171, 172]. The following sections delve into related literature, our study design, and our findings. Next, we discuss the role of participants,

focusing on how their perceptions and interpretations shape their emotional responses and annotations. We then provide recommendations for practitioners and researchers in the HCI and AI communities to work towards collaborative practices for collecting physiological emotion data while keeping key stakeholders in mind.

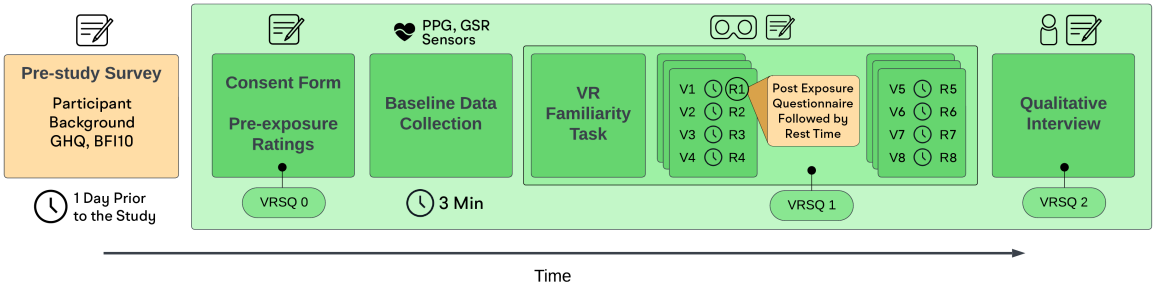


Figure 5.1: The figure depicts the data collection procedure used in this paper in its execution order. In the diagram, the following abbreviations correspond to the respective components: GHQ (General Health Questionnaire), BFI10 (Big Five Inventory-10), VRSQ (Virtual Reality Sickness Questionnaire), PPG (Photoplethysmography), EDA (Galvanic Skin Response), V (video or Stimulus), and R (Rest). The Pre and Post-Exposure Ratings include the Positive-Negative Affect Scale (PANAS) and the SAM Scales.

5.1 Methodology

To investigate our RQs, we conducted the data collection experiment in virtual reality using VR 360° short videos. The Institute Review Board (IRB) approved the study for ethical considerations. At the end of the study, participants received merchandise goodies worth $4 as a token of participation. The following sections describe our methods.

5.1.1 Participant Selection

Our study comprised 37 healthy participants (21 males, 15 females, 1 undisclosed) aged 18-33 (M=23.1, SD=4.02). Recruitment was conducted through institute-wide email calls and promotions within social media circles. Exclusion criteria encompassed individuals with experience or a history of heart issues, heart arrhythmia, high blood pressure, medical conditions affecting equilibrium, visual or auditory impairments, neurological ailments,

cognitive challenges, psychological issues, or diagnosed depression [155]. Additionally, participants with low proficiency in the English language were not included in the study to avoid the impact of language. Our participant demographic and data summary from our pre-study survey are presented in Table 5.1.

Category	Details and Count
Total Participants	**37**
Gender	Female **(15)**, Male **(21)**, Prefer not to say **(1)**
Playlist	Playlist 1 **(6M, 4F)**, Playlist 2 **(4M, 4F)**, Playlist 3 **(5M, 4F)**, Playlist 4 **(6M, 4F)**
Age	Range **18–33**, Mean **23.1**, SD **4.02**
Education	Senior High School **(4)**, Bachelor's Degree **(24)**, Master's Degree **(8)**, Doctorate **(3)**
Awareness about VR	Yes **(27)**, No **(10)**
Usage of VR	Never **(17)**, Rarely **(14)**, Sometimes **(4)**, Often **(1)**, Very often **(1)**
Average Screen Time	Less than 2h **(2)**, 2–4h **(5)**, 4–6h **(5)**, 6–8h **(8)**, 8–10h **(11)**, >10h **(6)**
General Health Assessment	Healthy **(20)**, Distressed **(17)**
Personality Characteristics	Agreeableness (Low **3**, High **34**); Extraversion (Low **5**, High **32**); Conscientiousness (Low **9**, High **28**); Neuroticism (Low **13**, High **24**); Openness (Low **5**, High **32**)

Table 5.1: Demographic information of the study participants. See Section B.4.2 for details on playlists, health assessment, and personality characteristics.

5.1.2 Stimulus Selection and Data Collection

We employed a between-subjects study design (Figure 5.3b) with two independent variables. The first, *VideoSet*, refers to the two sets of (N=8) VR 360° videos presented to participants. The second variable, *VideoOrder*, pertains to the sequence in which a video set is presented to participants. For this experiment, a total of 16 videos were utilized. The videos were selected from a publicly available 360° VR dataset [153], which was also used in prior work. To curate the video subset, we applied a heuristic protocol, selecting four videos from each category of the circumplex model [72]. The heuristic involved choosing videos with

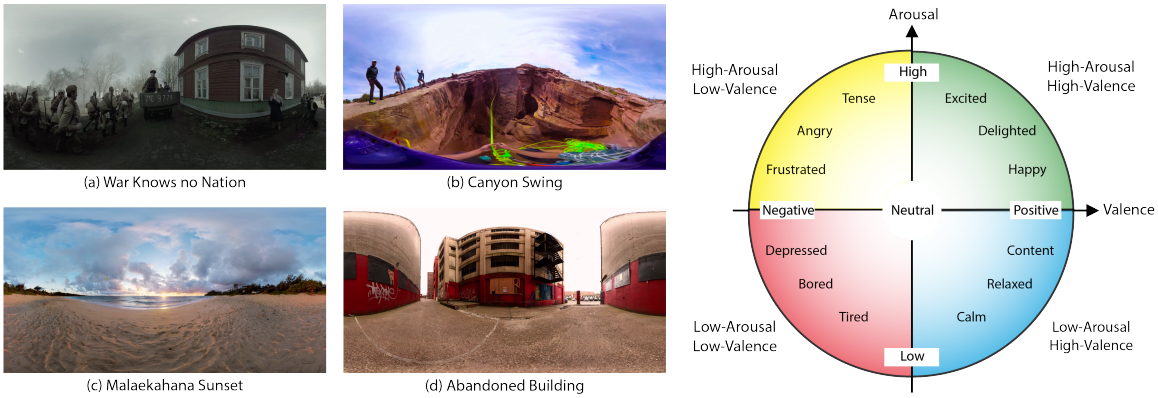
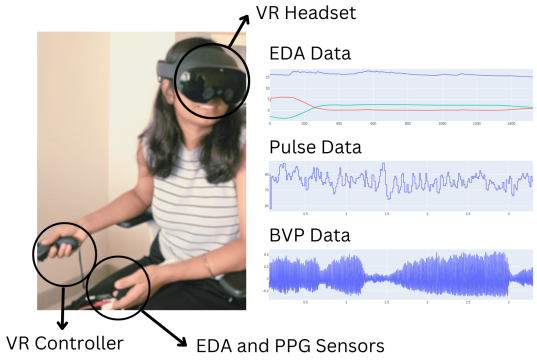


Figure 5.2: The figure shows stills taken from the 360° videos capturing different environments shown to the participants as a part of the experiment methodology. The stills capture different valence-arousal combinations such as (a) Low-Valence-High-Arousal (LVHA), (b) High-Valence-High-Arousal (HVHA), (c) High-Valence-low-arousal (HVLA) and (d) Low-Valence-Low-Arousal (LVLA), the database is publicly available [153].

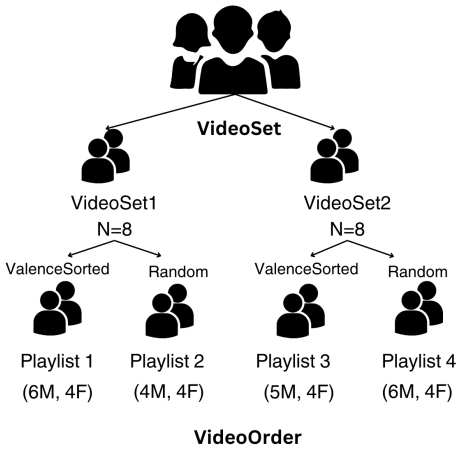
maximum distance from the origin to enhance coverage and diversity within the subset. The 16 videos were then divided into two subgroups of N=8 videos each, considering factors such as pilot feedback, total experiment time, and VR exposure to prevent participant fatigue or motion sickness. The subgroup selection involved arranging videos based on their valence ratings (which represent the degree of pleasantness or unpleasantness associated with an emotional state). This was done to determine if a specific sequence of emotions (from unpleasant to highly pleasant) impacts the overall emotional response of the participant, compared to a random presentation of stimuli that is independent of the valence rating. We chose valence-based ordering because it directly reflects the emotional response triggered by the stimulus and is easier for participants to identify and distinguish. In contrast, arousal is more subjective and harder for participants to differentiate [173, 174, 175]. To simplify the experiment for participants, we prioritized valence-based ordering. Subsequently, alternate videos were paired from each quadrant to create two VideoSet. This step aimed to establish playlist normalization, ensuring a balanced experimental setting. The selected videos were downloaded from database[1] using youtube-dl[2] tool in the equirectangular panoramic format

[1]https://stanfordvr.com/360-video-database/
[2]https://github.com/ytdl-org/youtube-dl

with a resolution of 3840 x 2160 pixels. All the videos were edited to fit within a 3-minute timeframe, as per prior work [155]. To investigate the impact of video order on emotional responses, we organized the videos into two distinct orders: Valence Sorted Order, where videos were arranged based on valence ratings within a VideoSet, and Random Order, where videos were arranged randomly, irrespective of their ratings. After applying these orders, we created four playlists. The sorting technique employed for the study involved arranging videos from low negative valence to high positive valence, facilitating the determination of the emotional properties as positive or negative. The four playlists (see Table 5.2) are as follows- *Playlist1: VideoSet1 - Random Order*, *Playlist2: VideoSet1 - Valence Sorted Order*, *Playlist3: VideoSet2 - Random Order*, and *Playlist4: VideoSet2 - Valence Sorted Order*. Participants were allocated playlists in a gender-balanced manner through random assignment (see Figure 5.3b). The videos (see Figure 5.2) were presented to participants in a custom Virtual environment with separate scenes using Unity Engine.



(a) The experiment setup and illustration of collected physiological signal data



(b) Between subject Methodology

Figure 5.3: Study Design

Experiment Procedure: Our experiment, as depicted in Figure 5.1, lasted approximately 1 hour and 15 minutes. Prior to data collection, an email was rolled out inviting individuals to participate in the study. Individuals who opted in for the study received an email confirming their availability. Twenty-four hours before the experiment, participants were given a pre-

study survey to collect crucial participant information. Since the nature of participation was opt-in, this was considered implicit consent for the pre-study questionnaire. This questionnaire covered - *Background Information*: Questions regarding gender, age, and familiarity with VR technology on a 5-point Likert scale. Open-ended questions were included to understand any prior exposure to VR. *General Health Assessment*: To assess psychological well-being over the past week, we utilized the twelve-item General Health questionnaire (GHQ-12) [176]. This non-medical screening tool helped gauge the emotional state of otherwise healthy participants with no diagnosed mental health conditions in the week prior to data collection. *Personality Assessment*: Participants' personality types were assessed using the Big Five Inventory-10 (BFI-10) [177]. The inclusion of GHQ-12 and BFI-10 in the pre-study questionnaire was influenced by previous data collection research [39, 53, 178]. Other participants' parameter characteristics, like average screen time, were collected to understand participants' consumption of digital media and their technology awareness.

On the data collection day, participants were briefed about the study without revealing its objectives. Subsequently, they read the privacy policy and the associated risks and signed the consent form. Following the briefing, a consent form was presented to the participants. Consenting participants were directed to complete a *pre-exposure form*. This form included questions about current emotional states using standard PANAS [74], SAM [179] scales, and Virtual Reality Sickness Questionnaire (VRSQ) [152] to assess initial fatigue and motion sickness symptoms before their VR exposure. The PANAS scale gathered positive and negative affect readings on a 5-point scale, covering ten positive and ten negative emotions [74]. The SAM scale was utilized for dimensional ratings of Arousal (refers to the intensity associated with emotion), Valence (positivity to the negativity of emotion), and Dominance (degree of control over emotion). Participants were encouraged to seek clarification from the experimenter if they needed assistance understanding the form. After completing the pre-exposure form, Photoplethysmography (PPG) and Electrodermal Activity (EDA)

sensors were attached to the participant's non-dominant hand fingers. Subsequently, 3 minutes of baseline data was collected, during which participants were instructed to sit and relax. Studies have shown the benefits of a shorter baseline collection period [180]. Consequently, we adopted a 3-minute period for baseline data collection, aligning it with our stimulus duration of around 3 minutes to ensure a balanced amount of physiological data collected. The choice of PPG and EDA sensors was based on their widespread use in emotion recognition. Within the VR environment, participants initially sat in a waiting room to acclimate to the technology, exploring the surroundings by looking around. Following this, participants transitioned to familiarizing themselves with the VR controller. Engaging in a simple game, they used the controller to pick up and throw a ball into a box within the VR room. After completing the VR familiarity task, participants were instructed to choose the assigned playlist and commence watching the videos in that playlist. Following each video, participants completed the *post-exposure form* (same as pre-exposure form) to document their emotions during the viewing. Once the form was completed, participants rested before proceeding to the following video.

Playlist Number	**Reference Video Number: Name of Video (in order)**
Playlist 1	P1V1: The Displaced; P1V2: Happyland 360; P1V3: Jailbreak 360; P1V4: War Knows No Nation; P1V5: Canyon Swing; P1V6: Redwoods Walk Among Giants; P1V7: Speed Flying; P1V8: Instant Caribbean Vacation
Playlist 2	P2V1: The Nepal Earthquake Aftermath; P2V2: Zombie Apocalypse Horror; P2V3: Abandoned Building; P2V4: Kidnapped; P2V5: Mega Coaster; P2V6: Malaekahana Sunrise; P2V7: Puppies Host SourceFed for a Day; P2V8: Great Ocean Road
Playlist 3	P3V1: War Knows No Nation; P3V2: Redwoods Walk Among Giants; P3V3: Happyland 360; P3V4: Speed Flying; P3V5: Instant Caribbean Vacation; P3V6: Jailbreak 360; P3V7: The Displaced; P3V8: Canyon Swing
Playlist 4	P4V1: Kidnapped; P4V2: Malaekahana Sunrise; P4V3: Zombie Apocalypse Horror; P4V4: Puppies Host SourceFed for a Day; P4V5: Great Ocean Road; P4V6: Abandoned Building; P4V7: The Nepal Earthquake Aftermath; P4V8: Mega Coaster

Table 5.2: List of videos in each playlist in the order presented to participants.

5.1.3 Experiment Setup

The study was conducted in the institute's research lab. Our experiment setup is depicted in figure 5.3a.

VR and Questionnaire Setup: Meta Quest Pro was used to show the VEs. The headset has 2 x LCD panels with 1800 x 1920 pixels per eye, a refresh rate of 90Hz, 106º Horizontal × 96º Vertical Field of view. Additionally, it incorporates eye relief adjustment, lens spacing features, and Spatial audio support. ASUS TUF laptop was used to run the project with 11th Gen Intel core and NVIDIA GeForce RTX 3050. The OpenXR plugin integrated the Meta Quest pro headset with the Unity application. OpenXR plugin helped us with hand gestures and controls for interacting with the Application User Interface. An iPad Pro tablet was used to complete the pre and post-exposure forms.

Physiological Measure Apparatus: A 4-channel Biopac MP36 [181] was used for continuous collection of PPG (SS4LA Hardware module) and EDA (SS57LA Hardware module) data using separate channels. BSL 4 software from biopac was used for data acquisition. The EDA sensor was attached to the index and middle fingers [155] of the participants' non-dominant hand, utilizing EL507 Electrodes for collecting users' skin electrical conductance. The PPG sensor was attached to the participant's non-dominant hand ring finger. Before attaching the sensors, Isotonic Gel was applied, and the electrodes were firmly fastened to ensure minimal noise in the collected data.

5.1.4 Interviews

We conducted a concluding semi-structured interview with the 37 participants post-exposure to the 360° VR stimulus. The interviews were conducted in English and Hindi as required after obtaining consent to record audio. During interviews, our objective was to grasp participants' emotional responses, comprehend the reasoning behind these feelings, gain insights into their understanding of scales employed for emotion annotations, and assess the impact of the experimental design and setup on their emotions. The interviews began with

an opening question, *"How was your overall experience?"* We then assessed participants' emotional engagement to study RQ1 by showing brief segments of each video on a computer screen to aid recollection. Following this, we explored the dominant emotion in each video, prompting participants to articulate their emotional responses and reasons behind annotation by referring to their subjective ratings. To investigate RQ2, we asked questions such as: *"How easy was it for you to transition emotions between videos? Did breaks between videos aid emotional transitions? Did preceding videos influence your emotions towards subsequent ones? What prevented you from feeling the expected emotion tied to the stimuli? Did you face any challenges filling out the pre- and post-exposure questionnaires?"* Subsequently, to understand the impact of VR 360° videos as an elicitation medium, we focused on various aspects of viewing experiences in VR. This encompassed elements such as camera angles, VR content stitching quality, different perspectives, overall video quality, and impact of real-life camera footage in comparison to computer-generated imagery (CGI) (Real-life footage such as "Happyland," filmed in a real-life slum area in the Philippines whereas computer-generated videos featured scenes created using visual effects and computer-generated-imagery, such as "War Knows No Nation," which used CGI to depict soldiers from various countries worldwide). We also delved into the auditory experience in VR, considering background music (such as calm music, energetic music, lyrical songs, and other types of background audio like birds chirping, sea waves sound, etc., as per the stimulus content), audio clarity, and the spatial orientation of audio cues. The linguistic experience of stimuli was examined, including factors like the language of narration and the presence or absence of subtitles. Additionally, we considered the influence of the lab setting and participants' awareness of it on their experiences. Our interviews followed a semi-structured approach, allowing us to adapt the questions based on participants' feedback.

5.2 Data Analysis

For analyzing our interview data, we first transcribed the recorded audio data into English using the Google Cloud speech-to-text API [3] as an initial draft. This draft was then checked for transcription and spelling errors by the first and second authors, who manually matched the text to the audio recordings to correct any mistakes (such as if heavily accented words were identified correctly or if there were any transcription mistakes in general) made by the Google API. We also manually added speaker identifiers (participant and interviewer) and identified text segments corresponding to particular videos. We have adopted realist epistemology [182] for analyzing our qualitative data, wherein we have reported the emotional experiences and perceptions of the participants while experiencing the stimulus and labeling the emotions. Subsequently, we employed inductive thematic analysis [183] on the refined transcripts from the previous steps. In the initial phase, the first two authors performed open coding on the interview transcripts line-by-line. Codes were reviewed and aligned to ensure consistency, with all authors collectively involved in the analysis process. Sample codes from our open coding included *"Novelty of VR," "connection with past," "future aspiration,"* and *"expectations from VR"*, etc. These open codes were then clustered based on emerging patterns over multiple iterations to derive axial codes, such as *"drawbacks of VR," "participant's perspective,"* and *"relationship with stimuli"*, etc. Finally, we used selective coding to refine the axial codes and identify the final themes that guided the structuring of our findings. Throughout this process, the research team utilized tools like Miro-board [4] and Google Sheets for conceptualizing and theme construction. We performed statistical tests on both EDA and PPG data for our quantitative data analysis. For EDA, we extracted the mean EDA values. For PPG, we extracted HRV RMSSD (Root Mean Square of Successive Differences) from the signal data collected during baseline and stimulus periods. These features were selected as indicators of parasympathetic

[3] https://cloud.google.com/speech-to-text
[4] https://miro.com/

nervous system activity [37]. Initially, we tested the data for normality using Q-Q plots and the Shapiro-Wilk test. Since the data did not follow a normal distribution, we applied a two-tailed Wilcoxon rank-sum test to compare the distributions of two independent groups for our analysis and the Kruskal-Wallis test for more than two groups.

Category	Key Insights
Participant perspective	- Self perception - Prior knowledge and experiences with stimulus - Logically reasonable and emotionally relevant stimulus - Motivation towards experiment
Experiment Design	- Order and length of stimulus - Exposure duration according to the emotion of stimulus - Interval between different stimuli - Choice of self-annotation method
Virtual Reality	- High immersion and presence - Need for Interaction within stimulus - Presence of human-like elements - Quality of video and audio elements

Table 5.3: Summary of the Findings

5.3 Findings

In this section, we present the implications and influence of participant perception, experiment design, and experiment setup choices on a participant's emotional response and annotation. The summary of our findings is presented in Table5.3.

5.3.1 Unveiling the Influence of Participant Perception

The Influence of Participant's Self-perception

Our interviews revealed that participants *self-perception* played an important role in how they responded to a stimulus. On various occasions, participants pointed out that their typical behaviour patterns and personality traits in daily life have shaped their emotional response to a stimulus shown in the experiment. We found that participants who self-perceived

themselves as emotionally stable or in control of their emotions were less responsive towards the stimulus than participants who did not have a strong self-perception of being emotionally stable. A participant shared:

> *"Not much change in emotion I would say. Not sure but I feel I have control over my emotions to an extent, I think so I am normally neutral in most of the situations."* **[P39, Playlist3]**

Further, it was also observed that participants with strong self-perceptions of being emotionally stable consciously chose not to respond to a negative stimulus, as explained by a participant **[P33, Playlist 3**] *"I know this is sad, but I won't cry over it"*. Similarly, individuals who identified as nature or pet enthusiasts displayed heightened emotional responses to stimuli involving natural settings or animals. For example, those who preferred beaches appeared more serene when viewing videos of beach scenes. In contrast, participants who were neutral towards puppies demonstrated a subdued reaction, whereas those with a fondness for puppies exhibited a more pronounced emotional response. Overall, participants' self-perceptions have shaped their responses to a stimulus and their annotations irrespective of the targeted stimulus emotion.

The Influence of Perception towards the Content of Stimulus

All participants mentioned that their viewpoint toward the content of the stimulus was a major reason behind eliciting a response. Past experiences were frequently linked to the content displayed in the stimulus, resulting in increased engagement among participants. Recalling familiar life scenarios made participants feel more emotionally connected to the stimuli. For example, participants mentioned instances such as trekking to locations resembling the stimulus and experiencing earthquakes in the past. One of the participants explained:

> *"Like I was sad...because one of my friend is in army, so I felt like he is going on*

war...yeah so one part where the woman is outside the train and the army was like going- from the window she was waving" **[P29, Playlist1]**

Similarly, prior knowledge about the content shown in the stimulus also contributed to the emotional response. Participants reported that prior knowledge has impacted their emotional engagement; for instance, prior knowledge about World War II and that it had already happened in the past has led to emotional detachment towards the content. A participant [**P14, Playlist 2**] explained that *"No, because I am aware that it happened [World War 2], so it did not feel like it's happening right now"*. According to our participants, cultural background also had a subtle impact on how much they related to content. A participant felt a stronger connection to the people of Nepal, as depicted in *The Nepal Earthquake Aftermath* because they live in a geographic region closer to Nepal, while a sense of detachment was reported by participants while watching the stimulus with the story of people in Ukraine. Personal beliefs are another factor, as mentioned in our interviews, that played a role in how participants perceived and annotated a stimulus. For instance, people who value social causes or believe in inequality of resources were influenced more by stimuli like *Happyland 360* than those who do not have strong opinions about poverty. Aspirations were another common factor often associated with the stimulus. In case of stimulus wherein adventure sports or new countries were shown *Canyon Swing, Speed Flying, Great Ocean Road*, participants reported being excited as they aspire to indulge in the adventure sport or travel to that country. Next, participants revealed that cognitive processes like logical reasoning and attention toward stimulus have shaped their response to a stimulus. Participant [**P23, Playlist 3**] explained, *"One thing that I could not understand logically was that I felt that the rope is so far, how come I am hanging in the middle now. I didn't understand that thing, but the rest was fine"*. We also observed logical reasoning, like an earthquake is a negative situation, and thus, the annotations should also be negative in our interviews.

The Influence of Perception towards Experiment

Through our interactions and observations, we found that participants' perception of the data collection experiment also shaped their emotional response to a stimulus. We observed in our interviews that some participants were more interested in experiencing VR technology than others, and thus, their emotional responses were mostly excitement or boredom, depending on the stimulus. For instance, [**P3, Playlist 3**] responded to our question on *How was your overall experience of the study? - It was nice. First, the videos were a little boring, but it had some interesting ones*. Another participant shared her excitement:

> *"The overall experience was pretty cool, the VR headset seemed perfectly fine and it was really immersive... I was on the ship, in the VR headset, it was very aesthetic and actually it made me more enthusiastic and excited to see virtual objects that seem to be real but aren't, so it was really amazing"* **[P12, Playlist1]**

Participants mentioned getting confused in some particular stimulus that didn't provide any context on what was to be shown in the stimulus for example, stimulus like *Abandoned Building* just started with scenes from an abandoned area with no instructions to participants on what they are expected to do or feel. This lack of clarity on what is expected within a stimulus has created confusion and also made participants ponder on what was the targeted emotion of the given stimulus. They reported that this confusion has propagated to their self-reports as they could not understand their emotions clearly. Such a situation is often raised in videos where participants were expected to sit calmly and listen to music, while participants were expecting some cognitive load or some storytelling due to their experiences in the previous stimulus. [**P31, Playlist 1**] stated on that *"I was confused in this one because nothing was happening, yeah. But then I realized that nothing was gonna happen... I was just looking around the buildings and stuff"* after watching *Abandoned building*. The time of the experiment and mood before the experiment were other defining factors for the

perception of the experiment. A participant reported feeling tired and less attentive to the stimulus due to their workload on that particular day and time before the experiment, which made them less interested in the experiment.

The Influence of Stimulus on Participant's Self-report and Physiological Signals

To study whether there was a change in annotations as per stimulus, we performed statistical analysis on arousal and valence scores. Our test for **change in arousal in high vs. low arousal stimulus**, showed **significant difference** (Test Statistic = 2.691, p-value = 0.007**), and for **change in valence in high vs. low valence stimulus** showed **significant difference** (Test Statistic = 7.697, p-value = 0.000**), suggesting the overall annotations have changed according to the change in stimulus. Furthermore, to study the impact of change in stimulus on the physiological signal data of our participants, we have performed statistical tests on our participant's data. On testing for changes in HRV RMSSD and EDA mean across different stimulus categories, we found no significant differences. Specifically, we observed no significant difference between HRV RMSSD and EDA mean values of High Valence vs. Low Valence stimuli and High Arousal vs. Low Arousal stimuli using a two-tailed Wilcoxon rank sum test. Additionally, when comparing the circumplex categories (LVLA, LVHA, HVLA, HVHA) for stimuli, no significant difference was detected in physiological signal values using the Kruskal-Wallis test. Suggesting that there were no significant changes in the HRV RMSSD and EDA mean features of the participant's data when stimulus categories were changed.

The Influence of Participant's Characteristics on Physiological Signals

For analyzing GHQ data, we calculated the GHQ score by summing the responses to the 12 questions, resulting in a total score ranging from 0 to 36. A lower score indicates better mental health and less psychological distress, while a higher score suggests elevated levels of psychological distress over the past week. In our data, 17 out of 37 participants have a

Category	Measure	Significance	p-value	Test-statistics
VideoSet1 vs. VideoSet2	Arousal **(H14A)**	**S**	0.000**	6.57
	Valence **(H14B)**	NS	0.372	-0.893
	EDA Mean **(H14E)**	**S**	0.003**	3.007
	HRV RMSSD **(H14F)**	NS	0.945	0.069
Playlist 1 vs. Playlist 3	Arousal **(H15A**)	NS	0.26	-1.126
	Valence **(H15B)**	NS	0.943	0.071
	EDA Mean **(H15E)**	NS	0.358	0.919
	HRV RMSSD **(H15F)**	NS	0.738	-0.334
Playlist 2 vs. Playlist 4	Arousal **(H16A)**	NS	0.359	0.917
	Valence **(H16B)**	NS	0.122	-1.547
	EDA Mean **(H16E)**	NS	0.5	0.674
	HRV RMSSD **(H16F)**	**S**	0.021*	2.306

Table 5.4: Summary of results of statistical tests for analyzing the impact of Video Set and order on both Subjective measures and Physiological data. In the table NS: Non-significant, S: Significant, **: less than alpha 0.01, *: less than alpha 0.05

high GHQ score, which suggests that these participants have elevated psychological distress in the past week (in real life outside our study environment). To test the impact of GHQ scores on physiological data (collected during baseline+stimulus), we conducted a two-tailed Wilcoxon rank-sum test to compare the EDA mean and HRV RMSSD values between participants with High GHQ scores vs. Low GHQ scores. The results showed no significant differences in EDA mean or HRV RMSSD between the two groups. We also conducted individual tests on the extracted values for the baseline and stimulus data separately but found no significant differences in these subsets either. This suggested that GHQ scores don't have any impact on the physiological data on our subset of participants. We conducted a series of tests to analyze the impact of different personality characteristics on physiological data. First, we calculated personality scores for each category: Agreeableness, Extraversion, Conscientiousness, Neuroticism, and Openness (more details in Table 5.1). We then applied a two-tailed Wilcoxon test on the combined data (baseline+stimulus), baseline-only, and stimulus-only data. For participants with High vs. Low Agreeableness, no significant difference in physiological data was found. The same result was observed for Extraversion. However, for Conscientiousness, we found a **significant difference** in HRV RMSSD data

(baseline+stimulus) between High and Low Conscientiousness participants (Test Statistic = -3.263, p-value = 0.001**) but no significant difference in EDA mean data. For Neuroticism, a **significant difference** was observed in EDA mean data (baseline+stimulus) between High and Low Neuroticism participants (Test Statistic = 3.818, p-value = 0.000**). Additionally, a **significant difference** in HRV RMSSD was found in the stimulus-only data for Neuroticism (Test Statistic = -2.120, p-value = 0.034*). Lastly, for Openness, we observed a **significant difference** in HRV RMSSD data (stimulus+baseline) between High and Low Openness participants (Test Statistic = -3.865, p-value=0.000*), while the EDA mean was non-significant.

5.3.2 Deciphering the Experiment Design's Impact and Significance

Order, Length and Choice of Stimulus matters

In our interviews with participants, we noticed that experiment design played an important role in emotion data collection. Design choices like the order in which the stimulus would be shown to participants and the length of the stimulus mattered to our participants. In our interviews, participants reported anticipating something to happen in a calming stimulus because the previous stimulus they had seen involved a lot of actions (e.g., observed in playlist 3 - P3V1 (War knowns No Nation) followed by P3V2 (Redwood Walks among Giants)). This suggested an impact of order in the form of anticipation and expectation from the content of the stimulus, impacting an otherwise relaxed emotional response. A participant stated:

> *"At first I was like, kind of alert, like in that question [PANAS] because I thought something was going to happen, because after the first video, I thought in the second video something is gonna happen like some jump scare or someones gonna come or something will happen. But then I realised it is a calming situation."* **[P12, Playlist4]**

The stimulus length was another important factor influencing the participant's emotional response. Participant [**P9, Playlist 2**] stated *"Yeah, it was calm...But then, I think it went on for too long...So, this got boring"*. Instances were noted where participants expressed boredom during lengthy stimuli, dissatisfaction with insufficient exposure, and an inability to feel any emotion during short exposures. Notably, participants found it easier to connect with brief positive stimuli, whereas more time was deemed necessary for negative stimuli. One participant articulated that the stimulus duration was too brief for her to establish an emotional connection. This highlights the nuanced relationship between stimulus length and participants' emotional responses. Additionally, we performed statistical tests to analyze the impact of VideoSets and the order of stimulus (playlists) on participants' physiological data and their annotations (valence and arousal). We found a **significant difference** in the Arousal score and EDA mean of participants in VideoSet1 vs. VideoSet2. Suggesting that VideoSets have an impact on arousal scores and EDA data. As for playlists, we found a **significant difference** in HRV RMSSD for participants in Playlist 2 vs. Playlist 4. This may or may not be due to playlist order since HRV RMSSD can also be influenced by the participant pool in the playlists. More details are provided in Table 5.4.

Challenges in Annotating the Emotions

Besides the order, length, and choice of stimulus, the annotation method was another crucial element in the data collection procedure. In our interviews, participants mentioned that they faced issues annotating their emotions on the SAM scale when they faced mixed emotions like fear and excitement [In Speed Flying], while PANAS was better for annotating mixed emotions. Another common comment by participants was that they found it hard to quantify their emotions in numbers and preferred interviews to express their emotions as stated by [**P9, Playlist 2**] *"this is really hard to answer, yeah, you can articulate it but actually quantifying it is a very difficult task to do so, overall you can say, okay, four or five or three [referring to SAM scale], I was a bit exciting, more exciting, but how much that is really*

difficult". Most participants noticed that few emotions were part of the PANAS scale, like, *hostile or guilty*, but they never felt those emotions throughout the experiment, questioning the need for including them in the questionnaire. A participant compared the two scales and explained:

> *"So I think partly name of the emotion [PANAS], because I was able to like correlate with words more than quantify a metric to my emotion, and like with the diagrams [SAM Scale] and all that, how am I feeling."* **[P25, Playlist2]**

Participants also pointed out a need for randomness in the self-report questionnaire as they were repetitive after each stimulus, giving them a chance to fill it out without much thought. Gaps or rest periods outside the VR environment between the consecutive stimuli were reported to help transition from one emotion to another, and thus, they preferred filling their annotations outside the VR environment.

5.3.3 A participant's (Virtual) Reality - Implications of Experiment Setup

Suitability of VR as an Elicitation Medium

Through our interviews, we observed that there were certain aspects of VR as a medium that impacted the overall experience of the participants, both in a positive and a limiting manner. Many participants mentioned that they could feel being present in the setting where the stimulus was set. This heightened sense of presence not only improved participants' overall experience but also contributed to a more profound emotional engagement with the stimulus. For example, as reported by [**P2, Playlist 1**], *"I just wanted to go back in time and see how people were doing things... I can't go there, but I was able to feel all that"* [referring to soldiers in war]. Participants also mentioned that when they immersed themselves in the VR content, they became somewhat unaware of their real-life surroundings and reported being more engaged with the content. A participant explained:

> *"And I'm a huge roller coaster fan in general, so this was the first time I actually*

felt like I was in it, like I could like physiologically feel like okay I'm taking a turn- yeah my heartbeat was... and I don't know there's this one feeling, like something happening in my stomach but in a good way, like in an excited way"

[P18, Playlist4]

Many participants reported feeling the urge to interact with the stimulus, but since it was a video-based stimulus, they could not do so. Participants also reported feeling that their actions were restricted because it was a passive stimulus. Some participants mentioned that in addition to the visual stimulation, they would also want some physical sensation that compliments the presented stimulus to feel more present in the stimulus. For example, [**P27, Playlist 4**] mentioned, *"I felt like I was there, I could- there was a person sitting behind as well, there was some wind, but I could not feel the wind, like I could listen to the wind"* , since they were immersed in the video but their experience got hindered due to a lack of the physical stimulation that comes with the wind brushing past. Some participants had biases toward VR, which hindered their overall experience. As [**P3, Playlist 3**] mentioned, VR seemed to be *"a little on the fake side"* to them. VR was a first-time experience for many participants, and they reported having elevated expectations from VR, which was influenced by the amount of knowledge they had about the medium. [**P17, Playlist 3**] mentioned that they were curious as to how much can one experience within a VR setting and were wondering, *"somebody can be sitting in their chair, in their house, and could be feeling like they are in the woods surrounded by such huge trees"*. Some participants also mentioned the opposite, that due to their past experiences with VR, they were pretty familiar with the platform and had realistic expectations from the various stimuli presented to them. A participant explained:

"VR usually means there is a purpose to the VR content that is created, and here I felt like it was just sort of being around nature. That was the only context, and for that I think I should be there, like I expected smell, I expected, like, the environment, the atmosphere, maybe some animal sounds, but those things

weren't there [in VR stimuli]" **[P21, Playlist1]**

Constraints of Stimulus in VR

Besides the overall experience that the participants had with VR as a medium for stimulation, our interview questions specific to each video being presented gave us insights as to what elements of the stimulus were responsible for the experience that the viewer had. Most participants mentioned that good video quality added to their viewing experience. In addition, we observed that many participants agreed that despite the poor video quality for certain stimuli, they were willing to ignore it if they found the stimulus engaging enough. [**P3, Playlist 3**] mentioned, *"I noticed the poor quality, but I liked it anyway. Video quality was not that important for me; it was the content of the videos"*. Many participants did not like the presence of quick camera cuts and fast-paced scene changes since it did not allow them to explore their surroundings. [**P19, Playlist 3**] mentioned, *"this felt like immersive because I was given time to like, explore around, [and see] what's happening"*. In addition to fast cuts, participants felt that unrealistic angles and non-human camera point-of-view (POVs) created a sense of detachment from the presented stimulus. POVs that were more human-like were appreciated. For example, in the stimuli *Happyland 360*, [**P39, Playlist 3**] referred to the eye-level placement of the camera and mentioned, *"I actually felt connected because most of the children- they were looking into the camera so it felt like they are looking at me, and we have a direct eye contact"*. Stimuli with video stitching issues and *black holes* at the bottom of the recording hindered the participants' experience. For instance, when asked *"The camera angle was such that you were the head of the person, did that create any impact?"* for the stimulus *Kidnapped*, [**P4, Playlist 4**] answered, *"No, because I looked down and I could see like a black hole there"*. The presence and absence of human elements within the video stimulus was also a factor, and participants reported that human interactions within the stimulus increased their sense of being present in the stimulus scenario, as explained by a participant:

"It was more [real] because of the environment...like if I see down there is debris and there is someone else [in stimulus] also standing with me, so I could relate because he's there, I'm also there. So the environment or the other person made me feel that I was there [in the video]" **[P27, Playlist4]**

Audio was also an equally important aspect that all participants focused on. For stimuli that had a narration or a human talking in a language that the participant could not understand, the viewer reported having lower engagement levels due to the linguistic barrier. Many participants mentioned that the background music and ambient sounds, if done in moderation and matched the context of the stimulus, added to their sense of being present. A participant explained:

"It helped, it was going with the whole experience because it was a calming kind of audio so the experience is soothing and the audio is also suiting, so it makes the whole overall impact more suiting." **[P27, Playlist4]**

Participants reported having a narration introducing the stimuli, and the context was very important. It was also recorded that the narration element of the stimulus should be subtle and not *overdone*. Participants reported being disengaged when the narration dictated what emotion they should feel. [**P19, Playlist 3**] reported that for the stimulus *Instant Caribbean Vacation*, they felt that *"The narration, I think, killed it off, it felt like an advertisement. Like, when she is speaking, I am not able to feel anything"*. Many participants mentioned that unrealistic audio can make the stimulus feel like a movie, thus detaching them from it.

5.4 Discussion

Our findings highlight the significant impact of participant-specific factors and experiment decisions, including the use of VR-based stimuli, stimuli choices, and the data labeling process, on data. These aspects resonate with participant-centric data work and the collaborative technological development themes within the HCI community. In the subsequent

sections, we delve deeper into our findings and have discussed on their implications on data quality. We have then offered recommendations to HCI and AI researchers for future data work and experiment designs.

5.4.1 Challenges and Opportunities for Participant-Centric Data Collection

Our findings have highlighted two major aspects that play a crucial role in quality emotion data collection - 1) The role of participants and their meaning-making processes, and 2) The role of data collection setup and experiment choices. Previous studies on emotion data collection have typically treated participants as a collective with shared characteristics, often defined by criteria such as similar backgrounds (age, gender, education and nationality), health conditions (both physical and mental), language proficiency, susceptibility to motion sickness (in the case of virtual reality), and personality traits [184, 53, 185]. Further, the annotations are limited to self-reporting using objective scales like Likert, SAM, or some specific questionnaires, observer's rating, stimulus label, and physiological changes [37, 166, 186, 36]. However, our findings revealed that participant context encompasses more than just the experimental setup (e.g., room temperature, sensors used, stimulus choice, and stimulus order), demographics, and physical activity. It also includes finer contextual details such as self-perception, relationship with the stimulus, and attitude towards the experiment. These factors significantly modulate emotional responses, leading to varied annotations for the same stimulus. Contextual factors such as personal preferences, prior knowledge, experiences, aspirations related to the stimulus, logical reasoning, attitude towards the experiment, and expectations from the experiment are important guides to annotations. Each of these factors can decide how different one participant's annotations would be from another participant. Personal preference towards content can lead to favorable positive annotations to an otherwise negative stimulus. Similarly, prior knowledge about a stimulus can lead to an emotional detachment or low emotional reaction to the stimulus in the present. Prior experience of what is shown in a stimulus tends to increase the relatability with a

negative stimulus while it tends to reduce novelty from a high arousal positive (exciting) stimulus. The aspirations and expectations towards a stimulus can have similar effects as personal preferences since a stimulus aligning with expectations or aspirations can lead to a positive experience while a stimulus against them can cause a lack of interest among participants. Similarly, the attitude towards the experiment is another important factor that can decide how participants annotate and experience the stimulus. In our case, we found participants were interested in experiencing VR and thus were excited by the initial stimulus irrespective of the content because they were experiencing VR, not the stimulus and its content. Logical reasoning also tends to impact the annotations and often leads to human biases, such as not emotionally engaging in the stimulus. The participant-specific factor can thus lead to annotations that are biased by elevated or uninteresting experiences and may not be true representatives of emotions felt by participants and thus can reduce the quality of collected datasets. These contextual factors can also lead to an imbalance in the emotion data collected (For instance, if all participants felt only positive emotions because the setup was exciting or if they showed a lack of interest in negative stimuli because it was based on a past event like World War II), suggesting that equal distribution of stimuli according to the targeted emotions doesn't exactly lead to the collection of balanced amount of data for all emotions. Our statistical analysis also revealed that participants were able to annotate their emotions according to changing stimulus categories, while their physiological responses did not show significant changes. This suggests that participants rely more on their self-understanding and experiences for interpreting their emotions rather than on their physiological reactions, which suggests that annotations are a response to perceived emotions guided by participants' meaning-making. Our findings also aligned with Ricoeur's framework [187], which suggests that participants' annotations are shaped by their self-narratives and cognitive reflections rather than just physiological changes. This perspective highlights the importance of considering participants' mental and cognitive abilities in interpreting emotions instead of relying solely on self-reports, physiological data, and

contextual information like activity data and personality traits [53, 39, 184]. Our results also align with the appraisal theory [188] and the social constructivist theory of emotion [189], which suggest that individuals' explanations, interpretations, past experiences, cultural background, and context play a pivotal role in the experience of emotions, regardless of accompanying physiological changes [190].

Furthermore, we have also identified the positive and limiting effects that the choice of elicitation medium can have on the emotional response of participants. Our findings revealed that utilizing elicitation mediums like VR for emotion elicitation does have a significant impact in increasing the presence and immersion [191] and thus better elicitation. Nonetheless, we found an opportunity for further research in designing data collection experiments within HCI communities that should consider the bias that expectations (for and against the technology) related to VR technology may introduce in emotion data collection. Moreover, our findings highlighted the importance of designing stimuli where participants can have close-to-real-life experiences with scenarios that make sense in real life, such as human-level camera angles, the presence of human elements, and active design elements. Lastly, we identified that while designing experiments for data collection studies in laboratories, careful attention should be given to the ordering of stimuli. While we did not find any statistical difference between playlists, qualitative data suggested the impact of order in some specific scenarios. Elements such as whether the stimulus has a story, includes narration, or requires participant action versus no action are crucial. These factors influence the arrangement of stimuli and the instructions to be provided to participants for a smoother emotional experience. Our results suggested that the length of stimulus is correlated with the emotion and intensity of emotions. For instance, emotions that don't require a deeper connection with a stimulus can be shorter in length, while the emotions that take a longer time to develop should have longer stimulus. We also found that annotating emotions using objective scales was challenging for our participants. In contrast, they found interviews to be an easier medium for introspection and emotional interpretation, as they were prompted and

encouraged to reflect on their feelings. In the following section, we will discuss the avenues for accommodating participants' subjectiveness within data work and its implications for future work within HCI communities.

Incorporating Participant's Context

Participant's context is crucial to how they annotate emotions. These contextual factors are the major reason behind participant's annotations. For emotion recognition models to work better, it is important that an annotation method accurately and descriptively captures the emotions a participant has felt and the context behind them. Our findings present an opportunity for the HCI, and AI communities to develop collaborative experiment designs that acknowledge the burden placed on participants to interpret and quantify emotions using given annotation formats accurately and collect participants' context without adding much to this load. We recommend that researchers explore more descriptive methods of annotating emotional data, such as using text, audio, or qualitative questioning through chatbots or large language models. Moreover, creating stimulus-specific questionnaires to comprehend the factors contributing to emotional responses can also be explored. These questionnaires can be pre-designed using insights from pilot trials and qualitative participant interviews. For example, stimuli featuring elements like dogs should include questions about participants' attitudes towards dogs to incorporate potential influences on annotations. Researchers should also explore experimental design that includes prior participant profiling using either qualitative or survey-based methods. This step enables a comprehensive understanding of participants' personalities and emotional responses in their everyday lives. Capturing participants' personas and attitudes can be achieved by using questionnaires such as standard general psychological well-being tests [176], personality traits questionnaires [177], and experiment-specific reflective questionnaires [68]. This approach ensures a more nuanced exploration of individual characteristics, contributing to a richer and more meaningful interpretation of collected data.

Designing for Participant in the AI Pipeline

Our findings underscored the critical role of experimental design choices, advocating for collaborative and participant-centred approaches. Researchers from HCI communities should aim to design environments and stimuli that offer realistic psychological experiences [192], thereby minimizing the influence of lab conditions. Moreover, our study highlights the importance of authentic emotional experiences when engaging with stimuli. Currently, much of the research relies on open-source stimuli derived from publicly available videos, music, or films. However, these stimuli may not be specifically curated for training AI models. Future studies could explore the development of stimuli that integrate these elements more effectively for emotional data collection purposes. Another interesting direction would be to explore designing emotional stimuli with humans in the loop. To address the influence of stimulus order, researchers should consider designing emotion-centric experiments tailored to the specific application. This approach contrasts with a one-size-fits-all design, where participants are exposed to a broad spectrum of emotions. Instead, focusing on the application's specific emotional context ensures a more nuanced understanding of participants' responses. Researchers can find inspiration in studies that specifically addressed single emotions such as harmful stress [193], anxiety [194], happiness [195] and depression [196]. Researchers can look into using qualitative methods and pilot studies with participants to decide the length of stimulus in the case of laboratory settings. Our findings suggested different emotions require different exposure lengths, suggesting a need for deciding the stimulus length based on the emotion category rather than using the same length for all emotions. To handle the complexities arising from mixed emotions, multiple questionnaires in the annotation procedure of emotion data, and reliance on participants' responses, we recommend practitioners employ techniques such as combining discrete and continuous labels [64], utilizing unsupervised or semi-supervised methods to identify labels from the data, or chat-bot based annotations for more context on emotions [197]. To accommodate participants' biases towards the technology, we suggest future works to design more elaborate

familiarization training or acclimatization procedures along with instructions on what is expected from the current level of VR technology.

5.4.2 Data-Work supporting Model-Work

Our research highlights the difficulties associated with collecting physiological emotion data. In this section, we will delve further into another side of the spectrum - designing experiments with less burden on human participants. Currently, AI-based interventions lean heavily on supervised techniques that rely extensively on the availability of accurate data annotations. However, our analysis reveals the inherent challenge of obtaining annotations that precisely correspond with physiological changes. Factors such as the subjectivity of participants, reliance on participant's interpretations of emotions, and a lack of contextual information in objective labels contribute to the complexity of data collection. These challenges emphasize the need for nuanced approaches and to pursue more accurate and reliable physiological emotion data collection. In the following sections, we will explore opportunities for HCI and AI researchers to design data collection experiments that address these challenges in emotion AI data work.

Taking Holistic or Application-Centric Approaches

Collecting physiological emotion data is commonly regarded as a means to gain insights into human emotions. This process typically involves participants either annotating the emotions they experience during the data collection period or being deliberately exposed to emotions of interest as per the study design. Both approaches hinge on the participants accurately identifying and annotating their emotions. Although models trained on existing datasets have demonstrated some level of performance [37], their practical implementation remains challenging. This challenge arises from the complexities of accurately translating physiological responses into a reliable and realistic representation of human emotions. Researchers designing experiments should design methodology that approaches data collection from

a holistic view wherein contextual data such as participants' characteristics (age, gender, personality), mood, motivation, physical activity, lifestyle, daily routine, location, health conditions, food, and caffeine intake should also be captured, along with emotion annotations [198]. Moreover, data collection and model development are frequently approached as distinct phases in research within AI communities, while HCI communities have pointed out the importance of data-centric approaches wherein data collection is also treated as a crucial aspect of AI [199, 30]. Traditionally, the data-related tasks are carried out independently of considerations for the subsequent model work. This separation results in preprocessing, data aggregation, and data preparation occurring in a later stage [30]. To illustrate this compartmentalization in emotion data, consider the process of collecting physiological data in a laboratory setting. In this scenario, participants are intentionally exposed to various positive and negative emotions and scenarios. However, when preparing a model, the data is often categorized into a binary classification of stressful and non-stressful, oversimplifying the richness of the collected information for the purpose of classification. We suggest researchers to approach data collection from an application perspective; for instance, if the application detects harmful stress versus eustress (beneficial stress), the annotation questionnaires should be specifically designed to collect stress information from participants rather than utilizing a standard questionnaire like SAM and PANAS which contains questions on emotions that aren't necessary for future model work.

Moving towards Annotation-Free or Minimal Annotation Approaches

Annotations are critical in AI data, serving as a fundamental component. When annotations fail to represent the underlying data accurately, it may lead to seemingly impressive performance metrics but ultimately produce results that are meaningless or unreliable. Collecting annotations for emotion data is particularly challenging due to the numerous factors that can influence the annotations. To address these challenges, we propose that researchers explore semi-supervised approaches [200] and weakly supervised methods [201] to leverage avail-

able data more effectively. In the realm of physiological emotion data collected in real-world settings, there is an opportunity to design experiments that align with circadian rhythms [202]. For instance, capturing physiological data at night could be labeled as representing rest or homeostasis (internal stability), using the temporal nature of physiological changes to inform the labeling process. This approach can potentially help researchers working on emotion data to collect quality data while minimizing the dependence on participants for annotations.

5.5 Limitations

A significant limitation of our study is the generalizability of our findings due to our participants' cultural backgrounds and technology awareness. Despite employing diverse recruitment methods to ensure varied participation, our current demographic primarily comprises tech-savvy individuals aged 18-33 with a higher level of education. It is crucial to acknowledge that our findings may be influenced by participants' age, educational backgrounds, technological experience, and emotional self-understanding. While these limitations exist, it is essential to recognize that emotion elicitation is inherently challenging and strongly influenced by personalized factors. Our study underscores the significance of understanding and collaboratively involving human participants in emotion data work. Additionally, we stress that our study, conducted in controlled lab settings, may not readily extend to data collection in natural, real-world environments. However, despite taking place in controlled lab environments using passive virtual reality stimuli, we believe some of our findings apply to the broader paradigm of physiological emotion data-collection studies, while others remain specific to VR-based elicitation.

Part II: Participants' Perspectives on Emotion Data Collection in Everyday Settings

Physiological signal-based emotion tracking is still in its early stages, especially for continuous, reliable monitoring in everyday life [89]. While there is growing commercial interest and increasing integration of physiological sensors into consumer wearables, current systems are limited to stress tracking and lack the sensitivity, personalization, and contextual awareness needed to fully capture the complexity of human emotions in real-world settings. AI-powered solutions that combine wearable devices with mobile phone applications offer a promising approach to addressing these challenges, especially given recent advancements and the growing adoption of AI to tackle complex, real-world problems. However, a major limitation in developing such models is their heavy reliance on high-quality, labeled emotion data that accurately reflects the nuances of human emotional experiences [203, 36, 1, 204]. As discussed in Chapter 1, most existing emotion datasets are collected in controlled laboratory environments or through short-term studies using self-report methods or expert annotations, and they often lack participants' centric, nuanced self-reporting of emotions. As a result, these approaches frequently fall short of capturing the dynamic, context-dependent, and multi-layered nature of emotions as they naturally unfold [205, 1, 26, 206, 5]. Consequently, models trained on such data often struggle to generalize to real-world scenarios, limiting their effectiveness and reliability for end users [205, 36, 207, 203], suggesting a need for methods to accurately capture emotional annotations and contextual information in real-life settings. Prior work in real-life settings has primarily relied on self-assessment approaches, such as the Experience Sampling Method (ESM)—which prompts users to report their emotions at regular or random intervals throughout the day—and the Day Reconstruction Method (DRM), where participants provide retrospective reports of their emotional states via structured questionnaires at the end of the day [68, 208, 12, 209, 117]. While useful for capturing momentary or reflective self-reports, these methods often provide labels using predefined emotional scales (such as Self-Assessment Manikin

(SAM) [179] or scales based on six basic emotions [23]), offering limited insights into contextual information about emotional experiences. Additionally, previous studies have highlighted further challenges; for instance, the act of annotation itself may influence the user's emotional state, introducing measurement bias [210]. Moreover, participants often experience annotation fatigue, leading to low motivation and engagement over time [211]. Furthermore, prior research has also emphasized the need for ecologically valid, human-centric data collection approaches that reflect how emotions are naturally experienced in everyday contexts [36, 1, 205] These insights suggests a significant gap in studies that investigate in existing methodologies from participant perspectives, who are not only the sources of data but also its annotators [1, 205].

To explore these challenges and identify opportunities in everyday emotion annotation methods for physiological signal-based emotion AI research using wearables and mobile phone data, we designed this qualitative investigation. Our approach centers on examining the perspectives of diverse stakeholders, including users and non-users of emotion-tracking technologies and mental health professionals. Through this lens, we aim to understand the human side of emotion data collection and its implications for designing more effective and user-centric emotion AI systems. Specifically, we address the following research questions:

- **RQ1:** What are participants' perspectives on the challenges and opportunities for annotating (identifying and labeling) and tracking emotions in their daily lives?
- **RQ2:** What are the perspectives of mental health professionals on the challenges and opportunities in physiological emotion data collection for developing emotion AI interventions?
- **RQ3:** How can participants' and domain experts' perspectives be integrated to develop a holistic methodology for collecting physiological emotion data in real-life settings?

We employed a qualitative research method, including surveys (n = 75) and interviews (n = 32) with members of the public (with and without experience in therapy or counseling,

as well as those who have used wearables and mobile phone applications for tracking stress and emotions, or participated in emotion data collection studies), as well as focus group discussions (n = 3) involving 12 mental health professionals. Following our methodology, our study evaluated participants' needs from diverse perspectives, with a total sample of 119. In this study, *emotion* is defined in line with the "Theory of Constructed Emotion" as proposed by Lisa Barrett [71], which views emotions - "as individualized, context-dependent experiences constructed by the brain through the interpretation of bodily sensations (e.g., heart rate, arousal) in relation to past experiences, situational context, and learned emotional concepts, not as fixed biological responses". In contrast with prior research, which typically models *emotions* as changes in physiological responses and behavioral reactions, using physiological signals (e.g., heart rate, skin temperature, electrodermal activity), behavioral patterns, and self-reported data [212]. By adopting Barrett's perspective, this study emphasizes the context-dependent and dynamic nature of emotions, enabling a more comprehensive exploration of emotional experiences. Also, for this study, we have referred to structured methods, such as PANAS, SAM, or Likert scales, as objective methods, while unstructured methods, such as the option to write, audio-record, or add images, are referred to as subjective methods. This human-centric view of *emotions* helps us to study the subjectivity and variability of emotional experiences, challenging the assumption that specific physiological or behavioral patterns can map directly to discrete or dimensional emotion categories [5]. This part of Chapter 4 makes the following key contributions:

- **Annosense Framework:** We introduce Annosense, a novel framework comprising 15 actionable guidelines for collecting well-annotated wearable and mobile-based emotion data in everyday settings. These guidelines are derived from an in-depth analysis of user experiences, contextual challenges, and common challenges in emotion data collection.

- **Expert Evaluation**: We evaluate the Annosense framework through feedback from 25 emotion AI experts with professional and academic experience, making this the

first work to present evaluated guidelines tailored specifically for real-world emotion data collection.

- **Potential Implementation**: We identify next steps for designing participant-aware systems in practice, informed by the Annosense framework and a review of current tools, technologies, and applications.

- **Design Implications**: Through our findings and expert discussions, we offer design recommendations for future emotion data collection practices and AI algorithm development.

5.6 Methodology

We employed a qualitative research approach, utilizing surveys, interviews, and focus group discussions to gather diverse perspectives from our stakeholders. This study received Institutional Review Board (IRB) approval to ensure ethical compliance. Participation (Survey, Interview, FGD, and Guideline Evaluation) was entirely voluntary, and no compensation was offered to participants. The following subsections provide a detailed explanation of the methods we employed.

5.6.1 Survey

To answer our research questions, we surveyed individuals aged 18 and above, including people with or without experience with emotion tracking technologies. Our survey was developed as an exploratory, mixed-methods tool to investigate how users perceive, interpret, and prefer to annotate emotional experiences in their daily lives. It was designed with reference to established emotion theories [213, 71] and user-centered design principles [214, 215], including a mix of quantitative and open-ended questions to capture both structured responses and rich personal narratives (see appendix C.1 for detailed questionnaire). It comprised 27 questions organized into three main sections: **1) Demographic Details**: This

section collected basic demographic information about our study participants, **2) Understanding Emotional Awareness**: This section was designed to explore how participants perceive, differentiate, and articulate their emotional experiences and was grounded in the concepts of emotional awareness, emotional vocabulary, and emotional granularity [213, 71]. To assess emotional awareness, participants were asked questions such as, "How often do you take time to reflect on your emotions?", "When experiencing a strong emotion, how easily can you identify what emotion you are feeling?" and "How often do you feel mixed emotions?" Further, they were asked to reflect on which emotions they find easier or harder to identify and to list five positive and/or negative emotions they commonly experience, along with their impact on daily life. These responses helped us understand each participant's emotional vocabulary and how they articulate emotional states. To further assess emotional granularity— the ability to distinguish between similar emotions— participants were asked whether they could tell emotions like sadness and disappointment or anger and frustration apart, and to explain their reasoning. This provided insight into their ability to make fine-grained emotional distinctions linked to more effective emotional regulation and self-awareness. In addition, participants were prompted to describe any recent situations involving strong emotions and to identify the emotions they experienced, to evaluate further their ability to identify and express emotions linguistically [216]. We also included questions to assess the conceptual understanding of important terms like emotions and emotional intensity. To assess conceptual understanding, we included targeted items such as a multiple-choice question asking participants to define "emotion" (e.g., as a bodily sensation, mental state, or response to external events). Further, we also asked them to define "emotional intensity" in their own words. Further, participants were asked about their previous experience with emotion management and use of tools or real-life techniques, such as wearables, emotion-tracking applications, mindfulness practices, and journaling, that assist them in identifying, labeling, and regulating emotions. **3) Attitudes Toward Daily Emotion Annotation**: This section examined how users feel about incorporating

emotion annotation into their daily routines. It included questions assessing the willingness to annotate positive and negative emotions, preferred annotation methods, and perceived barriers. Example questions included: "Would you like to annotate your emotions daily?", "What factors are most important to consider when labeling emotions?", and "How easy do you find it to annotate your emotions daily?" Participants were also asked to express their annotation preferences (e.g., emoji, voice input, or descriptive text) and their preferred frequency to annotate emotions daily.

To maintain the quality of our survey responses, we included several consistency checks in our questionnaire, where users were prompted to explain their choices qualitatively for multiple-choice and Likert-type questions. Further, our survey contained 14 open-ended questions, which also enhanced the depth and authenticity of our survey data. Additionally, to validate our survey design for question clarity, logical flow, and completion time, we conducted a pilot with 6 participants before our data collection. Moreover, to evaluate the quality of the responses to our open-ended survey questions, we calculated completion rates to assess participant engagement, distinguishing between required and optional questions. Additionally, we examined word count statistics for each open-ended question, including range, mean, and standard deviation, as proxies for response depth and variation (see Table C.1). There were 14 open-ended questions in the survey, eight of which were mandatory. Overall, the completion rate for open-ended questions was high, with an average of 85% for required questions and 82.9% for non-mandatory ones. This indicates strong participant engagement, even when responses were optional. The quality of responses varied across questions, with some eliciting brief answers and others generating in-depth, detailed feedback. Following designing and testing, our survey was distributed digitally using Google Forms. Participants were recruited using convenience sampling methods [217], using social media platforms like WhatsApp and an email call within our institute. Before filling out the survey, we provided our participants with brief information about our study's aim, potential risks, benefits, and confidentiality policy, followed by an informed consent form. Our survey

did not collect any identifiable information to maintain anonymity. We got 77 responses to our survey, out of which 75 participants filled out the complete form; the demographic details of our participants are provided in Table 5.5. All the valid survey responses were exported into Google Sheets for analysis. We analyzed the closed-ended question (n=13) using descriptive analysis, such as calculating percentages, cross-tabulation, and visualizations. For open-ended questions (n=14), we performed thematic analysis [218] and generated codes such as *"Self-reflective practices"*, *"Challenges in Emotion Identification"*, and *"Emotional Literacy"*. Examples of themes we identified were *"Emotional Awareness and Regulation"* and *"Language and Emotional Expression"*.

Category	Details and Count
Age	Range: **18–41**, Mean: **24.9**, SD: **3.54**
Gender	Males **(46)**, Females **(28)**, Prefer not to say **(1)**
Education	Bachelor's **(42)**, Master's **(21)**, Doctorate **(7)**, Senior High School **(4)**, Vocational Diploma **(1)**
Occupation	Students **(22)**, Professionals/Business **(24)**, Not applicable **(29)**
Prior Experience	No prior experience **(38)**, With prior experience **(37)** (e.g., journaling, mindfulness, self-reflection/introspection, apps, and wearables)

Table 5.5: Summary of survey participants' demographics. Prior experience refers to participants' use of tools or techniques for emotion tracking or management. Participants could report more than one technique; no experience indicates that participants do not actively track or manage emotions in daily life.

5.6.2 Interviews

We conducted our formative semi-structured interviews with 32 participants. To guide our semi-structured interviews, we adopted the 5W1H framework [219]. This framework was particularly well-suited for our study since it was an early-stage design study, which aimed at exploring how individuals perceive, approach, and reflect on the act of annotating or self-reporting emotions in their daily lives. As emotion tracking is a deeply personal and context-dependent practice, we needed a method that could surface not only what participants do, but also why and how they do it, within the broader landscape of their

routines, motivations, and challenges. Prior to conducting our participants' interviews, we did pilots with 5 participants to understand the flow of our interview design and the relevance of questions. Our interview design was further guided by prior qualitative research on emotions [220, 121, 91]. Below is an explanation of our interview design: **1)"WHO**- are they?": Focused on understanding participants' emotional awareness, experiences, and familiarity with technology for emotion tracking. Questions explored how they perceive and manage emotions, their prior experience with emotion data collection and logging, and the perceived psychological impact of the process on their lifestyle. **2) "WHAT**- would they annotate?": Examined the types of emotions or emotional events participants considered worth annotating. Questions included privacy concerns and whether they would share detailed information about their emotions. **3) "WHEN**- would they annotate?": Addressed the timing and frequency of emotion annotation. Participants were asked about their preferences for real-time versus retrospective annotation, the contexts or scenarios where they felt annotation was most appropriate, what kind of prompts, and at what frequency they might prefer to be notified for annotating. **4) "WHERE**- would they annotate?": Explored the environments where participants would feel comfortable annotating their emotions, as well as locations they might avoid. **5) "WHY**- would they annotate?": Investigated participants' motivations for annotating emotions, including perceived benefits and potential challenges or barriers. We asked them to discuss the benefits they see in annotating both positive and negative emotions. **6) "HOW**- would they annotate?": Delved into preferred tools and methods for annotation, the time participants were willing to dedicate, and their expectations for simplifying or improving the annotation process. A detailed description of our interview questions is provided in Appendix C.2.

Our participant pool was well-educated, technology-friendly individuals aged 18 and above, with and without any experience of emotion tracking technology, recruited through convenience sampling [217], using social media platforms like WhatsApp, as well as an email call. We received interest from 32 individuals for the interviews. Before conducting

the interviews, we obtained digital consent from each participant through Google Forms sent via email. Along with their consent, we also collected information on their age, gender, education, current occupation, mental health conditions, prior experience with therapy/counseling, and wearables/emotion tracking/emotion data collection studies. Sixteen of our participants had prior experience with either participating in emotion data collection studies or using wearables for stress detection and emotion/mood tracking applications. The rest of our participants did not use technology-based mediums to understand their emotions and mostly relied on techniques such as self-introspection, meditation, exercises, communication with other people, or other mindfulness or coping techniques to deal with emotions. Details about our interview participants are summarized in Table 5.6.

Category	**Details and Count**
Age	Range: **19–43**, Mean: **26.96**, SD: **7.15**
Gender	Males **(15)**, Females **(17)**
Education	High School/Diploma **(3)**, Bachelor's **(17)**, Master's **(6)**, Doctorate **(3)**, Postdoctoral **(3)**
Occupation	Students **(19)**, Professionals/Business **(13)**
Attended Therapy	Yes **(13)**, No **(19)**
Diagnosis	Yes **(2)**, No **(30)**
Prior Experience	No prior experience **(16)**, With prior experience **(16)**

Table 5.6: Summary of interview participants' demographics. Diagnosis refers to self-reported mental health diagnoses. Prior experience includes participants' experience using tools or techniques for emotion or mood tracking and participation in emotion data collection studies. Participants could report more than one technique; no experience indicates participants who do not actively use technology to manage their emotions.

The interviews were conducted in English, either online or offline, based on the participant's preference. Each session was recorded using Zoom Pro, following verbal consent. The interviews began with a brief introduction to the study and familiarization with terms such as "emotions," "emotion annotations," and "emotion intensity" to ensure participants understood the terminology and process of emotion data collection. The definition used to explain emotion annotation to our participant is *"The process of identifying, labeling, and documenting emotional experiences, often to capture emotional data for research, self-*

reflection, or technological applications. It involves assigning labels (e.g., specific emotions like happiness, anger, or sadness) to emotional events using methods such as written records, mood-tracking apps, emojis, or voice recordings." Each interview lasted approximately 30 minutes. The sessions were transcribed using Zoom's built-in audio transcription feature. The transcribed documents were then exported to Google Docs and manually reviewed by the first and second authors for grammatical and transcription errors using the original voice recordings. Following transcription, we performed the inductive thematic analysis [183]. We began with authors 1 and 2 reading and re-reading the interview to familiarize themselves with all the data individually. Following this, they individually generated the initial codes for all the data, which included codes like *"Avoidance and denial as coping mechanism," "Understanding of basic emotions"*, and *"Challenges with using static Likert scales"*. Later, authors 1 and 2 grouped similar codes together to form potential themes. Following this, all the authors reviewed the themes together by reviewing the data within each theme to ensure it accurately reflected the data. The high-level themes included *"Emotional literacy and awareness," "Technology and concerns"*, *"Emotional Regulation and management methods"*, and *"Barriers to Identifying and Annotating Emotions"*. All authors reviewed and refined the themes iteratively to ensure they were coherent and distinct until saturation. The themes formulated in the process helped us to structure our findings (see section 5.7).

Category	**Details and Count**
Professional Title	Psychologist **(1)**, Clinical Psychologist **(2)**, Psychiatrist **(6)**, Peer Counselor **(3)**
Year of Experience	Less than 1 year **(2)**, 1–3 years **(4)**, 4–7 years **(1)**, 8–10 years **(3)**, More than 10 years **(2)**
Experience with AI and Wearable Technology	No **(8)**, Yes **(4)**

Table 5.7: Summary of focus group discussion participants' professional demographics.

5.6.3 Focus Group Discussion

To conduct our focus group discussions (FGDs), we utilized purposive sampling [221] to recruit mental health professionals. We reached out to our collaborators, including doctors and NGOs, who helped disseminate our call for participation along with an interest form. From the 18 responses we received, 12 professionals were available for the scheduled time slots. We conducted three separate FGDs: FGD1 included 4 professionals (1 Psychiatrist and 3 Peer Counselors), FGD2 included 3 professionals (2 Clinical Psychologists and 1 Psychiatrist), and FGD3 included 5 professionals (1 Psychologist and 4 Psychiatrists). All participants in the focus groups were from the same country and shared a common cultural background as interview and survey participants, minimizing variability due to cross-cultural differences. Our FGD was designed in line with the prior qualitative studies done with domain experts [91, 222, 30]. Prior to the FGDs, we obtained digital consent from each participant through Google Forms sent via email. In addition to consent, we collected information on their professional titles, years of experience in the mental health field, and familiarity with AI or wearable technology. Details about the participants are summarized in Table 5.7. All our FGDs were conducted online via Zoom Pro, with a single moderator leading each session following verbal consent to record the meeting. Each FGD began with a brief introduction of all participants within the discussion, followed by introduction slides presented by the moderator to outline the role of AI in healthcare, the definition of emotion AI, and a brief description of the current practices in AI for emotion data collection and labeling. This overview ensured that all participants had a shared understanding before the discussions began. Following this introduction, the discussion was organized into three main segments: 1) Current Practices for Assessing Emotional States, 2) Attitudes Towards Data and AI, and 3) Challenges and Opportunities in Emotion Data Collection and Recognition. In the first segment, professionals discussed their current methods for assessing the emotional states of patients and clients. The second segment focused on their initial impressions of using AI to understand and monitor emotions, including potential

benefits and drawbacks in clinical settings. In the final segment, participants reviewed the current practices of AI data collection, as described in the introduction, and provided their perspectives, recommendations, and insights based on their own practices for the future. A more detailed description of our FGD is provided in the appendix C.3. Each FGD lasted approximately 1 hour and 15 minutes and was conducted in English. The sessions were transcribed using Zoom's built-in audio transcription feature. The transcribed documents were then exported to Google Docs and manually reviewed by authors for grammatical and transcription errors using the original voice recordings. The first two authors then completed the familiarization, where they thoroughly read all the transcripts. Next, initial codes are generated by reading all the data systematically, highlighting data segments, and assigning brief labels that capture their essence, following inductive thematic analysis [183]. The initial codes included *"Positive attitude towards AI"* and *"Emotional labeling is a mix of subjective and objective labels"*. The authors 1 and 2 searched for themes by grouping similar codes, forming broader patterns. Following this, all the authors jointly reviewed and refined the themes to generate coherent and distinct themes for structuring the findings (see section 5.7). A few examples of our identified themes are *"Parallel source of information"* and *"Emotional ground truth"*.

5.6.4 Development of Guidelines

To develop our guidelines, we analyzed the data from each source (surveys, interviews, and focus groups) independently to identify recurring themes and patterns. Next, we grouped similar themes and organized them iteratively [183] under the three guideline stages *"Pre-data collection," "During data collection," and "Post-data collection"*, ensuring logical flow and coherence. We then cross-referenced our identified themes with methodologies and recommendations from prior studies to validate and expand our understanding [223, 36, 224, 168, 61, 62]. Finally, we synthesized the insights from participant data and literature to create an end-to-end framework for everyday emotion data collection that is practical,

evidence-based, and user-centered. This process resulted in 15 guidelines (named G#) divided into three data-collection stages, as presented in Table 5.11, 5.13, and 5.14. Further, we evaluated our guidelines for their validity in emotion AI research (see section 5.8).

5.7 Designing *AnnoSense* — An Everyday Emotion Data Collection Framework for AI

This section introduces *AnnoSense*, a framework comprising 15 guidelines designed to support robust emotion data collection in everyday contexts, enabling the development of AI models applicable to real-life scenarios. *AnnoSense* is structured into three phases: pre-data collection, during-data collection, and post-data collection. Within each subsection, we will present findings from our data surveys, interviews, and FGDs to demonstrate the data-driven origins of each guideline. To ensure clarity and ease of navigation, we begin by presenting our data observations, structured into thematic subsections. These are followed by a dedicated subsection—Derived Guidelines—which outlines design guidelines that are directly informed by and grounded in the preceding observations.

Survey Question	Response	Count	Percentage
How often do you take time to reflect on your emotions?	Never	2	2.7%
	Rarely	11	14.7%
	Sometimes	**27**	**36.0%**
	Often	25	33.3%
	Always	10	13.3%
When experiencing a strong emotion, how easily can you identify the emotion?	Very difficult	2	2.7%
	Difficult	12	16.0%
	Neutral	9	12.0%
	Easy	**40**	**53.3%**
	Very easy	12	16.0%
How often do you feel mixed emotions (experiencing multiple emotions at once)?	Never	1	1.3%
	Rarely	18	24.0%
	Sometimes	**30**	**40.0%**
	Often	24	32.0%
	Always	2	2.7%

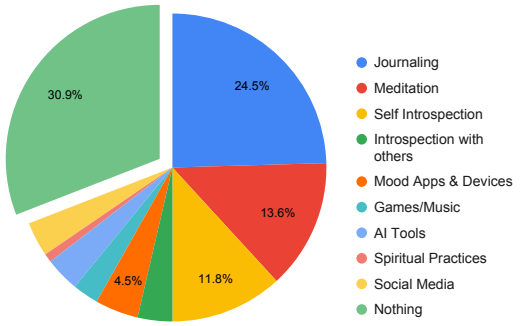
Table 5.8: Participant responses on emotional awareness and reflection (N = 75)

5.7.1 *"Two-way Communication"*: Pre-Data Collection Phase (G1-G6)

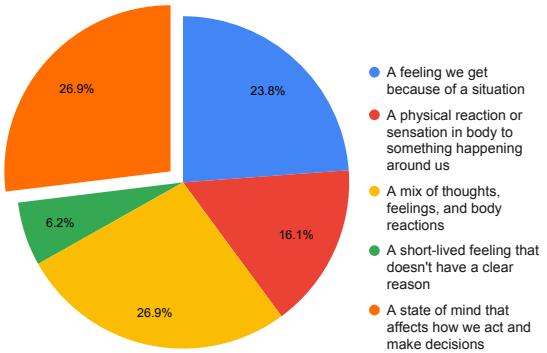
Data Observations:

To design our pre-study guidelines, we analyzed data from surveys, participant interviews, and focus group discussions to understand participants' specific needs prior to data collection.

1) Need for Prior Preparation and Training: Our survey findings indicate that participants generally believed that they possess moderate to high emotional awareness, with 69.3% reporting that they can easily identify their emotions during intense emotional experiences. However, their emotional reflection habits vary considerably, 46.6% reported to engage consistently in self-reflection, while a notable portion rarely or never does (more details in Table 5.8). Although many participants acknowledge experiencing mixed emotions, suggesting an awareness of emotional complexity, fewer than half use structured methods such as journaling, meditation, or introspection to process these feelings (see Figure 5.4a). A significant number (34 participants) reported using nothing at all or using suppression or avoidance techniques such as social media and video games, implying that emotional insight for many relies on instinct rather than intentional strategies. This trend was further observed in our interview participants. Participants reported using distraction techniques like watching movies, playing games, or avoiding emotions as a common method for dealing with emotions (mostly negative). As expressed by **(P8, Interview)** - *"I usually sleep. I usually watch television. I usually watch a web series. Nothing else means I can do anything, or I just play some video game. That is the only way of dealing with these emotions, like, sometimes when I'm too stressed"*. Moreover, participants also mentioned not talking about or reflecting on deeper negative emotions due to the stigma of sharing or acknowledging emotions. Participants have used statements like - *"Emotions make me feel weak"*, *"It's better to keep emotions inside"*, or *"Why should we track emotions? It is for people with mental disorders"* suggesting the deep-rooted stigma towards expressing, processing, or



(a) Survey results on emotion tracking practices.



(b) Participants' understanding of the definition of emotion.

Figure 5.4: Survey results for emotion awareness and management practices among our participants.

tracking emotions.

Furthermore, in our survey, when we asked participants to *recall a recent situation in which they experienced a strong emotion and describe both the context and the emotion identified*, to explore participants' emotional awareness, articulation, and the types of emotional experiences they tend to recall. A majority of participants (60%) were able to identify specific emotions tied to their experiences. Among these, anger, sadness, and anxiety were the most commonly reported emotions, suggesting negative emotions are commonly recalled by people. Mixed emotions were a notable part of the responses (9%), reflecting the complex nature of human emotions. Lastly, 21% of participants displayed uncertainty in expressing or identifying their emotions, with some responses showing emotional ambiguity or no clear emotion at all. This reflected the differences in participants' recall behaviors, where a majority of participants recalled negative events or were uncertain about expressing their emotions, possibly due to subconscious stigma or lack of vocabulary or awareness.

Further, in our survey data, we found that a large number of participants are only aware of basic emotions such as happiness (46), sadness (28), joy (26), and anger (27), as shown in Table 5.9. Interestingly, these primary emotions, such as anger (20), happiness (18), and sadness (13), were also frequently reported as easily identifiable (see Table 5.10). We also observed that several emotions appeared in both "easy" and "hard" to identify categories

such as, anger (20 vs. 10), sadness (13 vs. 8), anxiety (4 vs. 8), and happiness (18 vs. 4). This contradiction suggests the presence of distinct subgroups with varying levels of emotional literacy within our sample. Additionally, our survey data also revealed varying understanding among our participants about what they consider *emotions*, as shown in Figure 5.4b.

Positive Emotion	**Frequency**	**Negative Emotion**	**Frequency**
Happy	46	Sadness	28
Gratitude	38	Anger	27
Joy	26	Anxiety	16
Hope	21	Frustration	10
Love	18	**Motivation****	9
Peace	16	Fear	9
Excitement	14	Loneliness	8
Satisfaction	11	Guilt	6
Confidence	10	Jealousy	6
Motivation	9	Stress	4
Calm	9	Irritation	4

Table 5.9: Frequency of Top 10 Positive and Negative Emotions in Daily Life as Reported in our Survey. For positive emotions, the mean frequency of responses was 3.88 with a standard deviation of 1.8, while for negative emotions, the mean frequency was 2.48 with a standard deviation of 1.9. **Note: **Motivation** is contextually a positive emotion but was mentioned in the negative list—possibly reflecting low or lack of motivation.

2) Understanding the Participant Profile: Extending our investigation into emotional literacy, analysis of our survey question *"Can you differentiate between similar emotions (e.g., sadness vs. disappointment)?"* revealed significant variations in participants' emotional granularity capabilities. Results showed that 44.0% of participants explicitly reported difficulty differentiating between similar emotions, while only 20.9% indicated confidence in their ability to distinguish nuanced emotional states. The remaining 35.2% provided responses that were difficult to categorize definitively. Further, in our data, we observed several recurring themes: 1) sadness was characterized as a broader mood and disappointment as a more targeted emotion, 2) disappointment was frequently framed as a response to unmet expectations, and 3) they were differentiated based on perceived control, emotional

Emotions Easy to Identify		Emotions Hard to Identify	
Emotion	**Count**	**Emotion**	**Count**
Anger	20	Anger	10
Happiness	18	Sadness	8
Sadness	13	Anxiety	8
Frustration	6	Satisfaction	4
Joy	6	Fear	4
Anxiety	4	Happiness	4
Love	4	Depression	4
Loneliness	3	Positive	3
Disappointment	3	Jealousy	3
Hope	3	Guilt	3

Table 5.10: Comparison of top 10 emotions based on ease of identification as per our survey response.

intensity, and temporal duration. These findings further reflected the varying emotional abilities among our participants, suggesting that a one-size-fits-all solution to emotion data collection might not be sufficient for collecting quality data. Further, our discussions with experts also reconfirmed the *varying emotional literacy* as explained by an expert **(P1, Psychologist, FGD3)**, *"People tend to feel only 4-5 basic emotions and lack a vocabulary to explain their emotions and must be taught...A therapist tries to teach people about emotional awareness to improve vocabulary as it helps them identify emotions more clearly, along with their professional methods."* To overcome these challenges, domain experts within our FGDs emphasized efficient history-taking to understand emotion data reliably. Further experts also emphasized the importance of assessing various emotional aspects such as *emotional vocabulary*, *emotional range* (the spectrum of emotions a person can experience, express, and recognize), *emotional congruency* (the consistency between inner feelings and outward expressions), *emotional intensity* (degree of an emotional experience), and *emotional reactivity* (the intensity and speed of an individual's emotional response to a stimulus) to understand emotional data better. Finally, experts highlighted the importance of screening for conditions like alexithymia, which affects an individual's ability to identify and describe emotions. Finally, our participants' data and experts' discussions also revealed that

Guideline	Description
G1	**Selecting Participants**: **1.** Document the inclusion and exclusion criteria based on the study's objective and data requirements. **2.** Recruit individuals from diverse demographic groups (age, gender, culture, education, and occupation) in line with the study's inclusion criteria and data-requirements. **3.** Screen participants for alexithymia (difficulty identifying and expressing emotions) using standardized screening tools such as the Toronto Alexithymia Scale [225] or Perth Alexithymia Questionnaire [226], neurological disorders (e.g., cognitive impairments), and health conditions (e.g., cardiovascular issues or chronic illnesses) that might impact physiological signals, ability to identify and express emotions, and in line with the study's exclusion criteria.
G2	**Obtaining Informed Consent**: **1.** Clearly explain the purpose, benefits, potential risks, compensation, voluntary participation, time commitment, and key procedures of the study in simple, accessible language. Provide enough information to the participants without revealing details that could compromise the integrity of the study design. **2.** Clearly outline ethical approval and privacy measures, such as how participant data will be anonymized (e.g., removal of personal identifiers), compliance with relevant data protection laws (e.g., GDPR, HIPAA), secure data storage practices, and data sharing in the consent document.
G3	**Conduct Initial Calibration**: **1.** Conduct a baseline session or calibration trials (as per study requirements and resources) to familiarize participants with the devices being used in the study. **2.** Provide clear instructions on how to correctly wear the devices and ensure they are functioning accurately during data collection.
G4	**Provide Participant Training**: **1.** Organize practice sessions where participants label their emotions (e.g., joy, anger, sadness), identify subtle distinctions (e.g., frustration vs. irritation), and document contextual factors such as environment, social interactions, and cultural norms in real-time or respond to controlled stimuli, enabling researchers to clarify doubts and improve annotation accuracy. **2.** Educate participants about the broader impacts of emotion annotations and the data privacy measures in place to build initial trust and engagement. **3.** Offer expert-verified resources, including video clips, audio recordings, or books, to improve participants' emotional literacy and understanding of what is meant by emotion annotations.
G5	**Perform Detailed Psycho-Social Profiling of Participants**: **1.** Collaborate with domain experts to collect detailed history on characteristics like emotional range (spectrum of emotions a person can experience, express, and recognize), emotional congruency (consistency between inner feelings and outward expressions), emotional intensity (degree of an emotional experience), and emotional reactivity (the speed of emotional response), and emotional vocabulary. **2.** Collect contextual details such as past traumatic experiences, daily routines, work-life balance, family dynamics, emotional awareness, regulation habits, and potential stigma using standardized questionnaires or with the assistance of domain experts.

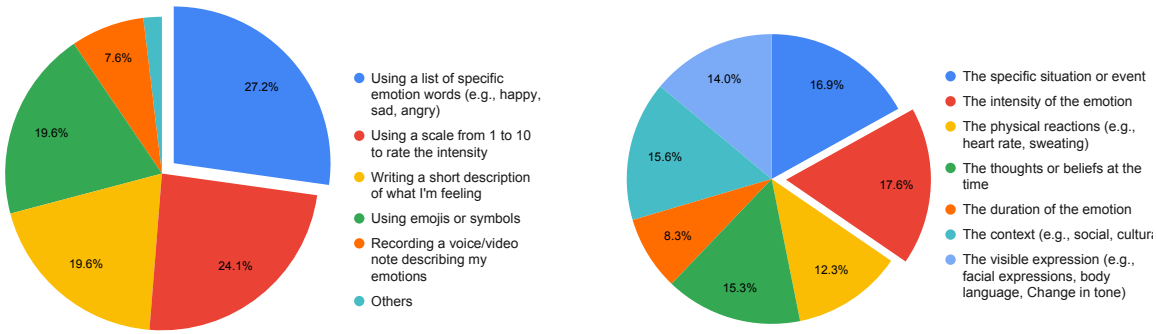
Guideline	Description
G6	**Collect Comprehensive Demographic and Medical Data**: **1.** Gather detailed demographic information (e.g., age, gender, education, socio-economic status, personality traits, and medical information) based on the specific needs of your research questions. **2.** Ensure that demographic data is relevant to the study objectives, is ethically approved, and captures any additional factors that may influence emotional responses.

Table 5.11: Guidelines for Pre-Data Collection Phase

emotions are deeply personal, and participants will find it challenging to share emotional details without assurance of privacy.

Derived Guidelines:

As reflected in our data, there were differences in participants' emotional awareness and attitude towards emotion management. This inspired our guidelines **G3, G4** on participant training and initial calibrations to ensure data collection methods are accessible and relevant to a diverse population. This is further crucial for collecting richer data. Discussions with domain experts further reinforced the need for participant training, and prior research has also shown that the varying ability in identifying and articulating emotions can adversely impact the quality of emotion data [205, 227]. Further to mitigate these impacts, we have added guidelines **G5, G6**. These guidelines are necessary for collecting detailed additional information to effectively contextualize emotion data [121]. Further, keeping in mind the wide range of hypotheses that inspire emotion data collection, we have added **G1**, which also includes exclusion-inclusion and screening guidelines inspired by data-centric AI and prior emotion data collection [228, 91, 36]. Diversity, necessary screening, alongside a comprehensive understanding of factors influencing emotional data, is crucial for ensuring the reliability of emotion assessments. Lastly, as expressed **(P3, Psychiatrist, FGD2)**, *"...the issue will be regarding the privacy part, how the data is being stored by the AI ... And you know who has access to it and how it is being used by the 3rd party. So overall, there are a lot of privacy-related challenges because there will be a lot of sensitive information.*



(a) Preference of annotation method.

(b) Factors that can impact annotations.

Figure 5.5: Survey results for attitude towards emotion annotation in daily life.

How we are tackling this will be an important point. And it should be communicated early on.", we have added **G2**. Elaborate informed consent was necessary alongside training and contextualization because stigma and privacy concerns, as visible in our data, can deeply influence the self-reporting behaviors. Consequently, together with our insights from data, alongside prior practices to collect quality data and emotion data collection methodologies, have informed our pre-dataset collection guidelines as provided in Table 5.11.

5.7.2 Understanding the needs of "*Data Source*": During Data Collection Phase (G7-G11)

Data Observations:

Within everyday settings, participants are typically prompted to annotate their emotions based on random, fixed time, or event-based triggers in response to changes in physiological or activity data [210, 61, 62]. These prompts often ask participants to complete surveys or questionnaires using predefined scales. However, these predefined surveys offer limited flexibility for participants to share additional context or express emotions to varying degrees.

1) Need for Adaptable Design: In contrast to these rigid methods, our survey data highlighted the diverse preferences participants have when it comes to emotion annotation (see Figure 5.5a). While 27.2% of participants preferred using an emotion list, 19.6% expressed a desire for an open-ended option to write about their emotions. Further, on examining participants' motivations behind their preferred annotation methods. *Ease of*

use emerged as the primary consideration (14.8%), closely followed by *expressiveness* (12.6%) - the ability to fully convey emotional experiences. Participants also mentioned *clarity* (8.9%) of methods and their ability to capture *emotional complexity* (8.9%) as important factors. These findings suggest that participants are seeking annotation methods that balance accessibility with expressiveness, allowing them to capture nuanced emotional states. Moreover, the relatively even distribution across annotation preferences points to significant individual variation. Further, in our interview data, we found similar patterns that highlighted the need for annotation methods that consider the transient nature of emotions. As one participant **(P30 - Interview)** explained- *"I would say the objective scales (likert, SAM or PANAS) will be easier use daily, but you should always give an option that if I am feeling extreme emotions — say if I'm extremely happy, extremely sad, or extremely angry —then there should be an option to write down or something."* Additionally, participants noted that during intense emotional moments, writing about the situation or their reactions would be easier rather than trying to identify and label specific emotions.

This finding suggested that participants may struggle to articulate intense emotional experiences, highlighting the need for structured guidance to help them navigate and understand complex emotional states. Mental health experts reinforced this insight, recommending adaptable annotation methods modeled after diary writing approaches. They specifically suggested incorporating probing questions about triggers, situational contexts, and emotional reactions. Such structured frameworks can significantly reduce the cognitive burden of subjective emotion annotation, particularly for individuals experiencing complex or overwhelming emotional states. Further, our interview participants noted that emotions with visible cues are easier to identify. However, identifying and articulating complex emotions (such as co-occurring, mixed, layered, or ambiguous emotions) is challenging. These emotions—like anxiety combined with fear or frustration intertwined with anger—were described as harder to pinpoint. These insights highlight the importance of designing interfaces that can facilitate the expression of both simple and complex emotional experiences. Such

an interface should provide participants with the required support, emotional vocabulary hints, reflective prompts, guided questions, and relatable analogies.

Question	Response	Count
Would you like to annotate your emotions daily?	Yes	28
	No	**47**
How easy do you find it to annotate your emotions daily?	Very difficult	5
	Difficult	22
	Neutral	**29**
	Easy	16
	Very easy	3
How frequently can you annotate your emotions?	Multiple times a day	17
	Once a day	10
	Few times a week	**20**
	Once a week	11
	Less than once a week	9
	Never	8
If you are going through a negative emotion, will you annotate?	**Yes**	**25**
	No	14
	Not Answered	36
If you are going through a positive emotion, will you annotate?	**Yes**	**21**
	No	20
	Not Answered	34
Do cultural or societal factors influence how you perceive emotions?	**Yes**	**44**
	No	31

Table 5.12: Survey Results on Emotion Annotation Practices and Perceptions

A few participants also suggested using more abstract and expressive methods for annotating emotions. They felt that predefined scales or specific words often limited how they could express their feelings. Instead, they proposed alternatives like sharing the songs they were listening to, quotes that reflected their mood, or photos of their environment. Some also mentioned sketches or free-form journaling. These methods, as explained by participants, allowed for a more personal and authentic expression of emotions, reflecting

the need for adaptable design. In addition to this, our survey data also revealed varying factors that can influence the identification of emotions (see Figure 5.5b). The intensity of emotional experience as felt by a participant emerged as the most frequently mentioned factor (53 mentions), closely followed by the specific situational context triggering the emotion (51 mentions). Contextual elements, including social or cultural factors, thoughts present during emotional episodes, physical sensations, and visible expressions (facial expressions, body language, vocal changes) are also mentioned as crucial. This even distribution of factors further reinforced that participants recognize emotion as a multifaceted phenomenon requiring multidimensional annotation approaches.

2) Participant's Agency, Learning and Participant-Aware Sampling: Our analysis of emotion annotation preferences and practices survey data (see Table 5.12) revealed significant resistance to daily emotion tracking, with 62.7% of respondents indicating reluctance compared to 37.3% expressing interest. This reluctance corresponds with perceived difficulty, as 36.0% found emotion annotation difficult, while only 25.3% considered it easy. Frequency data further reinforced these patterns, with only 35.9% of participants willing to annotate emotions daily, while 37.3% preferred weekly or less frequently. Further analysis of open-ended data and interviews revealed that participants wanted an annotation method that would prompt them according to their emotional intensities and pace. They also mentioned that the method should provide them feedback, insights, and an opportunity to learn from their data. Further, they mentioned that methods that only collect data without any learning engagements might not motivate them to annotate frequently. Participants also highlighted the importance of well-timed prompting methods per their personalized schedules. As mentioned by a participant (P26 - Interview) who uses an emotion logging application, the app frequently sends notifications when he begins working, distracting him. As a result, although he is willing to use the tracking technology, but he often does not annotate. This suggests that the varying needs of participants and assumptions, such as prompting users when they are not in motion, might not hold for everyone. For instance,

while some may find such prompts helpful during idle moments, others—like P26—may perceive them as intrusive, especially when they coincide with focused work sessions. Our interviews further highlighted that the timing and context of annotation must align with users' emotional states and willingness to engage alongside other contextual data, such as activity levels, behavioral cues, and physiological changes. Further, our data also revealed that many participants preferred non-digital alternatives, viewing digital tools as requiring extra effort and time. Participants highlighted the availability of real-life alternatives, such as writing with a pen and paper, sketching, sitting in silence, playing sports, or talking to friends, as the reason behind their preferences. This highlights the importance of participant-aware interventions incorporating user-agency in design and adapting to individual routines and preferences, rather than relying on one-size-fits-all strategies. Further, our survey data also revealed a significant component of cultural and societal influence on emotions. On deeper analysis, we found that these influences are shaped by negative connotations about sharing or expressing emotions, or stigma, and can hinder unbiased and balanced annotations. This suggested a need for an emotional literacy component in data collection methods.

3) Multi-perspective Assessments: Finally, our discussions with experts emphasized the importance of collecting emotional data from multiple sources, specifically for people with mental disorders or significant life events. They recommended combining self-reports with family members' input and regular evaluations by psychologists or psychiatrists. Emotional assessment is complex —even for professionals— so relying on a single source may lead to unreliable results. Experts also highlighted the value of integrating additional data streams. These included ecological activity data, social media behaviors, physiological signals, and other automated, objective measures. These sources can help complement and contextualize subjective self-reports. Experts also stressed the need for emotion assessment methods tailored to different groups. For the general population, tools should focus on promoting wellness and addressing everyday stressors. In contrast, more in-depth emotional assessments and professional evaluations are critical for clinical populations or those facing

Guideline	Description
G7	**Focus on Participant's Agency**: **1.** Use lightweight, non-intrusive wearable devices to avoid disrupting daily activities. **2.** Allow users to adjust the annotation frequency based on their preferences or schedules while ensuring a minimum frequency that balances data accuracy with preventing fatigue and disengagement. **3.** Set realistic expectations for emotional changes as per the research objective, for example conditions like depression don't show significant daily fluctuations, so daily recordings may not be necessary.
G8	**Develop Participant-Aware Sampling**: **1.** Trigger annotations by corroborating information on participants' characteristics (gathered in G5) such as, daily schedules, activity levels, physiological changes, and emotional profile while keeping G7 and research objectives in mind.
G9	**Design Adaptable Annotation Methods**: **1.** Offer participants the choice between structured annotation methods (e.g., SAM, PANAS) that use scales and unstructured subjective annotation methods (e.g., text, audio, images) as per their emotional intensity. **2.** For subjective annotations, adopt structured frameworks like the ABC model (Activating Event, Belief, and Consequence) to guide participants' responses. Alternatively, design LLM-based prompts [121, 119] customized to align with participants' unique emotional traits, as identified in steps G5, to provide tailored guidance. **3.** Provide participants with support in understanding complex emotions by offering tools such as emotion vocabulary lists, options to select multiple emotions simultaneously, visual aids like emotion wheels, reflective prompts, guided questions, and relatable scenarios or activity list to foster emotional clarity.
G10	**Incorporate Multi-Perspective Assessments**: **1.** Collect assessments not only from the participants themselves but also from trusted individuals in their support system, such as family members, peers, or mental health professionals, based on the participant cohort and study requirements. For example, clinical populations may require multiple assessments, whereas healthy individuals might need fewer. **2.** Integrate additional data streams, such as location, social media activity, phone usage, sleep patterns, and calendar events. **3.** Allow participants to select who and what data streams can contribute to their data based on their comfort and preferences.
G11	**Focus on Participant Engagement, Learning and Support**: **1.** Periodically reach out to participants to address any concerns, clarify expectations, motivate, and support. **2.** Use UI designs and prompts to encourage reflection, show growth, and provide supportive feedback to maintain engagement. **3.** Integrate interventions within the study that help participants enhance their emotional literacy over time. **4.** Integrate prompts that encourage reflection on positive outcomes or gratitude to offset the potential negative impact of recording difficult emotions. Additionally, provide access to mental health resources or emotional support tools for participants who may experience distress from self-reporting.

Table 5.13: Guidelines for During Data Collection Phase

significant life events to ensure accurate and meaningful insights. Further, our interview data also highlighted a set of participants who were skeptical about using technology for managing emotions and emphasized the need for a human touch. This finding reinforced the importance of incorporating multi-source assessment approaches, involvement of trusted people, and thoughtful data sharing mechanisms into emotion data collection methodologies. By integrating these elements, emotion tracking systems can complement rather than replace human interaction.

Derived Guidelines:

Our analysis revealed a preference for a balanced approach to emotion annotation. Participants valued having the flexibility to choose between structured, scale-based methods and unstructured, journal-writing methods based on the intensity of their emotions. The suggested need for this flexibility in our data collection approaches guided our addition of guideline **G9**. Within our data, it was also evident that participants frequently linked their emotions to specific environmental or situational cues. For example, loneliness was associated with the absence of companionship, stress with workload, fear with significant life events, and joy with time spent with loved ones. This underscores the need for tools that allow participants to articulate emotions by connecting them to contextual factors **(G9.3)**. Further, the need for user-agency to personalize the prompts per their schedules and emotional spectrum is also highlighted. Thus, designing methods with interfaces that could balance user-agency and participant-burden with data needs would be essential, guiding our inclusion of **G7**. Our data also highlighted a need to move beyond the context-aware sampling [124, 220], and adding a layer of participants' persona, cultural knowledge [229] to sampling strategies [121], as included in guideline **G8**. In addition to it, our discussion with experts and participants' interviews highlighted a need for adding multiple-perspective assessments and the option to multi-source data contribution [230] for improving the data quality. This supported our addition of guideline **G10**. Finally, our data observations sug-

gested a need for better participant support and components for improving emotion literacy over time in our data collection strategies for better participant engagement, guiding the inclusion of **G11**. Our detailed during data collection guidelines are presented in Table 5.13.

5.7.3 Learning from *Dynamic* Data: Post-Data Collection Phase (G12-G15)

Data Observations:

Following data collection, the post-processing stage involves several critical steps to ensure data usability and integrity. Our data highlighted several observations for the post-data collection stage.

1) Consistent Best Practices: The post-processing stage typically includes quality checks, consistent structuring, and preparation for data sharing to enable reproducibility and collaborative research [228]. While prior work highlights these best practices, data sharing remains inconsistent. For example, datasets such as WESAD [37], EEVR [3], and ASCERTAIN [43] provide not only the data but also baseline experiments to support emotion recognition research. In contrast, datasets like EMOGNITION [14] and GReX [57] focus solely on data release with quality checks, without offering baseline evaluations. While valuable, the absence of standardized benchmarks increases friction for downstream use and hinders fair comparisons across studies.

2) Handling Dynamic Data: Further analysis, as discussed in Section 5.7.2, revealed that participants have diverse needs and preferences regarding the sharing of their emotion data. Recognizing and addressing these needs is critical for fostering participant engagement and trust. Incorporating such preferences into data collection practices can enrich the resulting datasets, enabling the integration of information from multiple sources and annotation methods, including both structured and unstructured formats. Effectively managing this dynamic emotion data requires the establishment of a standardized data pipeline. This pipeline should include procedures for identifying and handling missing values, inconsistencies, and artifacts that could compromise data quality. Given the heterogeneous

nature of emotion data, validation must also extend to the annotation layer. This involves assessing the reliability of labels through cross-validation across various sources—such as physiological signals, self-reports, behavioral observations, and expert annotations where applicable. Such practices enhance the robustness and accuracy of the dataset, ultimately offering a more comprehensive and trustworthy representation of participants' emotional experiences. Normalization is another critical pre-processing step, particularly because individual differences, such as physiological signal ranges, emotional reactivity, or even environmental factors like temperature, can significantly influence emotional data. It is also important to document when and why normalization is applied to ensure transparency in the data processing steps. Normalization is crucial because it helps reduce bias and variation that could lead to inaccurate conclusions.

3) Grounding Emotion data: Our participant data highlighted that adopting more dynamic annotation approaches during data collection will likely produce annotations that differ from the standardized, fixed-scale labels typically used. These annotations will be more nuanced, context-dependent, and reflect participants' real-time emotional experiences. Domain experts supported this view. They recommended using both structured and unstructured data when labeling emotions. They emphasized that relying on just one type of data could miss essential nuances. When discussing emotional "ground truth," experts pointed out that exact accuracy is less critical than generating actionable insights. They stressed that aligning different data sources is a key indicator of accurate emotion labeling. Although these dynamic annotation approaches will capture richer and more authentic emotional experiences, they pose challenges for applying traditional AI models. This highlights the need to design newer systematic approaches to ground the emotion data.

4) Secure Data Handling: As discussed in Section 5.7.1, our participants have expressed a preference for keeping their emotions private or sharing them only with trusted individuals, such as family members or mental health professionals. Sharing emotional information with others or through technology was not the first choice for many participants unless

Guideline	Description
G12	**Secure Data Handling**: **1.** Store data securely using encryption and anonymization techniques, adhering to ethical guidelines (e.g., GDPR, HIPAA). **2.** Allow participants to request data reviews within a specified timeframe (e.g., 30 days), with researchers providing an overview instead of direct access to raw data to avoid misinterpretation and better confidentiality. Authorized researchers should handle deletion to maintain data security if deletion is requested. **3.** Clearly communicate any limitations on data review or deletion, such as once data has been anonymized or aggregated for analysis.
G13	**Data Quality Validation and Pre-processing**: **1.** Review datasets for missing values, artifacts, or inconsistencies. Depending on the study's goals and the extent of missing values, methods such as imputation, removal, or flagging of problematic data can be used to handle these issues. **2.** Cross-validate multiple data-sources (if available, G9 and G10) to improve reliability. **3.** Normalize data where individual differences (e.g., physiological ranges, personality traits) or environmental factors (e.g., time of day, activity) significantly affect results. Document when normalization is applied and why.
G14	**Holistically Analyzing and Grounding the Data**: **1.** Combine qualitative insights (e.g., text-based descriptions) and quantitative data (e.g., scale-based measures) to create multi-dimensional emotion labels that accurately capture the emotional experience within its context. **2.** Ground data on the reliability and relevance of the source (if G10 is applicable), such as expert assessments for emotional dysregulation, peer feedback for social interactions, and self-reports for subjective experiences. **3.** Combine psychosocial details (e.g., emotional traits, past experiences, daily routines) with emotion data to create a context-rich foundation for analysis and labeling, and document how these psychosocial factors impact emotion labeling to enhance transparency. **4.** Collaborate with domain experts to review and ensure the accuracy and consistency of grounded emotion labels.
G15	**Share Findings, Best Practices, Data Limitations and Usability**: **1.** Present key findings and any challenges faced during data collection, such as participant engagement issues, device inaccuracies, or contextual variability. Describe the study protocol in detail to ensure reproducibility. **2.** Highlight data limitations such as device reliability, data quality concerns, participant biases, or issues with ecological validity. **3.** Specify the intended AI applications for the data, like emotion detection, disorder diagnosis, or longitudinal tracking of emotional changes. Then, evaluate the data's suitability for each of these specific use cases.

Table 5.14: Guidelines for Post-Data Collection Phase

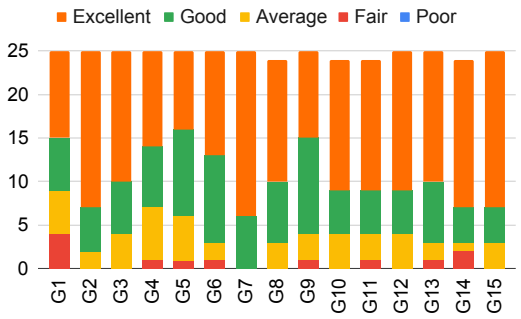
it significantly impacted their lives. Moreover, our survey (see Table 5.12), 62.7% of participants expressed reluctance to track their emotions using digital tools. Alongside time constraints and concerns about overthinking, emotional privacy emerged as a key reason for avoiding such tools. This highlights a strong stigma around tracking or sharing emotional data, as expressed by **(P21, Interview)**: *"If I am in real distress and I really want to get myself treated or understand the depth of my emotions, then I might provide access to my journal.* This suggests the need to maintain data security post-data collection.

Derived Guidelines:

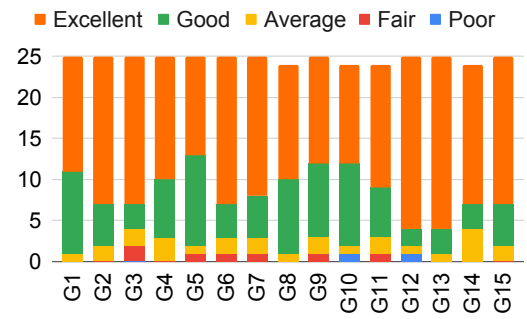
To address the privacy concerns, it is crucial that data collectors ensure participants feel confident in the secure handling of their data. Participants must also have the option to review or request deletion of their data, as outlined in **G12**. This guideline is essential because ensuring participants' trust is the foundation of ethical data collection, particularly when dealing with sensitive emotional information. By offering data review and deletion options, we respect participants' autonomy and privacy, addressing stigma and data misuse concerns. Subsequently, after data collection, validation of data quality [228] is an essential step, guiding the addition of **G13**. This includes support for heterogeneous data formats, standardized metadata schemas, and robust pre-processing pipelines that handle noise, missing values, and temporal inconsistencies. Further, the need for contextually grounding the collected dynamic data informed our guideline **G14**, which emphasizes the importance of holistically analyzing and grounding data. This guideline ensures that emotion labels reflect the complexity of participants' lived experiences rather than oversimplifying them for AI processing. Grounding the data involves assessing the reliability and relevance of its sources.

Recognizing psychosocial individual differences through standardized tools [231, 232] or through expert interpretations to contextualize emotional labels more accurately. Additionally, combining unstructured subjective responses (by quantifying them using either

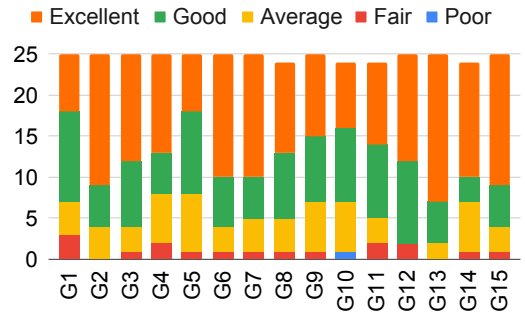
psycholinguistic analysis [233] or expert feedback) with structured, scale-based data provides a more nuanced and actionable grounding approach. This ensures that emotion annotations are participant-centered, addressing individual emotional experiences while also enabling the extraction of meaningful and useful emotion labels. Lastly, **G15** outlines the necessity of transparently presenting key findings alongside any challenges encountered during data collection, such as participant engagement issues, device inaccuracies, or contextual variability. By detailing these challenges, researchers offer transparency into the reliability and scope of the data, which is critical for ensuring the reproducibility and validity of the findings. Additionally, specifying the intended applications for the collected data—whether for emotion detection, disorder diagnosis, or tracking emotional changes over time—is crucial for guiding its use. Further, benchmarking and evaluating the data's suitability for the downstream task is equally essential for the future applicability of the dataset. Finally, our guidelines for the post-data collection phase are presented in Table 5.14.



(a) Clarity



(b) Usefulness



(c) Adaptability

Figure 5.6: The evaluator's scores for our guidelines.

5.8 Evaluation of Guidelines

The 15 guidelines were first internally evaluated by all the authors and fellow researchers/colleagues for clarity, usefulness, and adaptability. Following internal evaluations, we conducted an external evaluation involving 25 expert evaluators. These evaluators possessed expertise in emotion AI, physiological data collection, affective and ubiquitous computing research (detailed demographics presented in Table 5.15). This evaluation aimed to gather expert feedback on the clarity, usefulness, and adaptability of our guidelines. We employed purposive sampling [234] to select these experts, who were then contacted via email and social media platforms such as WhatsApp, Twitter, and LinkedIn. Each expert received a survey form encompassing informed consent, demographic questions (area of expertise and years of experience), and the 15 guidelines themselves, organized into three sections: *"Pre-data collection," "During data collection,"* and *"Post-data collection"*. For each guideline, we asked the experts to provide their ratings for clarity (description and communication), usefulness (practicality and goal achievement), and adaptability (real-world applicability across diverse contexts) on a 5-point Likert scale (1=poor to 5=excellent). Additionally, we asked the experts to provide qualitative feedback in the form of comments/suggestions for further refining the guidelines. Our method draws inspiration from Amershi et al.'s modified heuristic evaluation [235], where we adapted the principles of discount usability testing to evaluate our guidelines. Furthermore, to refine the guidelines, we conducted a descriptive analysis [236] of evaluator ratings as illustrated in figure 5.6.

Descriptive analysis of the expert evaluations reveals a strongly positive overall reception of the guidelines across all assessed criteria - clarity, usefulness, and adaptability. The guidelines were consistently rated highly for their clarity and usefulness, with the vast majority of evaluators rating them as "Good" or "Excellent" in these domains. While Adaptability also received predominantly positive ratings, a slightly higher proportion of "Average" and "Fair" scores in this category suggests a potential for refinements to enhance their perceived appli-

cability across diverse research contexts. Crucially, the consistent absence of "Poor" ratings across all guidelines and criteria indicated a robust framework without major perceived weaknesses. In addition to descriptive analysis, open-ended feedback was subjected to inductive thematic analysis [183], performed by the first author, to identify key suggestions for improvement. These suggestions, derived from expert feedback, primarily focused on enhancing clarity and comprehensiveness. Evaluators recommended adding more detailed explanations, illustrative examples, and definitions to make the guidelines more accessible. Furthermore, they emphasized the need to acknowledge the context-dependent nature of the guidelines, noting that their application may vary based on specific research objectives. In response to this feedback, we iteratively revised and rephrased the guidelines where needed to increase their adaptability to a wider everyday emotion research context. For example, Guidelines *#G1.1* and *#G1.2* were refined to explicitly state the importance of diverse recruitment while remaining aligned with specific study objectives. For *#G1.3*, to improve accessibility for interdisciplinary audiences, we incorporated references to screening tools like the Toronto Alexithymia Scale and added a definition of alexithymia. Similarly, Guidelines *#G3*, *#G4*, and *#G5* were revised to include examples and definitions, enhancing their overall clarity and broadening their applicability. Finally, all the authors then revisited and finalized the guidelines internally as presented in table 5.11, 5.13, and 5.14.

Category	**Details and Count**
Gender	Male = **13**, Female = **12**
Year of Experience	0–5 years = **12**, 5–10 years = **8**, 10+ years = **5**
Role	Researcher (Emotion AI/ Affective Computing/ HCI) = **23**, Data Scientist (Emotion AI) = **1** Researcher (Ubiquitous Computing/AI) = **1**

Table 5.15: Summary of Guidelines Evaluators.

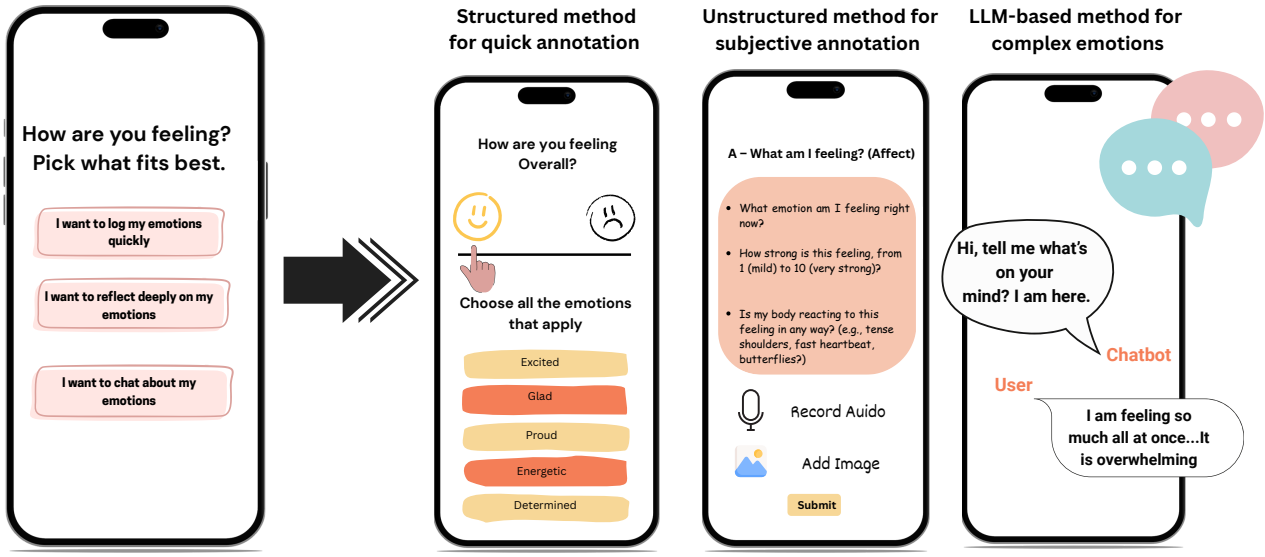


Figure 5.7: Visualization of the Participant-Centric Adaptable Annotation Approach. This approach offers flexible annotation options tailored to participants' emotional intensity and time availability. For quick annotations, a structured method using predefined emotion scales and lists is provided. In intense emotional experiences, participants can opt for a subjective, open-ended annotation guided by reflective questions. Additionally, large language model (LLM)-based support can facilitate meaningful annotation for users with lower emotional literacy.

5.9 Discussion

This section discusses how future research can leverage *AnnoSense* framework for designing participant-centric methodologies. First, we will discuss how to prototype tools based on our guidelines in section 5.9.1. Then, we will discuss the implications of pre, during, and post-data collection guidelines for future work.

5.9.1 Implementing *AnnoSense*: Designing for Participants

Building on the *AnnoSense* framework, in this section, we envision potential directions for prototyping new tools that can further support both real-life emotion data collection and the advancement of wearable and mobile-based AI solutions. To inform the design of potential prototypes, we reviewed the designs of currently available mobile applications and wearable technologies that support mood tracking, mental health monitoring, and

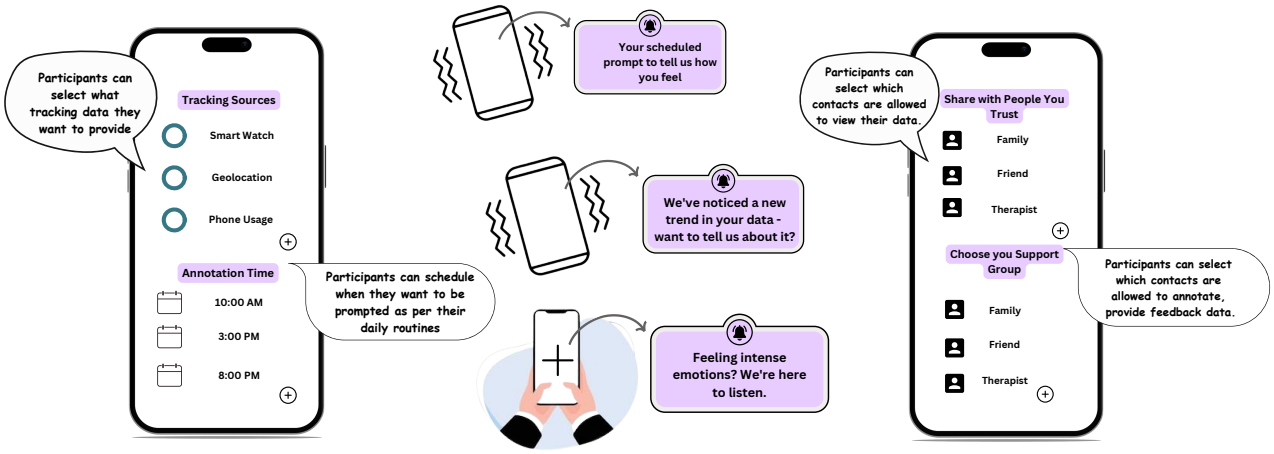


Figure 5.8: Visualization of Integrating Participant Agency into the Design Process. The first screen illustrates how participants can exercise agency by selecting preferred data sources and specifying suitable time slots for receiving prompts based on their individual schedules. The second screen presents three prompting strategies: (1) prompts delivered at user-specified times, (2) context-aware prompts triggered by physiological or behavioral indicators, and (3) user-initiated annotations during emotionally salient moments. To further support multi-perspective reflection, participants are also given the option to include input from trusted members of their support network.

emotion-related interventions. This included mood and journaling apps such as MoodPrism [237], MetricWire [238], Daylio [239], and MindLamp 2 [240]; well-being and mindfulness platforms like Headspace [241], Happify [242, 243], and Calm [244]; as well as wearable ecosystems including Apple Health [245], Samsung Health [246], Fitbit [247], Oura Ring [248], and WHOOP [249]. We also examined AI-supported mental health applications such as Wysa [250] and Woebot [251]. Additionally, we also reviewed well-know EMA frameworks like MindLamp [240], Beiwe [252], AWARE [253], PACO [254], mEMA [255], Experiencesampler [256], and MobileQ [257]. Our review identified several opportunities for designing future emotion annotation tools. Based on our review and the AnnoSense guidelines, we propose a set of prototype interfaces to accommodate users' needs.

1) Adaptable Annotation Interface: We propose a prototype for adaptive annotation interfaces that provides users with an opportunity to select an annotation method according to their emotional intensity or available time, as illustrated in Figure 5.7. This approach, in

Figure 5.9: Visualization of Participant Engagement, Learning, and Support Elements Integrated into the Design. The first screen displays personalized data insights derived from participants' inputs to foster self-reflection. The second screen offers curated, trustworthy information aimed at enhancing emotional awareness and literacy. The third screen illustrates how LLM-supported systems can be incorporated to provide contextually relevant guidance and emotional support.

contrast to traditional ESM methods, provides users with an option to select between quick scale-based annotations, detailed subjective annotations [121, 258, 118, 259], and chatbot-supported approaches [119, 220, 121]. It also offers users appropriate guidance through curated emotion lists based on their selected overall feelings during quick annotations. Diary-inspired reflection prompts based on psychological frameworks, such as the ABC model (A - Activating Event, B - Beliefs, C- Consequence) [260] or Cognitive Behavioral Therapy (CBT) thought record model (Triggering event, Automatic thoughts, Emotions, Evidence supporting, and Evidence against) [118, 220, 261, 262]. And an empathetic chatbot interface to support emotion annotations. Additionally, it supports a diverse range of users by offering options to record audio and upload images as part of their emotion annotations. Furthermore, these interfaces can be designed with an added redundancy layer to enhance data collection consistency. This redundancy can be implemented by adding a quick-annotation option in each annotation mode, ensuring a baseline level of information, followed by options to add deeper reflections. This structure helps maintain a consistent

data format while supporting varying user engagement levels. Moreover, the interface can also contain a curated list of activities that users can select to suggest the situational context of their data.

2) User-Powered Interface for Prompting and Multi-Source Assessments: We propose an emotion co-annotation platform where users can choose what data to share, set personalized prompting schedules, and invite trusted individuals to contribute their perspectives, as illustrated in Figure 5.8. These tools can allow users to set personalized prompting conditions (e.g., time-based, data-triggered [116, 91, 263, 264], or self-initiated) and control the granularity of the emotion data they wish to share. Overall, such systems can enhance user agency. Moreover, an additional Likert scale to provide feedback on the confidence of emotional assessments can also be incorporated [265] for users and other sources. This additional confidence assessment from various sources can help recognize the validity of data and provide users with an additional layer of reflection on their annotations.

3) Interface to Accommodate Learning and Support: We propose adding data insights, verified sources for emotional well-being and awareness guidance, and LLM-supported guidance systems to the data-collection applications. These design elements can enhance user engagement by motivating them to understand themselves better, as illustrated in Figure 5.9.

Lastly, future work can explore a range of mood tracking and self-reflection applications that prioritize adaptability, user privacy, data sharing, and personalized emotional insight. For example, iMoodJournal [266] allows users to select their mood from an extensive list of emotions and supplement entries with journal notes, images, and location tags. It supports mood log sharing while maintaining a strong emphasis on data privacy. Similarly, Apple Health [245] offers a "State of Mind" mood logging feature, which prompts users to first categorize their mood as positive or negative, then select specific emotions with the option for multiple selections. Users can also add contextual notes and identify potential causes of mood changes, such as activities and relationships, while choosing between

real-time and daily summary logging. Additional examples of feedback-oriented platforms include Mindsera [267], an AI-powered journaling app that provides emotion analysis and personalized suggestions to guide self-reflection and promote mental fitness. Another example is Daylio [239], which offers a quick and streamlined interface for logging moods and activities multiple times throughout the day. Overall, platforms that integrate mood tracking with everyday lifestyle have the potential to generate more accurate, actionable emotion data and offer valuable insights for both research and personal well-being.

5.9.2 Understanding the Implications of Pre-Study Guidelines

> *"I am made of little rooms full of thoughts, emotions & memories. You cannot define me by listening to me once. I'm too complex."* [Source: Unknown, Credit: Pinterest]

Prior research has often relied on one-way communication for everyday emotion data collection, borrowing heavily from traditional lab-based methodologies. However, everyday settings differ significantly from lab environments, as they lack the level of control typically available in laboratories. This lack of control introduces challenges in ensuring the quality and reliability of the collected data. To overcome these challenges, it is crucial to prepare a data collection pipeline that is robust to noise and bias in the real world. Beyond the lack of control, our findings also highlighted the diversity in participants' attitudes toward collecting and sharing emotional data, influenced by varying levels of emotional literacy. This diversity has been shown to impact emotion tracking among participants in previous studies [224, 211, 99, 94, 100]. Drawing from data-centric AI guidelines [228], ethical considerations for emotion AI [268], and past literature on emotion tracking, our findings emphasize the need for careful pre-preparation which involves: 1) **careful selection of participants** (G1), by clearly defining the inclusion and exclusion criteria, selecting participants from diverse backgrounds [269], and screening for possible conditions that could impact the data, 2) **preparing the participants** by elaborately informing (G2, G12), and training

(G3–G4) them about how to interact efficiently with devices and annotation methods involved in data-collection methodology and its benefits, 3) **understanding the participants emotional profile** by performing elaborate psycho-social profiling (G5) and comprehensive demographic data collection (G6). Including these steps in the data collection pipeline can enhance participant engagement but also minimize errors, reduce hesitations, and foster a sense of collaboration between researchers and participants, which was often missing in prior methods [62, 61, 270, 167]. This careful pre-preparation will ensure an inclusive experience for participants of varying levels of emotional awareness. Furthermore, gathering psycho-social profiles and comprehensive demographic data will allow researchers to collect broader context about emotional responses missing in prior context collection that was limited to activity levels, basic demographics, and personality traits [43, 62] and will further help researchers to tailor the data collection process to the participants' emotional traits and lifestyles.

5.9.3 *Go with the Flow*: Understanding the Implications of During Data-Collection Guidelines

Prior research on emotion data collection has typically focused on two approaches for collecting emotion data in real-life settings. 1) Designing real-life emotional scenarios, such as work-related stress [270, 167], group entertainment [57], or driving stress [87]. 2) Complete in-situ settings- where data collectors rely completely on participants' willingness to complete ecological momentary assessments (EMAs) or emotion questionnaires [62, 61]. These approaches often lead to data of a specific emotional scenario or incomplete data with limited contextual information. However, our findings emphasized the **inherent diversity in participants' attitudes** towards tracking and sharing emotion data and suggested designing sampling strategies that can accommodate this diversity. To address the challenges posed by the diverse engagement styles, we propose designing annotation methods that are tailored to the specific needs and capabilities of different individuals while also being flexible enough to support a broad range of participants. We recommend designing adaptable

methods that can accommodate the varying needs of people, as recommended in section 5.9.1. Additionally, for participants with lower emotional literacy, researchers can frame emotional tools as practical aids rather than self-reflective interventions (e.g., stress relief or productivity enhancers) to increase engagement. Subsequently, for clinical participants or participants dealing with emotional trauma or other life-changing events, researchers can investigate a multiple-assessment approach [230]. Furthermore, our findings show that participants preferred using objective methods (such as scales) and lower frequencies on neutral days, while subjective methods (like written descriptions) and higher frequencies were favored during periods of intense emotions. However, present approaches such as ESM (In-the-moment annotation) and DRM (after-the-fact annotation) [68, 208, 12, 209, 117] often overlook this fluidity in emotional experiences. These methods use either a fixed scale (e.g., SAM, Likert Scale) or questionnaires (e.g., STAI, PHQ-9) to capture emotion ratings within fixed or random time periods. This often doesn't provide users with an opportunity to label as per emotion intensity, thus leading to datasets that fail to capture the dynamic and contextual aspects of emotional experiences, instead treating emotions as discrete snapshots, to overcome these challenges we recommend designing systems that can adapt to users changing emotional landscapes (G9), as shown in Figure 5.7. Moreover, the timing of the annotation prompt can significantly affect the precision of annotations (G7, G8). For instance, in-the-moment annotation requires participants to assess and record their emotions as they occur, which can capture more immediate and authentic emotional states. However, this method can be cognitively demanding, as participants need to be aware of their emotions while balancing other activities in their environment [68, 208, 12]. In contrast, after-the-fact annotation allows participants to reflect on their emotional experiences once they have passed, which can provide a more thorough and considered response. However, this retrospective approach comes with its own cognitive challenges: memory bias and difficulty in recalling the intensity or nuances of past emotions accurately [209, 117]. This can lead to data that may not fully reflect the emotional state experienced

at the time, impacting the validity of the data for training AI systems. We recommend future works to design participant-aware sampling techniques and adaptable annotation methods that combine closed-end and open-ended questions (G9) as shown in Figures 5.7 and 5.8. Further, an interface for adding contextual metadata, such as associated events or environmental factors, alongside emotional labels [271, 270], should also be added. Finally, our data also shed light on the **psychological influences that self-reporting emotions** can have on participants' daily lives. While self-reporting can foster self-awareness and provide emotional patterns, it can also influence the user's emotions in unintended ways. For instance, users shared that recording subtle negative emotions can amplify overthinking. Conversely, documenting positive emotions can foster a sense of gratitude. To overcome potential influences, we recommend designing supportive and non-judgmental annotation techniques (G9, G11). For example, adding reflective prompts to encourage users to frame negative emotions constructively, like "What can this feeling teach me?". Further, LLM-based structured journaling activities, guidance for mindfulness or relaxation exercises, and references to resources during distressing periods can be added to the applications [121, 272]. Further tools can integrate features that allow users to record emotions without immediate analysis and then review entries after a period of detachment, or ask participants to note a small positive event or something they feel grateful for (G11), can be added, as shown in the prototype figure 5.9.

5.9.4 Moving beyond the Traditional Data Modeling: Implications on Post-Data Collection Approaches

Our findings emphasize that the concept of **"emotional ground truth"** extends far beyond the survey responses typically gathered through standard questionnaires. However, current datasets often assume a universal definition of emotions and one-to-one mappings between emotion data and filled surveys [71, 5, 205], to label emotion data. Moreover, prior work on developing models for physiological emotion data often applies simplistic labeling

approaches like categorizing emotions into discrete groups based on objective labels, such as predefined emotion categories (e.g., happy, angry) or scales (e.g., 1 to 5). These methods often do not use additional contextual data [37, 43, 44, 36], while modeling the AI algorithms leading to the development of models that cannot be adapted in real-life [89] or clinical settings [273, 274]. The continued use of such approaches can be attributed to several factors: 1) Simplifying emotion categorization reduces the complexity of emotion recognition models, making them easier to develop, train, and implement. 2) Discrete emotion categories are easier for participants or experts to label, lowering the annotation burden. 3) Standardizing datasets based on these categories facilitates generalization across various AI applications, such as sentiment analysis and video emotion recognition. 4) The influence of early psychological theories, such as basic emotion theory, has strongly shaped these practices. However, our findings based on interviews with domain experts challenge these assumptions. Experts argue that actionable insights and contextually relevant data should take precedence over overly generic labeling (G13, G14). They suggested that emotional ground truth is not a simple, one-to-one mapping from data to labels. In fact, it's a composite representation that varies according to the user profile.

Insights from both participants and experts have shaped our guidelines for collecting emotion data that is dynamic, layered, and actionable. Unlike traditional methods, our approach captures emotions in real-time and across varying contexts, resulting in data that is fundamentally different in structure and complexity. This shift highlights the need for new labeling and validation techniques that can accommodate the richness and variability of the collected data. Traditional emotion datasets often collect inputs, such as single-point self-reports, task-based annotations, or expert labels, resulting in relatively uniform data structures that are easy to label. These inputs are then reduced to binary or discrete categories (e.g., "happy" or "stressed") by binning self-reported or task-driven labels to fit downstream tasks. For example, in the WESAD dataset [37], emotional states are classified into stress versus no-stress categories based solely on experimental stimuli, without incorporating

participant self-reports. Similarly, GLOBEM [59] focuses on depression detection as a downstream task, and ASCERTAIN [43] performs arousal-valence classification based on self-reports. In contrast, our approach captures emotion as a dynamic, evolving state, influenced by contextual, physiological, and subjective factors as discussed in section 5.7. This results in data that is more variable and multidimensional. For instance, each emotional annotation in our system may include a combination of quick scale ratings, emotion labels, option text/audio/image data, confidence scores, and contextual metadata. The structure and depth of these annotations can vary based on the user's engagement and the intensity of the emotional experience. Such variability introduces both opportunities and challenges: while the data offers a more accurate and holistic view of emotional states, it also complicates traditional labeling and validation pipelines, which typically assume uniform input formats.

To effectively handle our dynamic data, we propose a set of new validation schemes that go beyond traditional practices. **1) Triangulated Validation**: In this technique, we can assign a final emotion score or label by combining information from multiple sources, such as scale-based self-reports, emotion-list, physiological signals, AI-generated or text annotations, images/audio annotations (optional), and contextual, peer, or expert feedback. Each of these sources can first be evaluated for coherence and reliability in a given context, and a confidence score can be assigned to them. For example, if a user provides a confident self-report, it might carry a higher weight of 0.9, while physiological signals with strong indicators could be weighted at 0.8, and AI-based reflections with uncertain text data might be assigned a lower weight, such as 0.5. All emotion representations are then aligned into a common format, such as a valence-arousal score or a set of discrete emotion categories. Finally, the final label can be computed using a weighted aggregation, such as a weighted average for numerical scores or a confidence-weighted majority vote for categorical labels. This ensures that more reliable sources contribute more to the outcome. Overall, this approach allows for a more robust and context-aware emotional label, addressing the limitations of relying on any single data source. **2) Semantic Validation**: Given

the redundant nature of information collected through multiple methods, such as scale-based self-reports, emotion names from the list, or optional text, should be validated for semantics. This means making sure that elements like emotion labels, confidence scores, and multimedia content (such as text, images, or audio) align coherently. For example, if an annotation includes the emotion label "joy," but the accompanying text expresses sadness or the image shows someone crying, a mismatch may need to be addressed. Similarly, if users rate their emotional intensity as very high but give a very low confidence score, that inconsistency could indicate confusion or noise in the data. This form of validation can add a layer of reliability in collected data, which is often missing in traditional data. **3) Contextual validation**: This involves checking whether the data fits logically within the context in which it was collected. This validation technique is similar to traditional approaches of checking the data contextually. **4) Annotation Agreement Validation and Co-Development:** This validation focuses on assessing the consistency of emotion annotation when multiple annotators (participants, experts, and peer groups) are involved in labeling the same content. Since emotions are highly personal and subjective, it's common for different users to interpret the same situation differently. This step helps identify how much agreement or disagreement exists among annotators. Techniques like inter-annotator agreement metrics (e.g., Cohen's Kappa or Krippendorff's Alpha) can be used by future works to quantify the level of consistency across annotations. This approach also provides a framework for emotion data collectors, mental health professionals, and emotion AI experts to co-develop and evaluate new tools and methodologies with users. By bringing together multiple stakeholders in the data validation process, the resulting systems can benefit from diverse expertise: users' lived experiences, clinicians' domain knowledge, and technical experts' implementation capabilities. This collaborative approach ensures data collection tools are not only technically sound but also clinically relevant and ethically implemented. Consequently, these validation techniques can validate emotion data more robustly, and they also align with domain experts' guided strategies of finding *congruence*

in emotional assessments. Further, they provide a platform to add clinical and participant insights to traditional data, thus adding an opportunity for co-development with experts while keeping participants in the loop. Lastly, our findings underscore the critical need to design AI algorithms that prioritize actionable outcomes [91], such as identifying meaningful patterns—like recurring emotional states—over simplistic emotional categorizations [101, 100]. This shift is essential for developing systems that align more closely with real-world applications. For example, in therapeutic settings, recognizing patterns in emotional states over time can help identify triggers or trends in mental health, providing valuable insights for personalized interventions. However, for clinical settings, tracking changes in symptoms can be targeted over nuanced emotional changes. By focusing on actionable outcomes, algorithms can also provide deeper insights for decision-making, enabling stakeholders to address the underlying causes of emotional responses rather than just classifying emotions into predefined categories. This approach moves away from a rigid framework of labeling emotions, embracing a more dynamic and context-sensitive model of emotion tracking.

5.10 Limitations

This study aimed to explore people's attitudes and preferences toward tracking and monitoring emotions in everyday settings for emotion AI data collection and interventions, and provide a set of guidelines for future emotion data collection methods. A limitation of our work was that most of our participants were well-educated, tech-savvy individuals familiar with AI, emotion monitoring, and wearable technologies. Thus, our findings might not generalize well to people with low literacy and less experience with emotion tracking. We also recognize that our findings might be influenced by the participants' demographics, group composition, and backgrounds, since all our participants belonged to the same cultural background and country. To address this, we have included a diverse audience of users and non-users of emotion-tracking technology with varying levels of technological familiarity, emotional awareness, and demographic profiles (age, gender, occupation, education).

Moreover, it is also important to recognize that different research objectives may encounter unique challenges when adapting these guidelines. Thus, we recommend that future work customize these guidelines to their specific needs for better adaptability. For instance, participant training and psycho-social profiling can be significantly more challenging when working with clinical populations compared to undiagnosed, healthy counterparts. Individuals with diagnosed mental disorders may require tailored approaches to ensure ethical considerations, comfort, and engagement throughout the data collection process. To address these complexities, we recommend engagement with mental health professionals to navigate the sensitivities associated with clinical populations. This is particularly crucial given that many emotion monitoring interventions are designed to target individuals with diagnosed mental health conditions. By including professional supervision, researchers can better align their methodologies with the needs of clinical audiences, creating a more inclusive, ethical, and effective data collection process [91] tailored to diverse participant groups.

5.11 Summary

This chapter examined participant perspectives on physiological emotion data collection across both controlled laboratory environments and everyday contexts, highlighting how human factors shape annotation quality and dataset reliability. In Part 1, a VR-based laboratory study with 37 participants revealed a clear disconnect between expressed emotions, physiological responses, and participant-provided annotations. The findings showed that individual perceptions, interpretation of emotional experiences, and experimental design choices significantly influence how emotions are elicited and labeled. These insights underscore the need to move beyond simplistic experiment design toward a more participant inclusive experiment protocols and labeling technique to improve data validity. Building on the data collected in part 1, we introduced a participant-aware modeling approach for emotion labeling and classification in Chapter 5 that improves generalization by explicitly accounting for participant variability during model development.

Part 2 extended this investigation to real-world contexts by exploring the perspectives of the public and mental health professionals on everyday emotion annotation. The study identified several critical influences on annotation quality, including the fluid and context-dependent nature of emotions, social stigma surrounding emotional disclosure, and varying levels of emotional literacy. These factors highlight the limitations of traditional questionnaire- and scale-based approaches when applied outside controlled settings. Based on these findings, the chapter introduced the *AnnoSense* framework, which outlines principles for more holistic, context-sensitive, and participant-centered emotion data collection. Furthermore, drawing on the guidelines derived from AnnoSense (discussed in Chapter 6), we designed an application prototype to support more user-aware, contextually grounded emotion data collection. Collectively, the two studies emphasize that improving physiological emotion recognition requires not only advances in sensing and modeling but also deeper engagement with participant experiences, motivations, and contextual realities.

Chapter 6

Designing for Real Life: Participant-Centric Emotion Data Collection

In recent years, mobile- and wearable-based ecological momentary assessments (EMAs), experience sampling methods, and digital phenotyping systems have been widely adopted to collect emotion self-reports in real-life settings at regular intervals throughout the day [275, 276, 277, 60, 128, 258]. These approaches aim to capture multiple snapshots of emotional experience by prompting users repeatedly in situ. However, much of this prior work emphasizes maximizing the quantity of collected annotations and is largely designed around researchers' data needs rather than participants' lived experiences [2, 96]. As a result, participants are often presented with fixed questionnaires or scale-based prompts that constrain how emotions can be reported [277, 58, 11, 88].

However, my previous exploration in earlier chapters and other participant-centric works have highlighted key limitations of these approaches, noting that emotional experiences are inherently transient and vary in intensity over time [275, 2, 1, 3]. Static, scale, or questionnaire-based methods are often insufficient for conveying the full range and nuance of emotional experiences. For example, complex emotional states, such as sadness intertwined with or masked by anger and frustration, often require more expressive space and reflection to articulate than more straightforward emotions like happiness. Additionally, participants and mental health professionals have noted that individuals often lack precise emotional vocabulary to express themselves when dealing with complex, overlapping, and abstract emotions, further limiting the effectiveness of predefined scales [2]. Beyond limiting participant expression, these constraints also affect downstream AI models trained on these restrictive labels. As these restrictive data labels often fail to reflect participants' true emotional landscapes, which vary widely in intensity, reactivity, and expressive range across individuals [2], impacting the performance of emotion recognition models in real-

life settings [3]. Moreover, by treating emotional experience as homogeneous, existing systems overlook inter-individual variability, resulting in training data that may be poorly representative of how emotions are actually experienced across users with varying emotional profiles in everyday life.

With the advent of large language models, recent work has started exploring AI journaling or end-of-day diary systems to support emotional reflection and mental well-being [120, 258, 121]. While these systems demonstrate the value of open-ended, narrative expression and LLM-based scaffolding, they are primarily designed to support mental health and reflection and have not been used for systematic emotion data collection in the past. As a result, they were mostly designed for participants' reflections, with limited attention paid to their downstream utility for collecting more in-depth emotion data. Moreover, these approaches typically rely on a single interaction modality, such as text-based journaling or conversational chat interfaces, offering users limited flexibility in expressing emotions across intensity and time availability. Thus, this one-size-fits-all design overlooks the fact that users' emotional intensity and expressive capacity vary across moments and individuals. Furthermore, across EMAs, AI-journaling or dairy studies, and other emotion data collection tools, prompting schedules are often determined by researchers using different methods that can be either context-centric or interval-based, offering users with limited agency to choose when to report emotions, depending on their availability and schedules. Taken together, these limitations underscore the need for more dynamic emotion annotation tools that can adapt to users' emotional states and availability, while also capturing contextual and experiential details often missing from static annotation techniques. Developing adaptable annotation mechanisms is further critical, as recent work [278, 3, 50] has shown that elaborate self-reports and contextual data, alongside sensor data, can enable more accurate physiological emotion recognition models.

Overall, our prior discussion points highlight the need for more flexible participant-centric, multimodal approaches to collect emotional self-reports that can balance user burden

with opportunities for expression while remaining scalable in real-world deployments [279]. Motivated by this, we designed a multimodal emotion self-reporting prototype system that supports flexibility in how users express their emotions. Furthermore, to explore how such flexibility shapes user behavior and emotion-reporting data, this paper presents a one-week feasibility study using our EMA prototype (see section 6.1 for more details). We deployed our application in the field for seven days with 33 participants. The study is guided by the following research questions:

- **RQ1.** How do users engage with multimodal emotion logging in everyday contexts?
- **RQ2.** How does multimodal logging support varying levels of expressive elaboration, emotional complexity, and contextual grounding in emotion self-reports?

Our formative in-field deployment investigates how users engage with an EMA system that supports flexible prompting schedules and multiple modalities for in-situ emotion logging. Furthermore, our findings demonstrate that supporting modality switching is not merely a usability enhancement but a mechanism that enables users to adapt expression to situational constraints and cognitive load. In summary, our results provide empirical evidence that multimodal and user-adaptive emotion logging systems are better suited to capturing the heterogeneity and situated nature of everyday emotional experiences than single-modality approaches.

6.1 Application Design: Overview

Formative user-centered research [280] and behavioral theories such as self-determination theory [281] consistently highlight that intrinsic motivation is strongly linked to perceived autonomy and agency. In data collection contexts, providing users with greater control over when and how they contribute data has been shown to improve both engagement and willingness to share information [282, 283]. In addition, prior work emphasizes the importance of adaptive and personalized designs to better accommodate diverse user needs

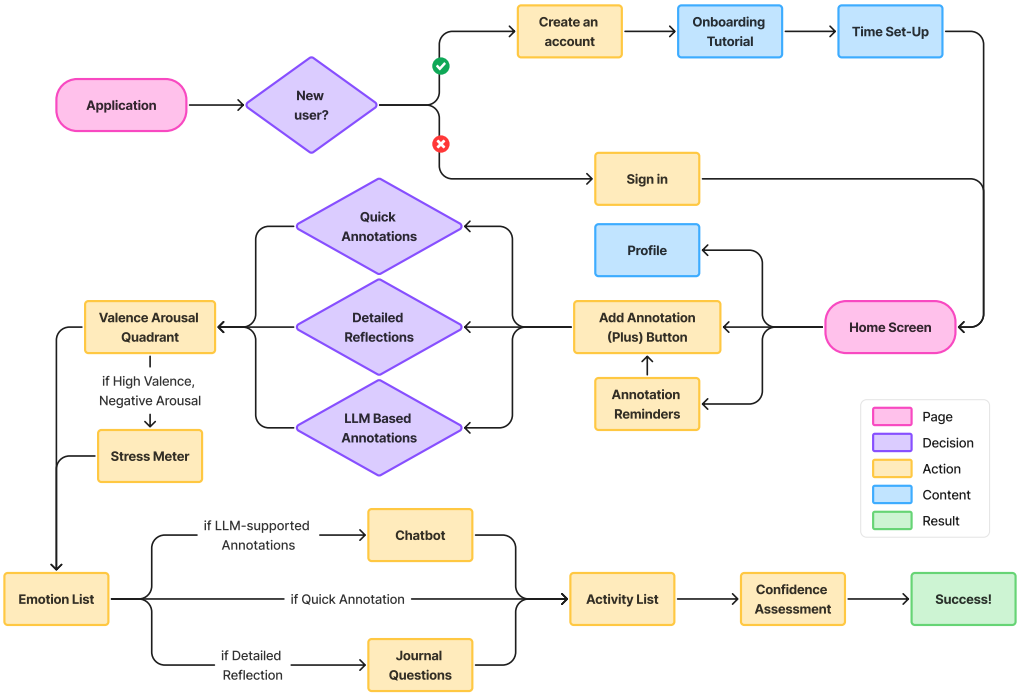


Figure 6.1: An Illustration of our Application Flow (Best viewed in color).

and contexts [284]. Guided by these principles, we developed our prototype. We took design insights from prior user-centered emotion logging systems [285, 2, 286, 1, 275, 118], as well as commercial applications such as Apple Health, Daylio, and Moodflow, for developing our prototype. Our prototype integrates the following set of design features aimed at supporting flexible and multi-modal emotion reporting:

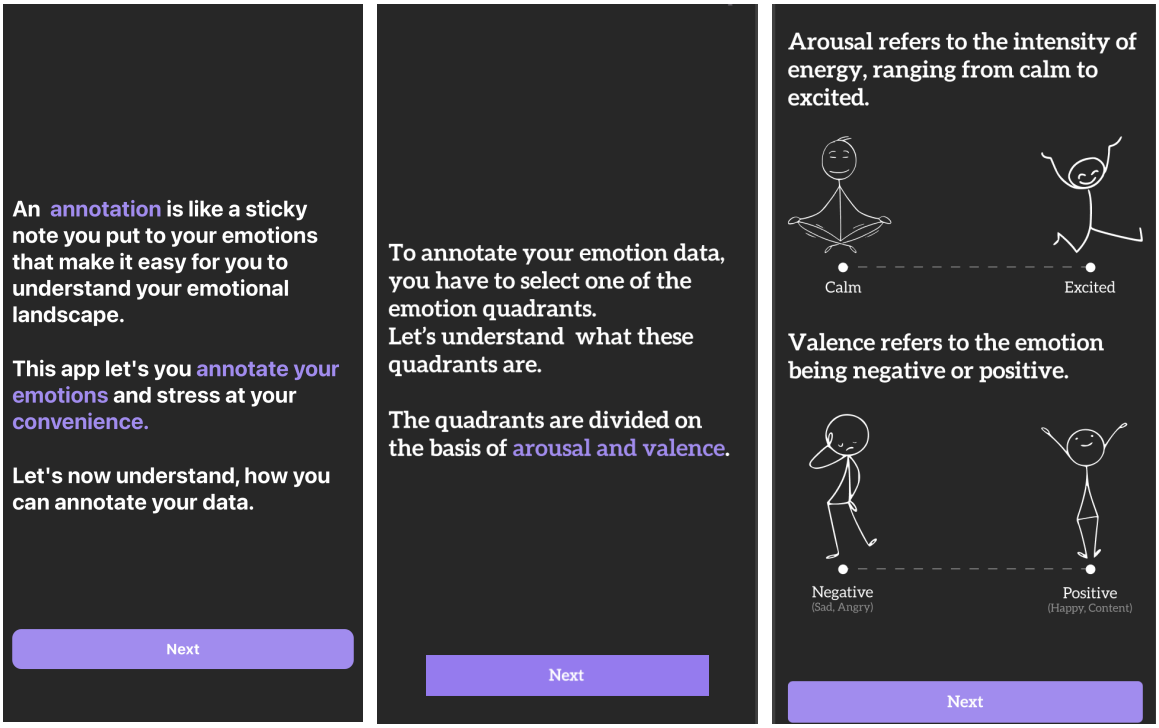
1. **User-configurable prompting and impromptu logging**: The system supports both user-defined reminders and on-demand logging, enabling individuals to record emotions at self-selected times as well as in-the-moment self-reporting. This design choice supports flexibility, accommodates varying daily routines, and reduces reliance on externally imposed schedules.

2. **Multi-modal self-reporting options**: To accommodate diverse expressive needs, the system offers multiple reporting modalities. This allows users to select the mode that

best aligns with their context, cognitive load, and preferred level of expression, while also accounting for variability in emotional vocabulary and articulation.

3. **Supporting contextual reporting**: The system incorporates multiple supportive mechanisms to enable richer contextual annotation.

Our application design is illustrated in Figure 6.1. More details on the technical implementation are provided in Appendix D.5. Next, we present the system design along with the underlying design rationales.

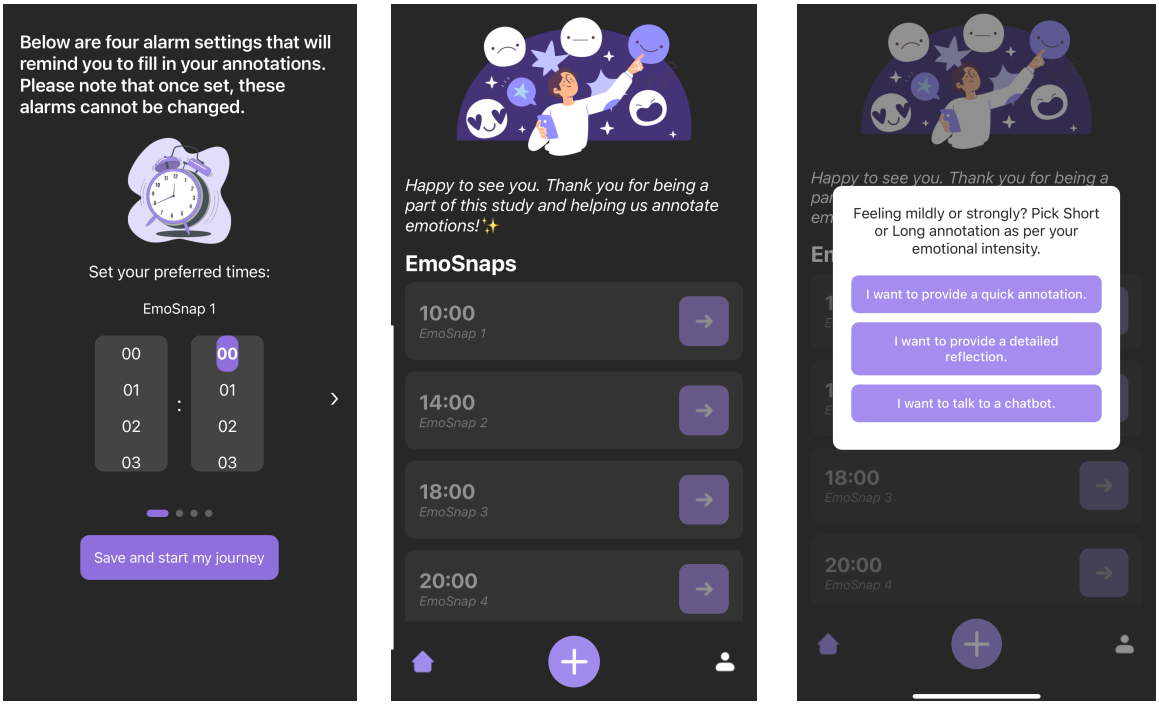


(a) (b) (c)

Figure 6.2: Tutorial screens guiding users through the self-reporting process: (a) introduction to annotations, (b) and (c) explanation of the Valence–Arousal quadrant with examples (Best viewed in color).

6.1.1 Onboarding Module

The onboarding module consisted of: (1) **Interactive Tutorial**, which introduced the concept of emotion annotation and the scales used in our application, including arousal–valence



(a) (b) (c)

Figure 6.3: (a) Time-slot selection screen where users choose four daily notification reminders. (b) Home screen displaying the selected times and a floating action button for on-demand emotion logging. (c) Modal window shown when the button is tapped or a notification is opened, offering three emotion self-report options (Best viewed in color).

dimensions, the four-quadrant system derived from it, and the perceived-stress scale (see Figure 6.2). The tutorial was added to support participants' training and reduce interpretation biases [287], and (2) **Time-slot Selection** (see Figure 6.3a), where participants scheduled four daily slots aligned with their routines. Notifications were delivered at these times, integrating self-reporting into daily life. Following prior works [19, 20], we limited reminders to four per day to balance data requirements with participant burden.

6.1.2 Home Screen and Modality Selection

After completing the tutorial and selecting their slots, participants arrived at the home screen, which displayed four prescheduled slots represented as "Emosnaps" (see Figure 6.3b). Each Emosnap was locked by default and unlocked one hour before the scheduled time,

remaining available for one hour afterward to allow flexible yet temporally relevant reporting. Once the window closed, the Emosnap was locked again to preserve temporal fidelity. We anchored prompts to user-defined routines to leverage established behavioral patterns, supporting habit formation through temporal consistency while preserving user autonomy in scheduling. The home screen also featured a floating action button for spontaneous entries, enabling participants to log their emotions whenever they felt something significant happened. This dual structure, combining scheduled prompts with on-demand logging, was therefore intended to support both structured recall and event-driven reporting, thereby aiming to increase ecological coverage of emotional experiences while preserving user autonomy and reducing interaction friction.

6.1.3 Self-reporting Methods

To study our research questions, in our prototype, we included three self-reporting approaches (see figure 6.3c): Quick Mode, Detailed Reflections, and LLM-supported Annotations. Each of these approaches was designed to provide users with multiple modes to express their emotions, depending on their context, cognitive load, and expression needs.

Quick Mode

The inclusion of this scale-based mode was motivated by its widespread use in traditional EMA systems as a lightweight and standardized mechanism for emotion self-reports. However, to address its known limitations, we extended it with a multi-select emotion list alongside the arousal–valence scale. This design allows users to report multiple concurrent emotional descriptors, thereby better capturing mixed, overlapping, and co-occurring affective states that are difficult to represent within a purely dimensional framework. In addition, the multi-select context list and confidence rating were incorporated to enrich each report with contextual and subjective uncertainty information. Together, these elements were intended to increase the interpretability of self-reports by providing additional information

for downstream modeling and analysis. It takes approximately one minute to self-report in this mode. The flow of this annotation method is as follows:

1. **Quadrant Selection**: Users start with the arousal–valence quadrant screen (based on Russell's Circumplex Model of Affect [134]. This color-coded screen has four quadrants, each representing a combination of arousal and valence. For example, the red quadrant (high arousal–negative valence) reflects emotions like anxiety or anger, while the yellow quadrant (high arousal–positive valence) represents excitement or happiness. The design was informed by prior emotion-assessment tools, like the SAM [179], Affect Grid [72], Geneva Emotion Wheel [288], and the Photographic Affect Meter [112].

2. **Stress Scale** (conditional): After the quadrant screen, users were shown the perceived stress scale (PSS) only if they selected the high arousal–negative valence quadrant, which corresponds to stress-related emotions. This scale, adapted from the widely used 10-item PSS [289], was included to measure stress intensity, often collected separately in emotion datasets [37]. To improve usability, each numerical value was paired with a short descriptive phrase, for example, seven represents moderately high stress.

3. **Emotion List**: Next, users were shown a curated list of emotions corresponding to their selected quadrant (see Table D.2). They could select multiple labels to describe their states (see figure 6.4a). This screen was designed to provide both a guided vocabulary and the flexibility to capture concurrent or overlapping emotions alongside quadrant labels [2].

4. **Contextual Factors**: After selecting their emotions, users chose from a predefined list of activities they had engaged in since their last log (see Table 6.1). The list, covering domains such as health, physical activity, medication, and environmental influences [63], was designed to help users reflect on possible triggers of their emotions.

5. **Confidence Rating**: Finally, users can rate their confidence in the accuracy of their annotation on a 5-point scale (see Figure 6.4d). This step was designed to encourage self-reflection [265] while also providing researchers with an additional indicator of data reliability [2].

Detailed Reflections

The detailed mode was designed to capture more complex and context-rich emotional experiences that cannot be adequately expressed through quick mode (see Figure 6.4b). This design draws on principles of reflective practice in HCI, where prompts can facilitate deeper sense-making while maintaining consistency across entries. These prompts can help users articulate their thoughts, situational triggers, and emotional interpretations without facing the cognitive burden of open-ended reflection. After completing the initial scale-based categorization (quadrant selection, stress scale (if applicable), and emotion list), users were presented with four open-ended journaling prompts, delivered across separate screens. The prompts were as follows:

1. How would you describe what you're feeling right now?
2. Did your body give you any clues about this feeling?
3. What do you think sparked this emotion?
4. Can you pin down the moment or thought that started it?

The prompts were grounded in the ABC model of journaling [260]. To support flexibility, users could skip any question they preferred not to answer. To accommodate diverse expression styles, the detailed reflection mode also included multimedia options, allowing users to record audio or upload images [276]. Finally, consistent with the quick mode, users could also log their current activity and rate their confidence in the reflection, ensuring coherence and comparability across data.

Contextual Factor	Examples or Description
Physical Activity	Performed some physical activity
Temperature Change	Change in temperature (e.g., AC to outdoors)
Medication	Took some form of medication
Food Intake	Had food recently
Caffeine	Consumed caffeinated drinks
Alcohol/Sugar	Consumed alcohol or sugary drinks
Environment	Noisy, crowded, or chaotic surroundings
Health	Feeling unwell or in pain
Supplements	Took vitamins or supplements
Recreational Substances	Used substances like nicotine
Menstruation	Menstruating (if applicable)
None of the Above	No relevant contextual factor

Table 6.1: Contextual Factors List

LLM-Supported Reflections

As a third self-reporting modality, the system includes a conversational interface that enables users to engage in dialogue while reporting emotions (see Figure 6.4c). The rationale for including this mode is to complement both quick-entry and structured journaling approaches by introducing interactive scaffolding for cases where emotions are ambiguous, evolving, or difficult to articulate. We hypothesize that, unlike one-way reporting, a conversational format can enable iterative clarification through prompts and follow-up questions, helping users progressively refine and externalize their emotional experiences. Given that most prior EMA systems rely on single-modality inputs, this mode is positioned as an exploratory extension to examine how the LLM-mediated annotation mode might support existing predefined format-based approaches for data collection. This design is further motivated by the recent adoption of LLMs in reflective journaling contexts, where they have shown promise in facilitating emotion articulation [220, 121, 120, 258]. To maintain design consistency, the chatbot option followed the same sequence as the other methods: quadrant selection, conditional stress scale, emotion list, chatbot interaction, and activity selection and confidence rating.

PID	Age	Gender	Education	Occupation	Mental Health Diagnosis	In Therapy	Emotional Event
P1	26	Male	Master's Degree	Student	No	No	Yes
P2	27	Male	Bachelor's Degree	Phd Student	No	No	Yes
P3	23	Female	Master's Degree	Phd Student	No	No	Yes
P4	23	Male	Bachelor's Degree	Founder	No	No	Yes
P5	29	Female	Bachelor's Degree	Software Engineer	No	No	Yes
P6	29	Female	Master's Degree	Student	Schizophrenia, Depression and Anxiety	Yes	Yes
P7	22	Male	High school	Student	Acute Clinical Depression and Anxiety	Yes	Yes
P8	23	Female	Bachelor's Degree	Research Associate	Anxiety	Yes	No
P9	25	Female	Master's Degree	Student	Anxiety	Yes	Yes
P10	28	Male	Master's Degree	PhD Student	No	Yes	No
P11	23	Male	Bachelor's Degree	Software Engineer	PTSD	Yes	Yes
P12	21	Female	High school	Student	No	No	No
P13	29	Male	Bachelor's Degree	Software Developer	No	No	No
P14	26	Female	Master's Degree	Home Maker	No	No	Yes
P15	24	Male	Bachelor's Degree	Software Engineer	No	No	Yes
P16	29	Female	Master's Degree	Phd Student	No	Yes	Yes
P17	30	Female	Master's Degree	Phd Student	No	No	Yes
P18	22	Male	Bachelor's Degree	Software Engineer	No	No	No
P19	27	Female	Master's Degree	Phd Student	No	No	Yes
P20	27	Female	Master's Degree	Phd Student	No	No	No
P21	27	Male	Bachelor's Degree	Phd Student	No	No	No
P22	21	Female	High school	Student	No	No	Yes
P23	21	Female	High school	Student	No	Yes	Yes
P24	22	Female	High school	Designer	ADHD	No	Yes
P25	34	Male	Bachelor's Degree	Freelancer	No	No	No
P26	21	Female	High school	Student	No	No	Yes
P27	26	Female	Master's Degree	Phd Student	No	No	Yes
P28	25	Male	Bachelor's Degree	Phd Student	No	No	No
P29	37	Female	Master's Degree	Manager	No	No	Yes
P30	22	Female	Bachelor's Degree	Research Associate	No	Yes	Yes
P31	20	Male	High school	Student	No	Yes	No
P32	22	Male	High school	Software Developer	No	No	Yes
P33	27	Female	Master's Degree	Phd Student	No	Yes	Yes

Table 6.2: Participant Demographics and Mental Health Background. Emotional Events refer to participants' recent experiences with significant emotional events.

Question	Response Summary (N=33)
Daily routine	Very structured: 3 (9.1%), **Somewhat structured: 22 (66.7%)**, Unstructured: 8 (24.2%)
Family dynamics	Supportive and emotionally open: 12 (36.4%) **Supportive but not emotionally expressive: 14 (42.4%)** Limited emotional support: 7 (21.2%)
Work-life Balance	**Well: 14 (42.4%)**, Moderately: 11 (33.3%), Poorly: 8 (24.2%)
Comfort Expressing Emotions	Very comfortable: 1 (3.0%), **Somewhat comfortable: 15 (45.5%)**, Neutral: 10 (30.3%), Somewhat uncomfortable: 7 (21.2%), Very uncomfortable: 0 (0.0%)
Concerned about others perception	Very concerned: 7 (21.2%), **Somewhat concerned: 11 (33.3%)**, Neutral: 10 (30.3%), Not very concerned: 4 (12.1%), Not concerned at all: 1 (3.0%)
Emotions as a sign of weakness	Strongly agree: 2 (6.1%), Somewhat agree: 3 (9.1%), Neutral: 5 (15.2%), Somewhat disagree: 9 (27.3%), **Strongly disagree: 14 (42.4%)**
Past Experience with Emotion Logging	**Yes**: 7, No: 26
Alexithymia (TAS-20)	**Low**: 30 (90.9%), High: 3 (9.1%)
Cognitive Reappraisal (ERQ-6)	**High**: 27 (81.8%), Low: 6 (18.2%)
Expressive Suppression (ERQ-6)	**High**: 21 (63.6%), Low: 12 (36.4%)
Resilience (BRS-6)	**High**: 22 (66.7%), Low: 11 (33.3%)

Table 6.3: Participant responses to daily support environment, emotional expression, and psychosocial measures.

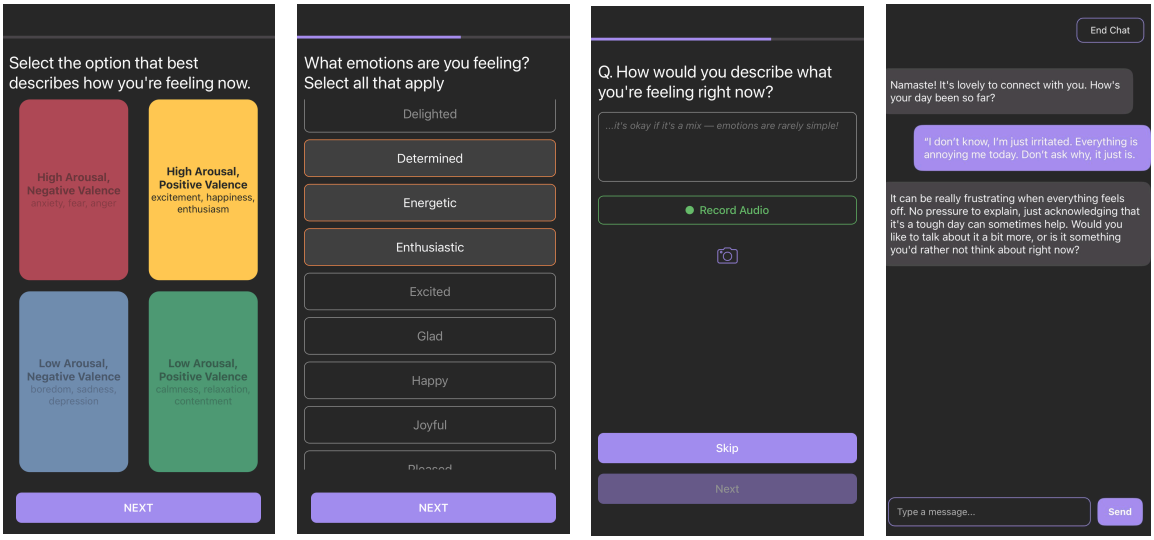


Figure 6.4: This figure displays the three reporting modes: (a–b) the quadrant screen and emotion list, (c) the detailed mode (Q1), and (d) the LLM-based mode (Best viewed in color).

6.2 Feasibility Study

Pre-Study Survey: We administered a pre-study survey via email to all interested participants, which included informed consent and baseline questions on demographics, mental health history, prior counseling, recent emotional events, daily routines, family and work–life context, and comfort with emotional expression. These factors helped contextualize participants' self-reporting behaviors, given known influences of routine, privacy, and social perception on EMA engagement [2, 29, 290]. The survey also incorporated three standardized measures: the 20-item Toronto Alexithymia Scale (TAS-20) [291], the 6-item Emotion Regulation Questionnaire (ERQ-6) [292], and the 6-item Brief Emotion Resilience Scale (ERS-6) [293]. These instruments captured individual differences in emotion identification, regulation, and resilience, which were important for interpreting how participants interacted with the emotion-logging system [2, 121]. More details added in appendix D.2.

Field Study Design: The study ran for three weeks, with each participant using the application for one week based on availability. Participants were 20–37 years old (M

= 25.42, SD = 3.98), including 14 men and 19 women. After completing the pre-study survey, all participants attended an **onboarding session** (in-person or online) covering installation, daily self-reporting procedures, and data privacy practices, reinforced later via email and a user manual [2]. During setup, participants selected four daily notification slots that aligned with their routines and were introduced to impromptu logging option via the floating-point button. Each participant used the app for a planned seven-day period, though some continued voluntarily for up to 11 days, citing its usefulness for tracking emotions. We analyzed all collected data to capture authentic engagement patterns. At study completion, participants completed a **feedback survey** assessing usability, relevance, and overall satisfaction [121] (More details added in appendix D.3), followed by an online **semi-structured exit interview** conducted via Zoom Pro (More details added in appendix D.4). The study received Institutional Review Board approval.

Participants Recruitment: We recruited participants using a mix of snowball sampling [294] and convenience sampling [217], leveraging institutional emails and social media. Of the 50 individuals who expressed interest, 35 enrolled and 33 completed the one-week study, with two withdrawing on the first day due to scheduling conflicts. All participants were over 18, enrolled voluntarily without paid incentives, and provided informed consent. Participants were not incentivized, allowing us to observe engagement that more closely reflects authentic, voluntary use. No exclusion criteria related to mental health history or prior emotion-logging experience were applied. Our goal was to capture varied psychosocial profiles and emotional experiences. Table 6.2 summarizes demographics and mental health history, while Table 6.3 details psychosocial profiles including support environments, emotional expression, alexithymia, regulation, and resilience.

6.3 Analysis

We adopted a mixed-methods approach combining descriptive quantitative analysis, mixed-effects modeling, and qualitative interpretation to examine how multimodal emotion logging

shapes both user behavior and the expressive characteristics of self-reports in everyday contexts. Given the exploratory nature of this study, our goal was not to evaluate long-term compliance but to examine how specific design features (scheduling flexibility, multi-modality, multi-select emotion list, and media sharing) shape (1) user experiences, data-sharing behavior, and (2) the characteristics of the resulting emotion data. To address RQ1, focusing on user behavior, we investigate the following sets of exploratory questions:

- **E1**: How do impromptu and scheduled logging approaches differ in terms of supporting user flexibility and the differentiation of reported emotional experiences?
- **E2**: How is annotation modality choice associated with temporal and affective context, reflecting users' adaptation to situational constraints and cognitive load?
- **E3**: How do individual characteristics (e.g., mental health history, emotional profiles, daily routines) relate to interaction patterns and modality preferences?

To examine E1–E3, we employed mixed-effects models with participant included as a random effect to account for repeated measures and inter-individual variability in logging behavior. We used this approach to appropriately model the nested structure of our data, where multiple observations are contributed by each participant across time and modalities. Fixed effects (e.g., scheduling type and annotation modality) and dependent variables were specified according to each exploratory question; further details are provided in the findings section. To explore RQ2, we examine how different expressive modes supported by the system shape the structure of emotion self-reports. Rather than treating "richness" as a scalar property, we conceptualize it as a multidimensional construct capturing how emotions are expressed and contextualized in user-generated data. We operationalize expressive richness along three dimensions: (1) Expressive elaboration: the extent to which detailed and chatbot-based modalities enable participants to provide descriptive accounts of their emotional experiences beyond scale-based quick entries. (2) Emotional complexity: the extent to which detailed and chatbot-based modalities provide space for expressing emotional states

that would not typically be captured in quick-mode entries, including mixed, overlapping, or evolving emotions. (3) Contextual grounding: the extent to which these modes support richer grounding of emotional experiences in situational context, including explanations of why users felt certain emotions and the nature of the events or circumstances underlying them, which are often absent in standard quick-entry EMA-style logging. To examine these dimensions, we conducted an inductive thematic analysis [295] of all text-based entries collected through the journal and chatbot modalities. Throughout the analysis, we employed constant comparison across modalities to identify systematic differences in how emotional experiences were structured in detailed journal entries and chatbot-mediated entries, with quick mode entries used as a baseline for reference. Additionally, we drew on descriptive statistics from the quick mode entries to contextualize the qualitative findings and support interpretation of differences in expressive patterns across reflective and chatbot-based modes relative to baseline quick logging. In addition, we performed a separate inductive thematic analysis [295] of exit interviews and open-ended survey responses to triangulate our understanding of user behaviors. Interviews were first transcribed using Zoom Pro's AI transcription feature and then manually verified for accuracy. For both the annotation data and participant feedback, three authors independently conducted open coding. The resulting codes were discussed regularly, with disagreements resolved through consensus. Codes were iteratively refined across multiple rounds of comparison, during which overlaps were merged, and irrelevant codes were removed. This process resulted in a set of higher-level themes. Together, these qualitative analyses, in combination with the quantitative results, structured the findings presented in this work.

6.4 Findings

In this section, we present our findings, which aimed to understand the influence of our features on participants' experiences and data quality. We have summarized our findings in table 6.4.

Theme	Key Finding	Implication / Interpretation
Flexibility in Scheduling	Participants preferred flexibility in scheduling their own routines and self-initiating emotion logs to better capture transient emotional states.	Allowing user-controlled logging may improve ecological validity and engagement.
	A generic logging schedule between 8–10 PM may not work effectively because participants have different sleeping and daily routines.	Fixed scheduling windows may reduce accessibility and participation for some users.
	When participants self-initiated impromptu logs, they were more likely to report negative emotional entries.	Self-initiated logging may be particularly valuable for capturing emotionally salient or distressing moments.
Choice of Modality	The choice of modality had no significant impact on emotional valence or intensity, with participants reporting similar emotions across all three modes.	Different logging modalities may be functionally equivalent for capturing core emotional outcomes quantitatively.
	Time of day did not significantly affect emotional valence or intensity.	Emotional reporting patterns appeared stable across different times of day.
	Qualitative findings suggested that participants preferred richer modes (e.g., detailed text or chatbot) when they had more time or when emotions felt particularly significant.	Richer modalities may be better suited for emotionally complex or meaningful experiences.
Participant Characteristics	Participants with higher self-reported concerns about sharing emotions showed higher response rates, suggesting the usefulness of short emotion annotation modes.	Lightweight logging approaches may lower barriers for emotionally reserved participants.
	Participants with higher alexithymia scores preferred richer modes.	Richer expressive modalities may help users who struggle to identify or articulate emotions.
Data Richness in Text-Based Modes	Text-based modes provided participants with space to express multi-layered and dynamic emotions.	Open-ended formats may better capture emotional complexity.
	Text-based modes accommodated contextual cues that might be missed in list-based activity selection.	Free-text input can preserve situational and experiential nuances.
	Chatbot modes supported self-reflection and sense-making, particularly when users were unsure what to annotate, and also functioned as a space for venting and self-understanding.	Conversational interfaces may provide both emotional support and richer reflective data collection.

Table 6.4: Summary of key findings across scheduling flexibility, modality preferences, participant characteristics, and data richness.

6.4.1 Understanding User Experiences and Data-Sharing Behavior

In our field study, we collected 505 logs across 221 participant-days from 33 participants. While the study was designed for 7 days per participant, actual voluntary use ranged from 1 to 11 days (M = 6.7 days), with no explicit author reminders or instructions to engage with the application after 7 days.

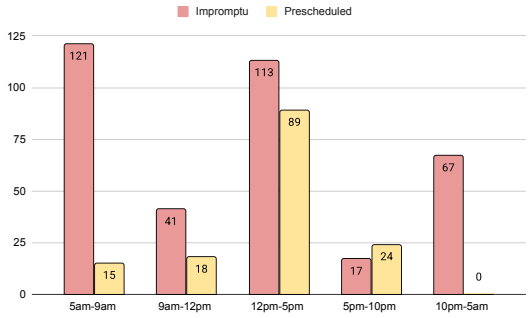
Effects of Scheduling Approaches on User Flexibility and Emotional Expression

We start by investigating user experience with our two distinct scheduling approaches (E1). To compare their usability for emotion logging, we calculated the scheduled response rate as the number of scheduled prompts answered, over the total number of scheduled prompts delivered ($\text{prompts}_{\text{Scheduled\_Delivered}} = 4 \times$ active participant-days). We considered scheduled prompts within a 1-hour tolerance window of the scheduled time as valid, since they matched our onboarding instructions. For impromptu logging, since participants had continuous access to the logging interface, we operationalized engagement as the impromptu response rate, defined as the proportion of participant-days on which any impromptu logging occurred, $\text{days}_{\text{Self\_Initiated}}$, over the total number of active participant-days. Participants demonstrated markedly different usage patterns between the two approaches. For scheduled prompting, participants achieved a 16.5% response rate. This meant 738 scheduled prompts (83.5%) went unanswered, representing substantial non-compliance with self-scheduled routines. In contrast, participants generated 359 impromptu logs, resulting in an overall daily engagement rate of 88.7%. To statistically compare these scheduling paradigms, we adopted an opportunity-level analysis framework. We defined each scheduled prompt as one response opportunity (N = 884) and each participant-day as one impromptu opportunity (N = 221), resulting in 1,105 total observations. We modeled response probability using a generalized linear mixed-effects model with random intercepts for participants:

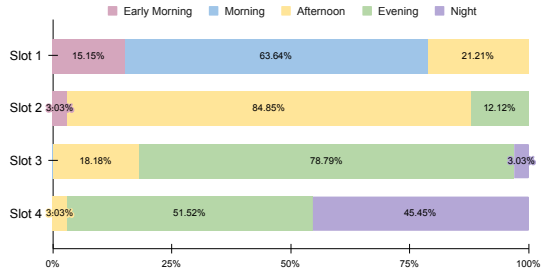
$$\textit{response} \sim \textit{schedule\_type} + (1|\textit{participant\_id})$$

Results from the mixed-effects model showed that scheduled prompts were associated with significantly lower response probabilities compared to impromptu opportunities ($\beta = -0.724$, $SE = 0.027$, $z = -26.99$, $p < .001$). This pattern was consistent with the observed response rates, indicating a substantial practical difference between conditions. The fixed effect of schedule type accounted for a considerable proportion of variance in response behavior (marginal $R^2 = .393$), while the full model, including participant-level random effects, explained 40.3% of the variance overall (conditional $R^2 = .403$). The relatively small difference between marginal and conditional R^2 suggests that schedule type contributed substantially more to response variability than participant-level differences. The model included 33 participant groups with observations ranging from 10 to 55 per participant ($M = 33.5$). Overall, these results show an association between schedule type and response probability (also illustrated in Figure D.1). In the qualitative analysis, participants emphasized that the combination of both approaches was useful. The prescheduled notifications facilitated habit formation, while the floating-point button (user-initiated) gave users the flexibility to annotate based on their emotional intensity and routine, thereby adding a much-needed layer of autonomy. A participants explained: *"I mostly used "+" button (**impromptu**). But sometimes you just won't remember that you can talk to somebody or you can write your emotions down somewhere. So the **notifications made me realize**, okay, okay, **there is an application** I can use to write my emotions down."* **(P30, 29, F)** We also observed a preference for the impromptu method during unstructured hours of the day, such as early morning or late at night, as evident in engagement logs (see Figure 6.5a).

Furthermore, we analyzed the slot-selection data (see Figure 6.5b). Our analysis revealed clear temporal patterns with evening hours (5–10 PM) being most popular (35.61%), followed by afternoon (12–5 PM, 31.82%), morning (9–12 PM, 15.91%), late night (10 PM–5 AM, 12.12%), and early morning (5–9 AM, 4.55%). This suggests participants favor notifications during active hours, from midday through the evening, with minimal interest in early-morning and late-night interruptions. Our feedback survey data revealed that



(a) Distribution of prescheduled and impromptu responses across time of day.



(b) Distribution of reporting times across five daily periods: Early Morning (5–9 AM), Morning (9–12 PM), Afternoon (12–5 PM), Evening (5–10 PM), and Night (10 PM–5 AM).

Figure 6.5: Temporal patterns of emotion reporting behavior across scheduling conditions and time-of-day distributions.

participants consider practical factors, such as natural breaks, transitions between activities, and changes in their environment, when selecting a slot. Some participants preferred longer intervals to allow meaningful mood variation (e.g., a 4-hour gap, noting mood would be unlikely to change sooner), while others aimed to "cover the whole day" or log during "productive hours" to capture more complete experiences. Device usage patterns also shaped selection, with participants choosing times aligned with their typical phone use. Furthermore, participants suggested adding an option to change prescheduled slots weekly or daily, noting the need for greater flexibility to accommodate their varying daily routines. As expressed by a participant:

> *"The schedule was fine. It suited me because I had the* ***option to choose my own slots****, so I chose the times that would be better for me. I* ***wouldn't have been able to annotate if there were pre-fixed time*** *slots. For instance, I wouldn't have been able to annotate at 9 AM, when I am usually in the metro. But because I had chosen my time slots, I was also able to do detailed reflections 2 or 3 times."*
> **(P33, 23, F)**

Next, to explore whether scheduling flexibility affected emotional expression, we used

three separate linear mixed-effects models that accounted for participant-level clustering via random intercepts. We operationalized emotional context through binary valence measures (negative = 0, positive = 1), binary arousal measures (low = 0, high = 1), and the count of emotions selected from the emotion list. The models followed these specifications:

$$\begin{aligned} valence &\sim schedule\_type + (1 \mid participant\_id) \\ arousal &\sim schedule\_type + (1 \mid participant\_id) \\ emotion\_count &\sim schedule\_type + (1 \mid participant\_id) \end{aligned}$$

We adopted this model because each participant contributed multiple emotion reports across both scheduling conditions, and observations are not independent. We specified three separate models to reflect distinct aspects of emotional expression: valence, arousal, and emotion count. This separation was necessary because these constructs capture conceptually different dimensions of emotional experience and may respond differently to scheduling manipulation. Results revealed a significant difference in valence only between conditions. Prescheduled logs showed higher positive valence compared to impromptu logs ($\beta = 0.096, \text{SE} = 0.047, p = .039$), corresponding to a 9.6 percentage point increase in positive emotional content (72.6% vs. 61.6%). Consistently, impromptu logs contained a higher proportion of negative valence entries (38.4% vs. 27.4% in prescheduled logs). No significant differences were observed for arousal ($\beta = 0.014, p = .758$) or emotion list selection count ($\beta = 0.072, p = .666$). Descriptive distributions further contextualized these patterns. While impromptu logs contained a higher absolute number of both positive and negative emotional instances due to greater overall volume (755 positive, 374 negative vs. 241 positive, 89 negative in prescheduled logs), proportional comparisons showed a more positive skew in prescheduled entries, reflected in higher positive-to-negative ratios (2.71 vs. 1.80). Our qualitative data also showed a preference for self-initiating logs for

negatively charged emotions. A participant reflected on this:

> *"It was eye-opening for me because sometimes you're not feeling your emotions. Sometimes **you're in a bad mood**, maybe even a good mood, but you don't realize it. And when the app notified...I remembered to pinpoint how I was actually feeling. And later it made me log when I felt the need to log an emotion (using impromptu approach)."* **(P23, 21, F)**

Overall, the findings indicate a clear preference for greater flexibility in emotion self-reporting, with scheduling conditions also shaping the emotional content of entries. Specifically, prescheduled prompts tend to elicit more positive valence, whereas impromptu logging captures a broader distribution of emotional experiences with a relatively higher proportion of negative valence. However, this pattern is confined to valence: no significant differences were observed in arousal or the number of emotions selected, suggesting that the influence of scheduling is selectively expressed in emotional valence rather than across broader dimensions of emotional reporting.

Effect of Emotions and Temporal Context on Choice of Modality

Our prototype was designed to support expressive affordances across a wide range of emotional experiences. Across the 505 emotion logs collected, most entries were made using the quick mode (433 logs, 85.7%), followed by fewer detailed entries (52 logs, 10.3%) and LLM-assisted logs (20 logs, 4.0%). Due to a technical issue in the data collection pipeline, LLM annotation data for four participants were not correctly recorded, potentially underestimating engagement in the LLM condition. Despite this limitation, the overall data indicate participants' strong preference for quick mode, as expected. To explore how providing modality choice supported emotional expression across situational contexts and emotional intensities (E2), we employed three linear mixed-effect models with random participant intercepts to examine differences in valence, arousal, and selected emotion counts

across the three modalities. The model specification was:

$$\textit{emotional\_outcome} \sim C(\textit{modality}, \textit{Treatment('quick')}) \tag{6.1}$$

$$+ (1 \mid \textit{participant\_id}) \tag{6.2}$$

We chose this model because it allows us to isolate the effect of modality choice (as a within-subject factor) on different emotional outcomes, valence, arousal, and emotion count, while controlling for inter-individual variability. We used treatment coding with "quick" as the reference condition to enable direct interpretation of each modality relative to the baseline interaction type. This specification provides a consistent and interpretable framework for comparing how different modalities influence emotional expression across contexts and intensities (E2), while maintaining robustness to participant-level heterogeneity. Across all three mixed-effects models, we did not observe statistically significant differences in emotional expression across modalities. For valence, neither detailed ($\beta = -0.078$, $SE = 0.071$, $p = .276$) nor chatbot entries ($\beta = -0.084$, $SE = 0.109$ $p = .438$) differed significantly from quick entries. The modality effect accounted for minimal variance in valence (marginal $R^2 = .003$), while the full model including participant-level random effects explained 6.9% of the variance (conditional $R^2 = .069$). Similarly, for arousal, both detailed ($\beta = 0.008$, $SE = 0.070$, $p = .907$) and chatbot entries ($\beta = -0.106$, $SE = 0.106$, $p = .322$) showed no significant effects. Again, modality explained very little variance in arousal responses (marginal $R^2 = .002$), whereas the inclusion of participant-level variability increased explained variance to 11.4% (conditional $R^2 = .114$). The number of distinct feelings reported was also comparable across modalities, with detailed ($\beta = 0.416$, $SE = 0.256$, $p = .105$) and chatbot entries ($\beta = 0.343$, $SE = 0.386$, $p = .373$). The modality effect remained small (marginal $R^2 = .004$), although participant-level differences accounted for a larger proportion of total variance overall (conditional $R^2 = .487$). Taken together, these findings suggest that modality choice had minimal influence on the structural

characteristics of emotional reporting. Instead, emotional expression remained broadly consistent across quick, detailed, and chatbot-based interactions, with participant-level differences contributing more substantially to variability than the modality itself. Next, we examined whether participants' modality preferences varied across temporal contexts, specifically across different times of day. To account for this repeated-measures structure while controlling for individual differences in baseline modality usage, we employed a mixed-effects model with participant-level random intercepts. We modeled time-of-day variation as a function of reporting modality, using the quick modality as the reference condition:

$$\textit{time\_of\_day} \sim C(\textit{modality}, \textit{Treatment('quick')}) \tag{6.3}$$

$$+ (1 \mid \textit{participant\_id}) \tag{6.4}$$

This model enabled us to examine whether certain modalities were more likely to be used during particular periods of the day while accounting for participant-specific reporting tendencies. The mixed-effects analysis revealed no statistically significant differences in modality use across times of day. Relative to the quick modality, detailed entries did not differ significantly in reporting time ($\beta = 19.17$, $SE = 39.88$, $p = .631$), and chatbot entries also showed no reliable temporal deviation ($\beta = 92.67$, $SE = 61.99$, $p = .135$). Although chatbot interactions appeared descriptively earlier in the day, the large uncertainty intervals and non-significant effects indicate that modality choice was not systematically structured by time-of-day patterns. The random intercept variance was comparatively large (σ^2 = 924.48), suggesting substantial participant-level variability in reporting times. This indicates that temporal logging behaviour was highly individualized, with differences between participants outweighing any consistent modality-specific temporal trends. To further examine temporal context, we analysed associations between day of the week and modality choice using

generalized estimating equation (GEE) models.

$$is\_detailed \sim C(day\_of\_week) \tag{6.5}$$

$$is\_llm \sim C(day\_of\_week) \tag{6.6}$$

Across most weekday comparisons, no statistically significant associations emerged for choosing either the detailed or chatbot modality over the quick modality. This suggests that modality selection remained relatively stable across the week rather than being driven by specific weekday routines or temporal rhythms. Taken together, these findings indicate that modality choice was not strongly determined by temporal context. Instead, participants appeared to use modalities flexibly throughout the day and across the week, with individual preferences and situational factors likely playing a larger role than consistent temporal patterns. Our qualitative analysis further reinforced this pattern, suggesting that modality choice was primarily shaped by in-the-moment situational factors and individual preferences rather than stable or systematic usage patterns. Participants frequently used the quick mode during busy moments or low-effort check-ins, while detailed entries were more often reserved for situations involving greater time availability or emotionally nuanced experiences. Overall, participants valued having access to both quick and in-depth reflection modes, appreciating the flexibility to choose the modality that best fit their needs and context. Some participants valued the chatbot for guidance when they had time, while others avoided it due to privacy concerns, lack of need, or discomfort with non-human interactions. Several participants also reported relying mainly on quick annotations because their emotional experiences during the study period did not feel sufficiently intense to justify more elaborate reporting. However, they noted that they would likely engage with richer modalities during periods of stronger emotional experiences or when deeper reflection was needed. As one participant explained, these contextual considerations directly shaped their modality choices:

"I use the quick annotations very frequently. I did not use the chatbot at all, and

I use detailed reflection ***when I have the time****, and I also use detailed reflection* ***when I could not understand which one of the 4 quadrants I fit into****. I then use the detailed reflection to analyze what I was actually feeling."* **(P33, 23, F)**

Effect of Individual Characteristics on Emotion Self-Reporting Behaviors

To investigate how individual characteristics might relate to interaction patterns and modality preferences, we tested the effect of participant characteristics on emotional self-reporting behaviors using three mixed-effects models with random intercepts for each participant. This approach accounted for individual baselines in daily response rate, modality choice, and emotional engagement. The model specification was:

$$emotional\_behavior \sim predictors + (1 \mid participant\_id) \tag{6.7}$$

We examined the individual characteristics as predictors, including demographics (age, gender), psychological traits (alexithymia, cognitive appraisal, expressive suppression, resilience), clinical characteristics (mental health diagnosis, therapy experience), contextual factors (daily routine, family dynamics, work-life balance), and emotional attitudes (comfort with expression, concerns about sharing, viewing emotions as weakness). Demographics captured broad population differences, psychological traits reflected emotion regulation capacities, and clinical characteristics accounted for prior mental health experiences. Contextual factors and emotional attitudes highlighted everyday environments and personal beliefs that could facilitate or constrain engagement. Together, these predictors allowed us to assess how both stable traits and situational conditions shaped participants' engagement, while simultaneously accounting for both within- and between-participant variability. Pre-testing, continuous predictors (Age, TAS Score, Cognitive Appraisal, Expressive Suppression) were standardized using z-score transformation, while categorical predictors (Gender, Therapy, Mental Diagnosis, Routine, Comfort with Expression, Concerns about Sharing, Emotion-

as-Weakness) were numerically encoded using label encoding. We defined two dependent variables: Daily Response Rate, calculated as the number of entries per participant per day; Modality Choice, coded as 1 for quick entries and 0 for detailed or LLM-assisted. Each outcome was modeled separately using mixed-effects models, with participant ID included as a random intercept to account for individual baseline differences. The two models were specified as follows:

$$\textit{daily\_response\_rate}_{ij} = \beta_0 + \sum_{k=1}^{14} \beta_k(\textit{predictor}_k)_i + u_i + \epsilon_{ij}, \tag{6.8}$$

$$\textit{logit}(P(\textit{quick} = 1)_{ij}) = \beta_0 + \sum_{k=1}^{14} \beta_k(\textit{predictor}_k)_i + u_i, \tag{6.9}$$

$$\tag{6.10}$$

where u_i represents the participant-specific random intercept, capturing stable individual differences, and ϵ_{ij} represents residual error. Random intercepts captured participant-specific tendencies, while residual errors accounted for unexplained variability within participants. For daily response rate, two factors emerged as meaningful predictors of daily annotation consistency. Participants who reported being less comfortable expressing emotions (21.2% of the sample) exhibited higher response rates ($\beta = -0.297, p = .050$). Similarly, greater concern about sharing emotions (54.5% of participants) was associated with marginally higher response rates ($\beta = 0.153, p = .057$), suggesting that choice-driven design could have the potential to support participants with varying expressive needs. Our qualitative analysis reinforced this; many participants described the structured and private nature of modes, particularly the quick mode, which was used most frequently ($n = 433, 85.7\%$), provided them a safe space for reflection as it did not require naming people or elaborating on events, allowing them to engage without fear of exposure. As one participant noted:

"Given that most of my annotations are about sadness or depression or anxiety, I think I'm not concerned about sharing the name of the emotion that I'm

Emotion Quadrant	Quick	LLM	Detailed
High Arousal, Positive Valence	96	2	11
Low Arousal, Positive Valence	188	10	20
High Arousal, Negative Valence	59	3	10
Low Arousal, Negative Valence	91	5	11

Table 6.5: Emotion types across annotation modes.

> *feeling...**I'm more concerned about sharing the details of why I am feeling that emotion**. If someone knows the details of why I'm feeling that particular emotion, then that is an issue."* **(P14, 27, Male)**

For modality choice, alexithymia (TAS scores ranged 28–72, mean = 50.39, SD = 9.95) emerged as a significant predictor. Participants with greater difficulty identifying and describing emotions were significantly less likely to choose the quick modality ($\beta = -0.064, p = .045$), suggesting they relied more on detailed entries to externalize or clarify their emotions more effectively. Similarly, stronger concerns about sharing emotions significantly reduced the likelihood of selecting the quick mode ($\beta = -0.050, p = .004$), suggesting that more cautious participants invested extra effort in logging their emotions. This may reflect internalized emotional stigma, with participants preferring detailed self-reporting to carefully process and contextualize their feelings while managing perceived internal or social judgment [296].

6.4.2 Understanding the Impact of Multimodality on Data Characteristics

Our analysis in the previous section highlighted that across both quick-mode and conversational reporting, participants were generally able to annotate broad emotional states ranging from calmness and comfort to tiredness, stress, and anxiety. Overall, within our collected dataset, we observed that Low Arousal, Positive Valence (LAPV) emotions were reported most frequently (see Table 6.5). However, our qualitative analysis of long-form journal entries and chatbot conversations further demonstrated the importance of incorporating

richer reporting modalities into emotion annotation workflows. We observed that both journal and chatbot-based entries consistently elicited substantially richer emotional narratives from participants. Rather than simply naming emotions, participants used these elaborative modes to explain triggers, bodily sensations, interpersonal tensions, motivational struggles, coping strategies, and evolving interpretations of their own emotional states. In many cases, seemingly simple labels such as "Depressed," "Hopeful," "Tired," or "Well" expanded into layered emotional experiences involving loneliness, cognitive overload, relational burden, excitement, uncertainty, guilt, or emotional exhaustion. Next we will discuss three overarching themes emerged from our analysis.

Emotional Experiences are Multi-Layered and Dynamic

A recurring pattern across the text-based entries was that emotional experiences were rarely singular or static. Participants frequently described emotionally mixed or internally contradictory states, often using the additional narrative space to explain emotional shifts and co-occurring feelings. Our analysis highlights three annotation patterns across both journal-style entries and conversational logs: (1) explicit multi-emotion selection followed by rich elaboration, (2) single-label compression followed by rich elaboration, and (3) narrative or "journey-like" articulation. In the first type of emotion annotations, participants explicitly selected multiple emotion descriptors (e.g., "Amused, Delighted, Energetic, Enthusiastic, Excited, Glad" or "Ashamed, Disappointed, Bored, Gloomy, Guilty, Tired, Worried"). While this appears to indicate high emotional granularity, our qualitative analysis revealed that participants frequently bundled emotionally adjacent states without clearly separating their causes or temporal ordering. For instance a participant (P3) reported their emotions as "Energetic" and "Excited", however when we checked their journal entry we found that reported excitement was after solving a problem, and they also felt behavioral indicators like bodily vibrations and cognitive "aha" moments because they were initially struggling to solve the problem. However, the list alone did not capture the progression from confusion to

insight to satisfaction. It was only through accompanying explanation *"I understand how to crack that problem ... because I discussed with my friend and had a realization moment"* that the emotional structure became legible. This suggests that multi-label selection increases breadth but not necessarily depth.

A second and more common pattern involves participants selecting a single emotion (e.g., "Tired," "Hopeful," "Relaxed," "Depressed") while providing rich narrative detail that significantly complicates or even redefines the initial label. In these cases, the emotion tag functions more as a starting anchor than a complete description. For example, participant (P30) selected the single label "Depressed," which on its own suggests a relatively static and uniform emotional state. However, their accompanying journal entry reveals a substantially more layered and embodied experience of distress that extends well beyond this categorical label. The participant situates their emotion within an interpersonal conflict with their spouse, describing a breakdown in communication ("*he is not at all ready to understand me*"), emotional exhaustion *("I feel like I'm done*"), and a perceived lack of reciprocal effort despite attempts to resolve the issue. Rather than a singular state of depression, the account reflects relational strain, accumulated frustration, and a sense of emotional depletion shaped by repeated unresolved interactions. This is further intensified through explicit bodily grounding in Q2, where the participant describes somatic manifestations of distress: "*My eyes are puffy. My face is puffy. My body is crying out loud*". Here, emotional experience is not only cognitive or relational but also materially embodied, suggesting that affect is being registered through physical exhaustion and stress response. In Q3 and Q4, the participant further localizes the emotional trigger to a specific interaction ("*Something he said yesterday" and "I had a fight with him*") which reframes the initial label of "Depressed" as the outcome of a conflict rather than a generalized emotion state. Overall, while the quick label "Depressed" collapses the experience into a single category, the elaboration reveals a multi-dimensional emotional configuration involving interpersonal conflict, perceived invalidation, bodily distress, and cumulative emotional fatigue.

The third pattern is that many participants do not treat emotions as discrete categories, but instead describe them as *processes unfolding over time*. Rather than stating "I feel X and Y," they construct a narrative of transition, moving from one emotional state to another, often without explicitly naming each stage. For instance, participant P7 described a successful sales interaction. Rather than directly articulating multiple emotional states, the participant situated the experience within a temporal sequence: "*Feeling delightful because closed a deal with a client,*" *"successful attempt of sales,"* and *"minutes after I closed the sale order."* While the reported emotion labels were High Arousal, Positive Valence states such as "Energetic" and "Pleased," the elaboration reveals a broader progression tied to effort, completion, and immediate emotional response. Also intermediate states such as anticipation, pressure, or relief are not explicitly named, yet are implied through the narrative structure. In a quick-mode entry, this experience would likely have been reduced to a static label losing the temporal nature of the emotional experience. Taken together across all three patterns, a consistent insight emerges that the act of elaboration consistently revealed additional layers to emotion self-reports involving bodily sensation, cognitive appraisal, social context, and temporal change. Moreover, many participants naturally defaulted to narrative descriptions rather than categorical combinations when given space to reflect. Taken together, these findings suggest that emotional reporting systems benefit from moving beyond fixed-label paradigms toward hybrid structures that support both lightweight categorization and open-ended narrative expression.

Contextual Narration Made Emotions Interpretable

Another recurring pattern across both journal and conversational entries was that emotional labels alone were often insufficient to understand what participants were actually experiencing. The accompanying contextual narration transformed otherwise generic affective categories into interpretable and situated experiences. We observed that the same emotional label could correspond to substantially different lived experiences depending on context, see

Table 6.6. Without contextual narration, these experiences would appear identical within a categorical annotation scheme despite arising from different causes and potentially requiring different interpretations. Similarly, labels such as “Tired,” “Relaxed,” or “Anxious” became meaningful only when grounded in participants’ ongoing circumstances. In one case, “Tired” referred to physical strain after walking for several hours. In another, it reflected mental fatigue caused by being stuck on a technical problem for an extended period. Although both entries shared the same surface-level label, the underlying experiences differed. Contextual narration also revealed the social and interpersonal structure of emotional experience. Participants often situated their emotions within relationships, conflicts, or responsibilities. For example, feelings of sadness or frustration were tied to loneliness, lack of emotional reciprocity, or unresolved arguments with partners or friends. These contextual details changed the interpretation of the emotional label from an isolated affective state to a response embedded within ongoing social dynamics. In several cases, contextual elaboration revealed emotional mixtures that were not directly reflected in the selected labels themselves. For example, participant P33 selected “Hopeful” as the primary emotion label, which in isolation suggests a relatively stable positive emotional state. However, the accompanying narrative described a more layered experience: *“I also feel excited about what is to come, a little stressed too because there are a lot of things on the table, but not too stressed, just excited stressed I guess.”* The participant further connected this emotional state to bodily awareness and preparedness: *“my mind feels aware and observant because the body knows there is a lot of work to do.”* Here, contextual narration reveals an emotional state shaped simultaneously by optimism, pressure, anticipation, and task awareness. The phrase “excited stressed” illustrates how participants often used contextual explanation to communicate nuanced emotional configurations that are difficult to represent through predefined labels alone. This also suggests that emotional labels often capture only the dominant or most socially recognizable affective state, whereas contextual narration reveals co-existing tensions and subtleties. We also observed that many of these contextual details extended beyond the

Emotional Label	Context Revealed Through Elaboration
Sleepy	Exhaustion after overnight train travel
Sleepy	Low motivation and feeling lazy while working
Sleepy	Grogginess immediately after waking up
Sleepy	Physical exhaustion following intense activity or long day

Table 6.6: Examples showing how the same emotional label ("Sleepy") corresponded to different lived experiences when participants elaborated on their emotional state.

predefined activity categories available in our interface. This reflects a broader limitation of categorical context lists commonly used in EMA systems, where predefined options often capture only generic or symbolic aspects of experience while missing personally meaningful situational details.

A Window for Self-reflection and Sense-Making

Beyond supporting emotional descriptions, we observed that conversational entries often functioned as spaces for self-reflection and emotional sense-making. In several chatbot interactions, participants were not simply reporting emotions, but actively trying to understand, organize, or reason through what they were feeling. Rather than treating chatbot mode as a medium to express, participants used conversation as a medium for exploratory reflection. For example, one participant (P1) initially appeared to express a relatively straightforward low-energy emotional state associated with tiredness and demotivation. However, through conversational elaboration, the participant described simultaneously feeling overburdened, emotionally exhausted, socially isolated, and frustrated: *"I am working on 2 projects and in both I have to babysit everyone, even the senior."* The participant further explained: *"All my friends are calling me [to] dump their trauma and frustration on me and I don't have anyone to dump trauma."* The interaction eventually revealed not only exhaustion, but also emotional labor, unmet social support needs, and a desire to escape monotony: *"I want to go on a light outing, some sort of dinner and break my monotonous life."* Importantly,

these reflections unfolded progressively through interaction. The conversational structure often support participants in unpacking emotions incrementally, often moving from vague statements toward more interpretable explanations. In several cases, participants themselves expressed uncertainty about their emotional states. For instance, one participant (P26) repeatedly questioned why they were feeling sleepy and bored in the morning, asking the chatbot: *"Ohh I want to know why I am sleepy."* Here, the interaction became less about reporting a known emotion and more about seeking interpretation. Similarly, another participant (P27) did not begin by describing a concrete emotional state at all, but instead asked the chatbot about "deep communication" and requested information about "mindfull talk." The conversation gradually shifted toward mindfulness exercises and reflective discussion. Such entries suggest that participants occasionally approached the chatbot not merely as an annotation interface, but as a reflective companion in case of unclear or evolving emotional experiences. Overall, we observed that conversational interfaces enabled participants to articulate emotions indirectly through discussion of situations. This differs substantially from quick-entry approaches, where users are expected to identify and select emotions immediately. Our findings therefore suggest that conversational emotional reporting may support forms of emotional awareness and self-interpretation that are difficult to capture through categorical self-report alone.

6.4.3 User Experiences with the Application

In this section, we will share our qualitative findings on how various features within our application influenced user engagement across diverse participant profiles. These features include a tutorial, quadrant screen, stress scale, emotion lists, multimedia inputs, activity tags, and confidence ratings. **(1) Tutorial:** According to our feedback survey, 78.8% of participants found the tutorial helpful, while 15.2% reported a neutral experience. Participants indicated that the tutorial was crucial for understanding the arousal-valence quadrant system. Several participants also suggested supplementing the existing tutorial with a more detailed

video explaining *"how emotion annotations can support emotional well-being"* would be beneficial, noting that this could enhance motivation for users with limited emotional literacy.

(2) Quadrant-Screen and Stress Scale: 69.7% of participants reported that the arousal-valence system was easy to follow, while 21.2% found it moderately easy. Additionally, some participants found the arousal-valence quadrant system challenging to use when experiencing multiple or neutral emotions. They suggested enhancements, such as the ability to select intersecting quadrants or to indicate primary and secondary emotions, to more accurately represent complex emotional states, underscoring the importance of flexibility and personalization in emotional self-reporting tools. Most participants found the inclusion of numerical phrases useful, but they suggested reducing the 10-item scale to 5 for easier quantification.

(3) Emotion List: The emotion list was widely used, with participants selecting between 1 and 11 emotions per entry (mean = 3.615, SD = 2.328), reflecting both engagement and utility in expressing mixed emotional states. When asked about the comprehensiveness of the list, 21.2% found it sufficient, 42.4% mostly sufficient, and 36.4% found it limited or restrictive. Many participants requested the option to include additional emotions, such as *disrespected, betrayed, confused, blank, neutral,* and *blessed*, to allow for more personalized and accurate self-expression.

(4) Audio and Image Entries: The multimedia feature was used less frequently than text-based entries. Usage varied across participants: P3 (26, M) used audio 33 times, P25 (23, M) and P30 (29, F) each used it 4 times, and P8 (22, F) used it once, combining audio with an image of a donut to capture a moment of joy. These patterns suggest that multimedia options enabled richer emotional expression for some users, while others used this feature minimally, likely due to personal preference or privacy concerns. As one participant explained:

(5) Activity Tagging: Participants appreciated the ability to track emotions in relation to daily activities, which helped them identify patterns between mood and routines (see

Table D.1). However, the activity list was perceived as somewhat restrictive. Participants suggested adding more common activities, such as *"Quick Walk," "Chit-Chat with Friends," "Meditation," "Had a Meeting," "Attended Class,"* or an *"Other"* option for adding new activities with greater flexibility.

(6) Confidence Assessment: we found mixed reaction for this feature (see Table D.1). Many participants found the feature helpful for self-assurance, while a few reported it increased cognitive load, suggesting it should be optional. Additionally, many participants requested new features, such as the ability to edit past entries and access to their data history, including visualizations of emotional patterns, to enhance their sense of control and ownership over their data. Together, these findings emphasize the importance of personalization and flexibility in designing self-reporting tools for a diverse audience.

Additionally, participants suggested improvements such as richer visualizations, more flexible notifications, an option to expand activity and emotion lists, stronger privacy features, added guidance for managing emotions, and an option to connect to mental health professionals if required. Overall, participants responded positively: 48.5% were satisfied, 44.2% moderately satisfied, 75.6% found it easy to use daily, and 57.6% wanted to continue, especially with added features like history and trend tracking. Qualitative responses highlighted the app's perceived value for emotional awareness and self-regulation.

6.5 Discussion

Our study aimed to examine how providing users with greater flexibility in emotion reporting influences their logging behaviors, data-sharing practices, and the characteristics of the resulting emotion data by offering multiple modalities and spaces for expression. Next, we will discuss our findings and their implications for designing future emotion logging systems.

6.5.1 What "Richer Emotional Data" Means?

Our findings indicate that a multimodal emotion logging system primarily influences the expressive depth and interpretability of emotional self-reports, rather than changing the underlying range of emotions reported. Across quick-mode entries, participants were able to capture a broad distribution of emotional states. The presence of a multi-select emotion list supported the labeling of co-occurring emotions, and the activity list supported basic-level contextualization. This suggests that lightweight structured reporting remains sufficient for capturing everyday emotional states in ecological settings. And this was also evident in our quantitative analysis, which showed that there was no distinctive pattern in users emotion logs across the annotation modes. However, our qualitative analysis of long-form journal entries and LLM-based conversational interactions reveals an important distinction. The richer modalities do not necessarily change what emotions are reported, but they substantially change how emotions are expressed, contextualized, and made interpretable. Moreover, richness does not stem from modality alone, but from whether participants choose to elaborate beyond categorical labels. Across journal and conversational entries, we observed three consistent patterns - (1) Label expansion: A single label (e.g., "Depressed") unfolded into layered accounts involving relational strain, emotional exhaustion, and somatic distress. (2) Label compression: Multi-label selections can increased breadth but did not reliably capture temporal or causal structure without narrative explanation. (3) Process-oriented narration: Participants frequently describe emotional experiences as narrative transitions that were not represented in static labels. Taken together, these patterns show that structured annotations capture categorical snapshots, while elaborative modes capture interpretive structure. Overall, these findings suggest that multimodal annotation modes can augment the traditional emotion logging. While the structured EMA captures what emotion is dominant, we propose that elaborative modes can support capturing why it is experienced and how it evolves. Moreover, these hybrid emotion logging systems can also support the development of more realistic artificial intelligent models of emotions that could better reflect how emo-

tions unfold in everyday life. Prior work has already shown the usability of text-descriptions in supporting emotion recognition [297]. In particular, contextual and process-level information can help bridge the gap between static emotion labels and dynamic real-world emotion trajectories, which are often missing in purely categorical datasets. Although these richer expressions still depend on users' expressive depth, contextual awareness, and emotional literacy, they align more closely with how individuals naturally communicate and make sense of emotions in everyday life. Furthermore, our results show a clear preference for self-initiated logging, suggesting that participants were more likely to engage with the system when they chose the timing themselves rather than responding to fixed prompts. This indicates that the choice-based design of our prototype helped accommodate differences in users' emotional sharing attitudes, allowing them to report emotions in ways that aligned with their availability, comfort, and momentary readiness. Overall, these findings suggest that while scheduled prompts and quick logs remain useful for ensuring baseline coverage, emotion logging systems should systematically incorporate user choice in both timing and interaction style. Doing so better accommodates the transient, contextual, and subjective nature of emotional experiences and supports more naturalistic patterns of emotion disclosure.

6.5.2 Considerations for Designing Multimodal Emotion Logging Systems

Our findings suggest several key considerations for designing multimodal emotion logging systems in the future. These considerations are grounded in how participants interacted with our prototype, their emotion data, as well as the challenges they reported in usability, motivation, and emotional articulation. Firstly, we saw in our engagement data that both the expressive modalities were less used in comparison with the quick entries. Within the expressive modalities, the chatbot based was used very rarely, only 20 captured instances. We observed that some people didn't use the chatbots at all. Our qualitative analysis of the feedback survey suggested that the conversational mode functioned less as a traditional annotation interface and more as a reflective scaffold that supported users in unpacking and

making sense of their emotions. Participants often used it to move from vague or compressed emotional labels toward more elaborate explanations. However, it also introduced clear design tensions. Participants reported concerns about time cost, particularly when emotional states were straightforward or when they had limited availability. In addition, some users experienced the system as non-human-like or repetitive, occasionally reiterating ideas already expressed by the user. This limited its perceived value in certain contexts where immediacy or emotional clarity was already present. This suggests that future systems should use LLM systems more like an on-demand interpretative tool rather than a fixed interface [258]. We also observed participants having concerns with sharing details about their emotional experiences due to privacy concerns, which was also highlighted in prior literature [268, 2], and thus, this also impacts the usability of expressive modes where users might be asked to share additional details. We thus recommend future systems to support abstract data logging, or maybe providing participants with training about sharing emotional details anonymously might help. A recurring challenge across modalities was sustained motivation. Participants frequently emphasized that their willingness to engage depended on emotional intensity, available time, and perceived relevance of logging at that moment. This reinforces the need for systems that support both low-friction quick capture and high-reflection deep capture, rather than privileging one mode. In parallel, differences in emotional literacy significantly shaped how participants engaged with the system. Some users required structured guidance (e.g., tutorials or emotion lists) to externalize their feelings, while others naturally engaged in narrative-rich expression. This variation reinforces the importance of designing for heterogeneous expressive capabilities, rather than assuming uniform ability to articulate emotions. These observations align with an idiographic perspective on emotion modeling [71, 298], where emotional expression is understood as individually situated rather than universally standardized. Overall, participants' feedback consistently points toward the need for flexible, optional, and user-controlled pathways of expression, rather than prescriptive workflows, suggesting that future systems should work towards balancing flexibility with

data needs.

6.6 Limitations

Our study has some limitations. First, it was a formative user study with a relatively small sample of 33 participants over one week, which may limit the generalizability of findings to broader populations or longer-term use contexts. While we aimed for diversity, most participants were well-educated, technologically proficient, healthy individuals who belonged to the same country and had a similar cultural background, so results may not extend to individuals with lower digital literacy, severe disorders, or other cultural contexts. In the future, we aim to involve larger, more diverse samples and longer-term deployments to further validate these insights. Moreover, despite implementing strong privacy measures, such as open-source LLM deployment on a private server, data anonymization, and secure storage, some participants still hesitated to share emotions, potentially reducing data richness. A further limitation of our study is that participants were not incentivized for engagement. While this allowed us to observe more naturalistic and voluntary usage patterns, it may also have introduced participation bias toward individuals who were already comfortable with emotional self-reflection or intrinsically motivated to engage with emotion-tracking practices. As a result, the observed engagement patterns may not generalize to broader populations who may require stronger external motivation or who are less inclined toward reflective self-reporting. Finally, the overlap between our study and the festival season has influenced user engagement, and we believe it has also highlighted the importance of situational context in user engagement.

6.7 Summary

In this chapter, I introduced a multi-modal annotation approach for capturing transient emotional experiences across diverse and dynamic emotional profiles. Moving beyond fixed emotional prompts, the system employs a choice-based design that allows users to

log emotions based on their current emotional intensity and availability. Together with other participant-centered features, this chapter's findings suggest that multimodal emotion logging can support the collection of richer, more nuanced emotional data. At the same time, I observed that sustaining long-term engagement remained challenging, even though the proposed approach enabled more transient, participant-agency–centered data collection. Building on these observations, the next chapter explores how post-logging guidance, data-driven feedback, access to log history, and more flexible, user-centered scheduling strategies can better support long-term engagement while maintaining the richness and quality of the collected data.

Chapter 7

Beyond Accuracy: Understanding User Needs in Emotional Well-Being Technologies

"My experience is what I agree to attend to. Only those items which I notice shape my mind – without selective interest, experience is an utter chaos."

— William James

Amid increasing societal competition, heightened productivity demands, and persistent social and economic uncertainty, concerns around mental health have intensified globally [299]. The World Health Organization has identified transforming mental health services as one of the most pressing public health challenges worldwide, reflecting both the scale of mental health needs and gaps in access to timely support [300]. In this context, ubiquitous computing research has increasingly explored how data-driven techniques, combined with mobile and wearable technologies, can support everyday emotional well-being beyond traditional clinical or therapeutic settings [301, 302]. Mobile and smartwatch-based applications now leverage artificial intelligence and continuous sensing to track emotions in situ, enabling emotional support for users in their everyday environments [303, 301, 122]. Advances in adaptive feedback and context-aware mechanisms have further expanded the potential for delivering personalized support that responds to users' moment-to-moment experiences [304, 305, 119, 306]. However, unlike other mobile health applications, such as step counting or calorie tracking, emotion-logging applications, despite being widely available for many years, have not achieved comparable levels of sustained adoption among users [307]. While these systems are designed to support emotional awareness and regulation [122, 86], their long-term utility and integration into daily life remain limited, with many users discontinuing use after brief periods [27, 100, 80]. Prior user studies on emotion-logging and assessment tools consistently report declining usage and low long-term compliance, indicating that initial adoption does not reliably translate into sustained engagement and

utility in real-world settings [308, 128, 224, 129, 96]. Similar trends have been observed in commercially available emotion and stress logging applications, where early use rarely leads to meaningful or long-term engagement [309, 310, 308]. Overall, prior work suggests that increasing the perceived utility and long-term adoption of these applications remains an unsolved problem [311].

Research has identified several factors that contribute to low engagement with mental well-being and emotion-logging applications. Borghouts et al. [311] categorize these into three levels. At the user level, engagement is influenced by individual characteristics such as mental health status, beliefs about emotional disclosure, digital literacy, and the ability to integrate logging into daily routines. At the system level, limited personalization, unclear usefulness, rigid interaction patterns, and insufficient guidance reduce sustained use. Finally, technology-level factors, including usability issues, privacy concerns, and social or cultural barriers such as stigma around emotional disclosure, impact engagement. Collectively, these findings indicate that disengagement is not due to a lack of interest in emotional well-being, but rather arises from mismatches between users' needs, values, daily contexts, and the design of existing systems. Prior work in ubiquitous computing and digital mental health interventions has explored technological strategies to improve user engagement with emotion-logging applications. Researchers have developed context-aware prompting to deliver reminders at opportune moments [312, 220], as well as micro-interval prompting to increase reporting frequency [128]. Efforts have also focused on increasing accessibility by shifting logging from smartphones to wearable devices, such as smartwatches, enabling more convenient, in-situ reporting [28]. In addition, gamification techniques, including badges, progress tracking, and rewards, have been employed to motivate continued interaction [313, 314]. While these approaches have shown promise in supporting initial adoption, sustained use, and meaningful integration into daily routines remain challenging due to several factors, such as real-life disruptors [315], and worldly barriers [311, 316]. Users' preferences for emotion self-tracking have also been studied qualitatively in the past [27, 29, 121, 210, 120,

81, 317], but most of these studies either focus solely on qualitative methods [316], or on user feedback after using an emotion logging tool [318], or on identifying barriers from literature [96]. Prior work has also emphasized the importance of involving users in the design and delivery of such interventions, suggesting that user-centered approaches could improve relevance and adoption [2, 101, 96, 86, 319]. However, user-centered design alone does not guarantee sustained engagement or improved value, and empirical evidence directly linking design practices to studying engagement remains limited [310]. Overall, despite growing interest in digital emotion self-tracking, user engagement, and human-centered design approaches, a persistent gap remains between how these systems are conceptualized by researchers and designers and how they are experienced and valued by users in everyday life. This gap points to the need for longitudinal investigations that examine how users' preferences, engagement patterns, and perceptions of value evolve over time as emotion self-tracking becomes embedded in daily practices.

In this chapter, we conducted a three-phase study to address the following research objectives: (1) to understand users' preferences for emotion-logging, including input methods and feedback for emotional awareness and regulation, and (2) to examine whether supporting these preferences leads to sustained, meaningful engagement and perceived value. The study was guided by the following phase-wise research questions, each aligned with a specific phase of the investigation:

- **RQ1a**: What types of input interfaces do users prefer for logging emotions?
- **RQ1b**: What kinds of short-term and long-term feedback do users want after logging their emotions?
- **RQ2a**: How do users engage with a user-informed, human-centric emotion logging application over time in a real-world deployment?
- **RQ2b**: What features, interaction patterns, and contextual conditions are associated with sustained engagement or disengagement?

- **RQ3a**: How do users perceive the value and impact of long-term emotion logging on their emotional awareness and well-being?
- **RQ3b**: How can emotion logging systems be designed to align with users' values to support long-term engagement?

Phase 1 focuses on understanding users' preferences for emotion logging, including input methods and feedback mechanisms, and addresses RQ1a and RQ1b. Insights from this phase informed the design of a human-centric emotion-logging application. **Phase 2** involves a four-week field study in which participants used the application in their daily lives and completed weekly and end-of-study surveys, enabling us to examine engagement patterns and the impact of user-preferred features on sustained use (RQ2a, RQ2b). **Phase 3** consists of a qualitative investigation using semi-structured interviews and open-ended surveys to explore why participants engaged with or disengaged from our application, and how they perceived its value and impact on their emotional awareness and regulation (RQ3a, RQ3b). An overview of our study design is shown in Figure 7.1. This chapter makes the following contributions:

- **A human-centered, adaptable emotion self-tracking system.** We design and implement a working prototype of a human-centered emotion self-tracking application that supports user autonomy, flexible logging, and choice-driven feedback, informed by users' design preferences.
- **Empirical insights from a real-world field study.** Through a four-week field deployment and qualitative investigations, we provide in-depth insights into how users engage with emotion self-tracking systems over time, revealing emotion-driven and selective engagement patterns shaped by internal motivation, prior coping strategies, and perceived value.
- **Design recommendations for meaningful engagement.** We derive design implications for future emotion self-tracking tools, emphasizing adaptive, low-effort, and

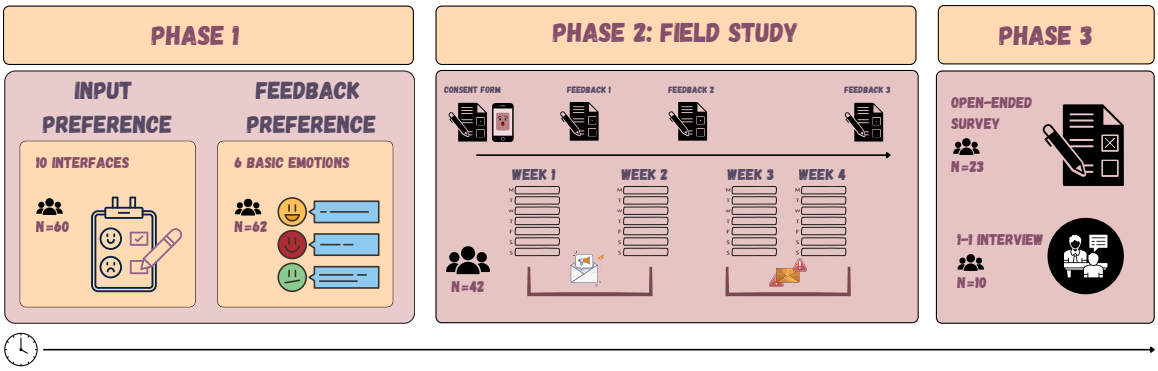


Figure 7.1: Overview of the three-phase study design and corresponding research questions.

Target Emotion	Scenario Description
Happiness	"You receive some surprisingly wonderful news that you weren't expecting at all, instantly brightening your day."
Sadness	"You realize that you have lost something that was very important and held a lot of sentimental value to you."
Anger	"You're waiting for something important, and you experience a significant and unfair delay or cancellation due to someone else's mistake."
Fear	"While you are walking, a car unexpectedly speeds through a stop sign you were about to cross, missing you by only a few feet."
Disgust	"You take a sip of a drink, expecting a familiar taste, but instead, you are met with a spoiled, sour, or foul flavor."
Surprise	"While cleaning or tidying up, you find something of value (either sentimental or monetary) that you were convinced you had lost forever."

Table 7.1: Situational scenarios used in Phase 1 Input Interfaces Study to elicit imagined emotional responses.

context-sensitive support that prioritizes actionable outcomes, emotional capacity, and long-term well-being over continuous use or compliance-based engagement.

7.1 Phase 1: Understanding User Preferences

7.1.1 Study 1: User Preferences for Emotion Input Interfaces

To address RQ1a, we began by examining users' preferences for input interfaces for logging emotions. Prior research has shown that users often experience difficulty articulating their emotions as they are felt [307, 1, 2, 26]. These challenges are frequently attributed to

individual differences in emotional profiles, including variations in emotional vocabulary, range, congruence, intensity, and reactivity [2]. These findings suggest that users who may appear similar at a surface level can differ substantially in how they understand and express their emotional experiences, shaped by a range of internal and external factors. As a result, different users may require different interface designs to effectively articulate similar emotional states. In the first Phase 1 study, we therefore investigated whether users' preferences for emotion-logging interfaces vary and whether there are commonalities or differences to consider when designing emotion self-tracking systems for diverse user groups. To examine users' preferences for emotion-logging interfaces, we developed a web-based evaluation platform that enabled the systematic comparison of ten distinct emotion-annotation interfaces. These interfaces were selected based on a review of commercially available emotion and mood tracking applications, as well as prior literature on emotion representation and self-tracking interface design. An overview of the ten interfaces is provided in Table 7.2. We employed a within-subjects study design, in which each participant interacted with and evaluated all ten interfaces. The study began with collecting informed consent, followed by an integrated demographic questionnaire that collected participants' age, gender, education, occupation, prior experience with emotion or mood tracking, and mental health background. After completing the demographic survey, participants sequentially interacted with each of the ten interfaces. For each interaction, participants were first presented with a randomly selected **imaginary scenario** from a set of six scenarios corresponding to Ekman's basic emotions—happiness, sadness, anger, fear, disgust, and surprise [23]. These scenarios were developed through a pilot study with five participants and refined through iterative discussions within the research team. All scenarios were intentionally designed to be generic and broadly applicable across diverse demographic groups. Table 7.1 presents the scenarios used in the study. Participants were asked to imagine themselves in the presented scenario and then use the interface to **log their emotional state**. The order of the interfaces was randomized for each participant to mitigate order effects. Following each interaction, par-

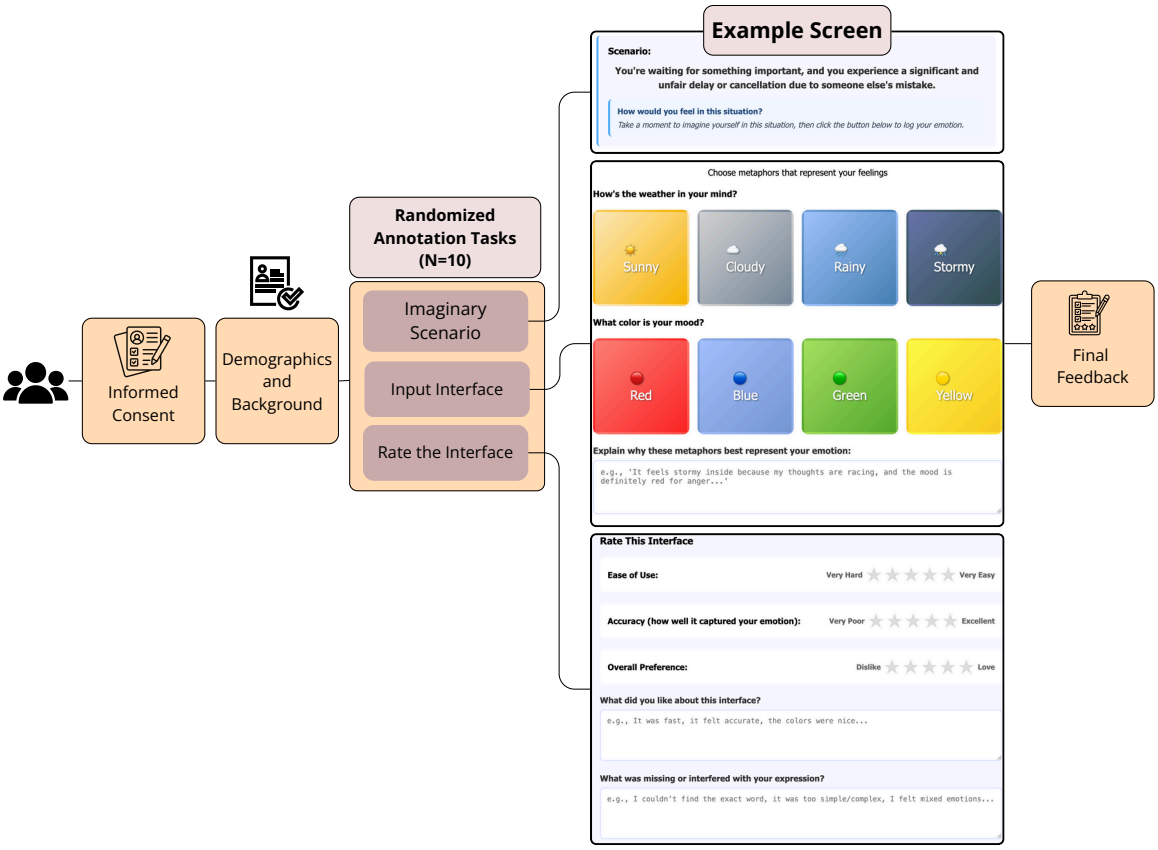


Figure 7.2: Overview of the web-based interactive study design used to evaluate emotion input interfaces in Phase 1.

ticipants evaluated the interface using a 5-point Likert scale assessing perceived ease of use, perceived accuracy, and overall preference. Participants were also given the option to provide open-ended feedback to elaborate on what they liked or disliked about each interface. The study concluded with a **summative assessment** in which participants ranked their most and least preferred interfaces, explained their reasoning, and provided suggestions for improvement and potential real-world usage contexts. An overview of the study procedure is illustrated in Figure 7.2, and we have added original screenshots of the website in the supplementary material.

Interface	Description
Basic Emotion Labels	A simple categorical interface presenting a grid of twelve discrete emotion words (e.g., Happy, Anxious, Frustrated), based on the evolutionary theory of emotion [23]. Users select applicable labels and indicate the overall emotional intensity using a numerical slider ranging from 1 to 10.
Emotional Space (Technical Terms)	A dimensional interface based on the Circumplex Model of Affect [134], represented as a 2D coordinate plane with axes labeled Valence (positive vs. negative) and Arousal (high vs. low arousal) alongside example emotions and colors in each quadrant. Users select one of four quadrants to categorize their emotional state.
Emotional Space (Simple Language)	A simplified version of the Circumplex Model of Affect [134] that retains the 2D structure but replaces technical terminology with intuitive language such as High/Low Energy and Positive/Negative Feelings.
Emotion Sliders	A continuous input method using two horizontal sliders representing emotional valence (Very Negative to Very Positive) and intensity (Very Mild to Very Intense) [134].
Emotion Tree Navigation	A hierarchical interface based on Parrott's Emotion Framework [320], where users progressively move from primary to more specific emotions.
Visual Metaphors	An abstract interface using symbolic representations rather than direct labels [321], including metaphors such as weather or colors, optionally accompanied by text explanations.

Interface	Description
Emoji Selection	An icon-based interface presenting emojis derived from the PANAS scale [322], with optional intensity ratings on a 1–10 scale.
Feelings Wheel	A digital adaptation of the Gloria Willcox Feelings Wheel [323], allowing navigation from core emotions to more specific descriptors.
Journaling (ABC)	A text-based reflective interface grounded in Cognitive Behavioral Therapy (CBT) [324], where users describe the Activating event, Beliefs, and Consequences.
Contextual Factors	An interface capturing links between emotions and contextual factors such as sleep, social interaction, work, or environment [2, 81].

Table 7.2: List of our ten emotion-logging interfaces evaluated in Phase 1, Study 1. Interface names were intentionally simplified and made user-friendly to avoid technical terminology and to ensure they were easily understandable to participants.

7.1.2 Study 2: User Preferences for Post-logging Feedback and Support

To address RQ1b, we examined users' preferences for post-logging feedback and support mechanisms. Prior work has shown that people often engage in emotion self-tracking to increase self-awareness and, over time, to better manage their emotions by identifying personal triggers and recurring patterns [2, 27, 325, 224]. Emotion self-tracking is also commonly recommended by mental health professionals through practices such as manual diary keeping or the use of digital applications. These practices are intended to help individuals develop emotional awareness, learn to label and regulate emotions, and maintain records that can support reflection or inform clinical conversations. Emotion regulation

is central to these practices, as they are intended to help individuals develop and refine strategies for managing and responding to intense emotional experiences. Therefore, many emotion logging applications also include feedback or supporting features to help users learn and support themselves over time [10, 319]. Designing and implementing user-centric emotion regulation tools remains an open challenge [122]. So, we conducted this study to understand user preferences for receiving feedback and support via emotion logging in an application. To enable this study, we designed a survey and delivered it to participants via Google Forms.

The survey began with demographic questions capturing participants' age, gender, occupation, and education, along with prior experience with emotion or mood tracking and self-reported mental health background. The main portion of the survey focused on participants' preferences for receiving feedback from an emotion-logging application across six basic emotions—happiness, sadness, anger, fear, surprise, and disgust [23]. For each emotion, participants responded to four categories of questions addressing (1) the preferred type and tone of feedback, (2) desired application features to support engagement or regulation, (3) how they typically experience or respond to the emotion in everyday life, and (4) opportunities for elaboration through open-ended reflection.

Closed-ended response options were tailored to each emotion and grounded in emotion regulation strategies and feedback mechanisms identified in prior literature. Rather than using uniform response sets, we designed emotion-specific options to reflect how different emotions are commonly experienced and supported in everyday life. For positive emotions such as happiness, options emphasized acknowledgment, savoring, and amplification strategies (e.g., celebratory feedback, gratitude prompts, reflection, or extending the moment), while allowing participants to indicate that no feedback beyond logging was needed. For negative emotions such as sadness, options reflected commonly studied coping strategies, including empathetic or validating messages, reflection prompts, soothing or mood-lifting activities, mindfulness exercises, social connection, or rest. Participants were also asked

to report their typical behaviors when experiencing each emotion to contextualize feedback preferences within existing regulation practices. Participants could select multiple options and provide open-ended responses across all emotions. In addition, participants reported their preferred modalities for receiving feedback or support to sustain emotion logging over time. We assessed preferences across audio-, visual-, and text-based feedback, as well as interactive, gamified, and social modalities, using example options grounded in prior work. These included soothing audio, visual summaries and trends, supportive text or chatbot interactions, interactive tools such as check-ins and mindfulness exercises, lightweight gamification, and optional social support. Participants could explicitly opt out of any modality and provide additional input. Overall, the survey aimed to better understand how digital feedback could align with, complement, or avoid disrupting users' natural regulation strategies.

7.1.3 Participant Recruitment

Participants for both phase 1 studies were recruited through a mix of snowball sampling [294] and convenience sampling [217], with the researchers sharing study flyers on their social media channels and the institution's email list. Study 1 received a total of 60 responses, while Study 2 received 62 responses. Both studies were IRB-approved and voluntary, with no rewards or compensation. Study 1 took approximately 30 minutes to 1 hour for each user to complete, as per their emotional profiles. Both studies were open to the public aged 18 or older and had no exclusion criteria related to prior mental health experiences or familiarity with emotion logging, as we aimed to understand the perspectives of a varied range of users. All participants were educated, spoke English, and were proficient with laptops and mobile phones to complete the studies. The demographics of participants in both the Studies in phase 1 are added in Table 7.3. This phase and subsequent phases are all approved by our institutional review board.

Category	Study 1 (N=60)	Study 2 (N=62)
Gender		
Female	28	24
Male	32	37
Prefer not to say	0	1
Age (years)		
18–25	27	26
26–35	31	32
35–45	1	1
46–55	0	2
55+	1	1
Education		
High School	6	3
Bachelor's / Graduation	29	32
Master's / Post-graduate	20	20
PhD / Doctorate	5	5
Other professional degree	0	2
Occupation		
Student	29	29
Software Engineer	18	26
Healthcare Professionals	3	0
Other Professional / Freelancer	8	7
Not working	2	0
Prior Emotion Tracking		
None	35	36
Manual journaling	13	12
Apps or wearables	11	13
Clinical journaling	1	1
Mental Health Background		
No history / concerns	39	43
Self-identified concerns	9	8
Therapy (past)	5	3
Therapy + medication	2	3
Diagnosed, not in treatment	2	5
Prefer not to say	3	0

Table 7.3: Participant demographics for Phase 1 studies.

7.1.4 Analysis

We employed a mixed-methods analysis across both Phase 1 studies. Closed-ended survey responses were analyzed using descriptive and basic statistical methods, while open-ended responses were examined using inductive thematic analysis [183]. In Study 1, we analyzed participants' ratings of each emotion-logging interface along three dimensions: ease of use, perceived accuracy, and overall preference. For each interface, we computed mean ratings for each metric and derived a composite score by averaging these means. We also calculated standard deviations to capture variability in participants' evaluations. These statistics enabled comparisons across interfaces in terms of both overall preference and consistency of use. Open-ended and summative feedback was further analyzed to contextualize these ratings and inform subsequent design decisions. In Study 2, we examined participants' selections of feedback strategies, tones, real-life coping behaviors, and modalities to support reflections across different emotions. For each emotion, we calculated frequencies and percentages to identify the most commonly preferred forms of support and to characterize variation across emotional contexts. Rather than converging on a single optimal strategy, this analysis surfaced patterns and clusters of valued feedback types and modalities, highlighting diversity in user preferences and informing the design directions for the next phase. Based on these analyses, we next present the findings from Phase 1.

7.2 Phase 1: Findings

7.2.1 Study 1: Findings

In Study 1, we focused on understanding users' preferences for different emotion input interfaces to support emotion logging. Overall, across all interfaces, participants consistently highlighted ease of use, perceived accuracy in expressing emotions, speed of interaction, visual intuitiveness and aesthetics, and the structured articulation of emotions as key factors influencing their preference for an interface. Across interfaces, Emotion Space – Simple

Language received the highest composite average rating across the three key dimensions i.e., ease of use, accuracy and preference (M = 4.41, SD = 0.76), followed by the Feeling Wheel (M = 4.23, SD = 0.97) and Emoji Selection (M = 4.23, SD = 0.94). These interfaces combined high perceived usability with relatively low variability in ratings, indicating both strong preference and consistency across participants. In contrast, the three lowest-ranked interfaces were Journaling (ABC) (M = 3.64, SD = 1.35), Visual Metaphors (M = 3.98, SD = 1.14), and Emotion Tree Navigation (M = 3.99, SD = 1.12). Overall, the journaling method was the least favored in all three dimensions, while the other two were easy to use, but users perceived them as less accurate and were overall preferred less. Participants consistently preferred interfaces that allowed quick, low-effort emotion logging. In table 7.4, we have shared participants' ratings across our ten interfaces.

Emotion Space – Simple Language was highlighted for its intuitive quadrant layout and simple language - *"Clear indication of emotion with a list and appropriate color in the background."*, *"The quadrant explained everything perfectly, and colors made the interface very intuitive."* While participants also felt that, although it is quick and intuitive, it lacked depth and is not very useful in case a deeper reflection is required. A participant mentioned - *"Lacks layering of emotions and no option for detailed discussions on my emotions."* In contrast, **Feeling Wheel** was valued for structured guidance and hierarchical selection of emotions: *"The layered structure helps narrow down and reflect deeply on complex emotions."* Participants also liked that they could see their emotions in layers, which made it much more useful for deeper reflections. But some participants felt a need to select multiple emotions and found it confusing when emotions were complex or when internal understanding was lacking, and suggested the need for multi-selection and an additional space to share intensity. A participant added -*"You can only select one path, which limits mixed or complex emotions. No intensity was measured, and tertiary options feel too limited or not fully representative."* **Emoji Selection** was praised for familiarity, visual cues, and expressive power. Plus, they liked that they could select multiple emotions to

overall reflect their emotions and had an option to share the intensity of their emotions. A participant mentioned -*"Love the visual representations, helps me check if I relate to the emoji or not"*, and another reflected *"Emojis are familiar and something we use in daily life"*. While the least-ranked interfaces were primarily ranked lower due to the cognitive effort required to understand them, the time required to share emotions, the ambiguity in sharing emotions, and limited flexibility. Interfaces such as Journaling (ABC) and Emotion Tree Navigation required multiple selections or reflective steps, making logging feel time-consuming and mentally demanding, particularly for quick, in-the-moment entries. Creative but abstract interfaces like Visual Metaphors and Contextual Factors were often confusing, as users struggled to map their actual feelings to the provided imagery correlation. Also, the interface where emotions were associated with numbers using sliders (Emotion Sliders, basic emotions) was also preferred less due to the complexity of assigning a number to emotions. As a participant expressed - *"I do not prefer to select emotions from a slider with a lot of options (like 1-10). This sends us into a deep thought about whether I should select, say, 7 or 8, which are both pretty close. Exactly quantifying the emotion is difficult. Moreover, we do not understand the implication of selecting 7 when it could have also been 6 or 8."* People who liked the Journaling (ABC) interface mentioned a desire to reflect deeply, but also stated that it might not be relevant in all emotional scenarios. As a participant mentioned - *"It's a good interface, as I can write down my feelings in detail. However, if I am in a rush, then this may not be the preferred interface."* Taken together, participants preferred interfaces that offer a good balance of intuitiveness, accurately capture a wide range of emotions, and are less time-consuming. These findings helped us in designing our application in the next phase.

7.2.2 Study 2: Findings

In Study 2, we examined participants' preferences for post-logging support across different emotional states, including the desired type, tone, and modality of feedback, as well as how

Interface Name	Ease of Use	Accuracy	Preference	Mean AVG	Ease of Use	Accuracy	Preference	Mean STD
Journaling (ABC)	3.48	3.85	3.59	3.64	1.46	1.25	1.35	1.35
Emotional Space - Simple Language	4.60	4.32	4.32	4.41	0.64	0.83	0.81	0.76
Emotional Space - Technical Terms	4.28	4.07	4.10	4.15	1.01	1.10	0.99	1.03
Visual Metaphors	4.18	3.80	3.95	3.98	1.13	1.19	1.11	1.14
Feeling Wheel	4.35	4.15	4.20	4.23	0.94	1.02	0.95	0.97
Emotion Tree Navigation	4.08	3.93	3.95	3.99	1.14	1.13	1.08	1.12
Emotion Sliders	4.38	4.00	4.02	4.13	0.96	1.10	0.93	1.00
Emoji Selection	4.35	4.16	4.16	4.23	0.87	0.96	0.98	0.94
Basic Emotion Labels	4.50	4.03	4.08	4.21	0.75	1.01	1.00	0.92
Contextual Factors	4.23	4.05	3.97	4.08	0.99	1.06	1.11	1.05

Table 7.4: Average Ratings and Standard Deviations of Emotion Logging Interfaces (Phase 1 Study 1)

participants currently manage or process their emotions (results are reflected in Table 7.5). Overall, we did not observe strong consensus around any single form of support type across emotions. Instead, participants frequently selected multiple support options, suggesting diverse and situational preferences and a lack of a singular, clearly defined support strategy. For **happiness**, participants preferred celebratory and affirming forms of feedback that allowed them to either share their joy or reflect internally. The most frequently selected option was "Help me feel a moment of gratitude" (33 selections, 53.2%), followed by "Give a motivational message" (30 selections, 48.4%) and "Celebrate with me" through a cheerful message or animation (27 selections, 43.5%). In contrast, a smaller subset of participants indicated that no follow-up was needed beyond logging the emotion ("No feedback needed", 7 selections, 11.3%), highlighting variability in desired engagement. The preferred tones are closely aligned with these forms of engagement. Participants most often selected a cheerful and playful tone (34 selections, 54.8%), followed closely by a calm and reflective tone (33 selections, 53.2%), indicating a preference for positive yet emotionally grounded feedback when experiencing happiness. These preferences were consistent with participants' reported real-life strategies for managing happiness. Most participants indicated that they typically *share their joy with someone* (46 responses) or *take a moment to feel gratitude* (37 times), followed by *engaging in energizing activities* (23) and *private reflection* (15 times). A smaller number reported simply *enjoying the moment without action* (10 times) or *eating good food* (2 times), reinforcing that both expressive and reflective responses to happiness are

Emotion	Preferred Support and Tone
Happiness	Celebratory and affirming feedback, including gratitude prompts, motivational messages, and cheerful acknowledgments (e.g., animations). Participants preferred feedback that felt *cheerful and playful*, while also valuing opportunities for *calm, reflective* engagement.
Sadness	Supportive and restorative feedback, such as gentle activity suggestions (e.g., walking, drinking water), mindfulness or grounding exercises, and empathetic validation messages. Preferred tones were *gentle, comforting, quiet, and empathetic*, emphasizing low-effort support.
Anger	Emotion-regulation–focused feedback, including calming exercises, physical release suggestions, and prompts to identify emotional triggers. Participants preferred a *neutral and factual* tone that supports control and de-escalation rather than validation or celebration.
Fear	Reassuring and socially oriented support, including encouragement to talk to someone, help identifying the source of fear, affirmations of safety, and grounding or distraction strategies. Preferred tones were *reassuring, supportive, and calming*.
Surprise	Lightweight acknowledgment or brief reflective prompts that allow users to process or savor the moment without deeper intervention. Participants preferred a *minimal and neutral* tone, particularly for positive surprise.
Disgust	Action-oriented and redirective feedback, such as suggestions to remove the trigger, shift attention, or engage in calming activities. Preferred tones were *neutral and suggestive*, focused on moving past the experience.

Table 7.5: Emotion-wise preferences for post-logging support and tone (Study 2)

common and should be supported. When participants experienced **sadness**, they preferred supportive and restorative forms of feedback. The most frequently selected option was *suggesting mood-lifting activities such as taking a walk or drinking water* (38 responses, 61.3%). This was followed by *mindfulness-based exercises* including grounding, breathing, or meditation (29 times, 46.8%), and empathetic validation messages such as *"It's okay to feel this way"* (28 times). Participants also valued reminders of past resilience (28 times) and *healthy distraction strategies* like listening to music, resting, or watching a comfort show (27 times). Fewer participants selected connecting with loved ones (21 times), suggesting that

while social support is valued, many users prefer low-effort, self-guided coping strategies when experiencing sadness. *Inviting reflection or journaling* was selected less frequently (19 times), indicating that while some users welcome deeper processing, it is not universally desired during sad states. Only a small number of participants preferred no feedback beyond logging (6 times), suggesting that most users expect some form of supportive response when experiencing sadness. Participants' real-life strategies for coping with sadness largely emphasized withdrawal and self-soothing. The most common responses involved spending *time alone* (36 times) and *resting or sleeping* (28), followed by *talking to someone for emotional support* (24 times). Many participants also reported using distraction (21 times) or continuing with necessary tasks without deliberate coping efforts (16 times). Comforting or soothing activities were mentioned by 17 participants, while fewer participants reported engaging in hobbies (4 times). Similarly, participants' preferred feedback tones were gentle and comforting, quiet and supportive, and empathetic, with some also valuing motivating, reflective, and suggestive styles of support. Overall, these patterns highlight a contrast with our proposed options, indicating a stronger preference for low-effort, restorative coping approaches over active or expressive strategies when experiencing sadness. A participant reflected on this - *"In an ideal scenario, I'd love to go on reflecting mode, when I'm sad. But it rarely happens with me. Usually, I hop onto Instagram, watch people's stories, or scroll. Sometimes, I go for a solo walk to soothe myself. So I would like the app to suggest to me good forms of distractions or exercises, and should later remind me to reflect."*

When experiencing **anger**, participants primarily preferred feedback that supported emotional regulation and release. The most frequently selected options were suggestions for calming activities such as deep breathing or stretching (31, 50%), followed by prompts for physical release or action (e.g., walking or squeezing an object) (29, 46.8%), and reflective prompts to help identify emotional triggers (25, 40.3%). These preferences are closely aligned with participants' real-life strategies for managing anger. Most participants reported taking time alone (35), using calming techniques (25), or engaging in physical release

(25). Others coped by writing out their thoughts (24) or talking the situation through with someone (13), reinforcing the emphasis on self-regulation and reflection over external validation when experiencing anger. Participants preferred a neutral, factual tone. Notably, participants expressed a stronger desire for app-based support when experiencing anger compared to other emotions. Many emphasized that anger feels particularly critical to manage because it can lead to harm to others as well as to oneself, whereas other emotional states were perceived as less immediately risky. When experiencing **fear**, participants primarily preferred socially oriented and reassuring forms of support. The most frequently selected options were suggestions to talk to someone (35 times, 56.5%) and help identify what was causing the fear (34 times, 54%), followed by distracting activities (26 times, 41%), affirmations of safety or coping ability (23 times), and engaging in talking the fear through (22 times). These preferences suggest that, unlike other emotions that may call for self-regulation or solitude, fear often prompts a desire for reassurance, sense-making, and interpersonal support. Most participants mentioned that advice on *"How to overcome your fears?"* is crucial and it is often hard to deal with extreme fear alone, and more specifically in the case of anxiety (when fear arises from an imaginary future), the need for someone to help or a toolkit of exercises and soothing music is necessary. For **surprise**, participants noted that such moments are relatively infrequent and typically do not require substantial support or intervention. When feedback was desired, participants preferred lightweight responses, such as prompts to briefly reflect on the surprising event or simple acknowledgment that allows them to savor the moment, particularly in cases of positive surprise. In cases of negative surprise, participants noted that the experience often served as a precursor to other negative emotional states (e.g., fear, anger, or sadness). As a result, they expressed a preference for support that helps them transition into identifying and processing the subsequent emotion. For **disgust**, we saw a preference for taking action to overcome it, identifying its causes, or engaging in calming activities to deal with it. The focus was on shifting the attention or removing the trigger. We also explored preferred feedback modalities: audio, visual, text,

interactivity, social, and gamified features. Participants favored soothing sounds or music playlists for audio, lightweight visuals like mood animations or calming images for visual feedback, and supportive or reflective text messages for text-based feedback. Interactive tools were preferred as short check-ins or links to self-care activities. Most participants did not favor gamified features, though a few appreciated mini challenges or achievement badges. For social features, participants preferred sharing with a trusted friend or access to an anonymous peer group to talk with. Access to mental health professionals was also valued. Overall, participants were open to multiple modalities, except gamification, which they found forced when dealing with emotions. These insights informed the design of the next phase, emphasizing user choice and flexibility, enabling participants to select preferred feedback types and modalities, and helping us better understand which approaches are most effective for supporting individual emotional experiences.

7.3 Phase 2: System Description

In this phase, we aim to study users' preferences, values, and points of friction in emotion logging through long-term, in-the-wild use of a research prototype. To do so, we developed a mobile application for both Android and iOS that serves as an investigative tool for observing how users engage with different emotion input, feedback, and support mechanisms over time. The goal of this system was not to design an optimal or complete emotion logging application, but to examine user engagement with participant-centric features, as articulated in RQ2a and RQ2b. The system design was informed by two sources: Phase 1 findings and insights from prior work on human-centered emotion tracking, self-reflection, and emotion regulation. Phase 1 examined how users prefer to log emotional experiences using different input interfaces and how they wish to receive feedback or support after logging, while prior work motivated the inclusion of features that reduce burden, support reflection, and accommodate varying emotional states. Based on these insights, the application provides multiple emotion input options and feedback modalities, allowing users to choose how and

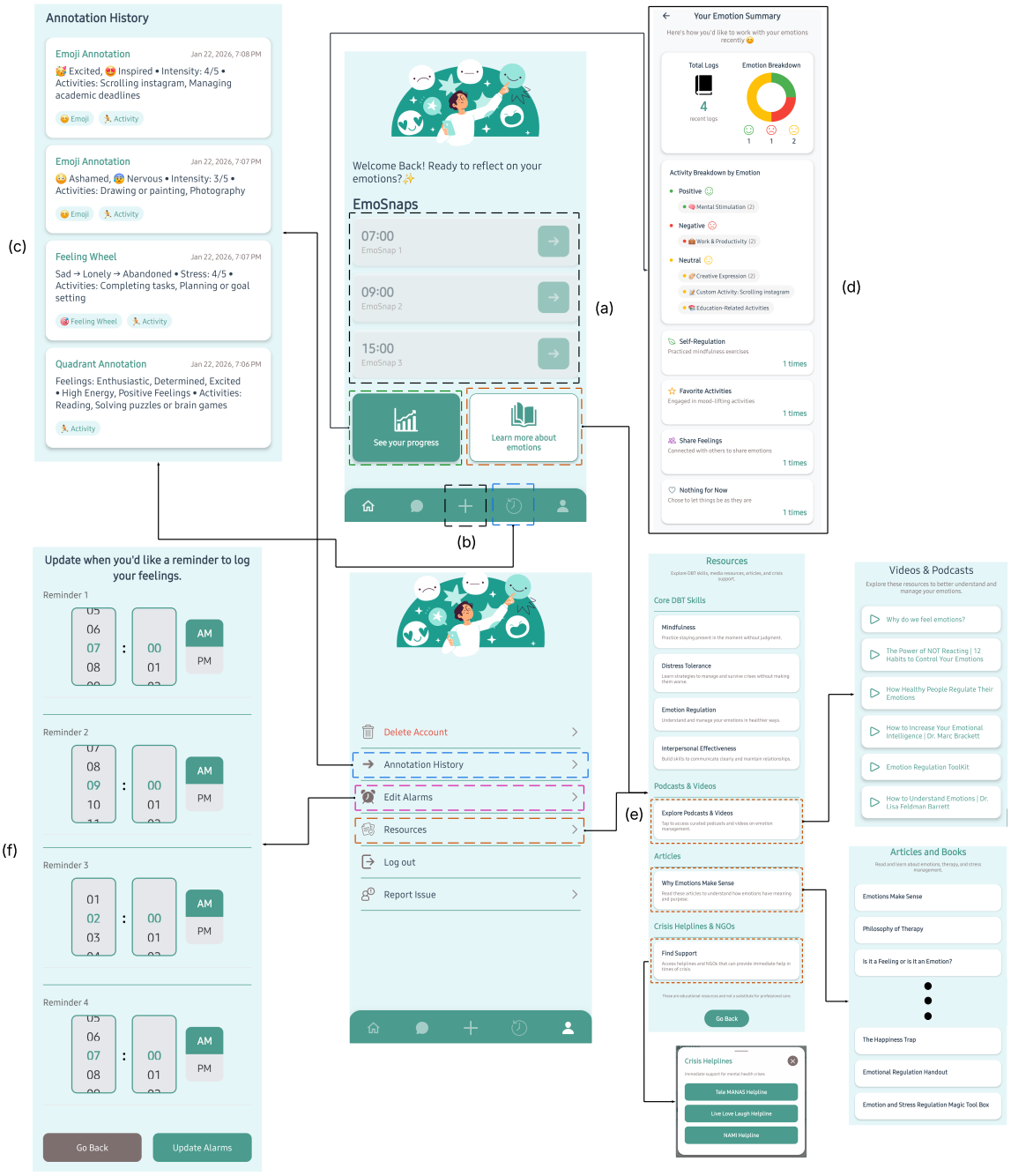


Figure 7.3: The figure illustrates our **"Application Overview"**. (a) The Pre-scheduled reminders (b) Impromptu logging (c) Annotation history, past record of all annotations made; (d) Progress Screen, (e) Educational Resources, (f) Reminder update screen (Best viewed in color).

when they engage. This choice-based design enables us to examine engagement behaviors, trade-offs between effort and expressiveness, and changes in usage patterns over a four-week field study, with the goal of understanding which design elements support or hinder sustained emotion logging. Our application consists of two primary design components: an emotion input interface for logging experiences, informed by Phase 1 findings, and post-logging feedback and support features, informed by Study 2, which focuses on providing feedback and support after logging. In the following sections, we describe the system in detail.

7.3.1 App Overview

We designed the application from a human-centric perspective, emphasizing user agency and choice based on our findings from chapter 6 and the previous phase. All features are designed to support emotion logging, post-logging support, and long-term feedback, while allowing users to decide how and when they want to engage. Rather than enforcing a single interaction flow, our system offers flexible input and feedback options that users can adapt to their routines, preferences, and emotional states, therefore enabling us to study their preferences in the long term. The application starts with a registration phase and a splash screen that explains the main features of our application: the three annotation modes, post-log processing, and feedback. After this, users can set their initial four notification slots according to their routines. Within our application, the **Home screen** functions as the central dashboard, providing users with an overview of prescheduled emotion reminders, access to past entries, progress summaries, additional resources, a chatbot, profile settings, and a floating action button for spontaneous emotion logging. The application supports two logging modes to meet different user needs. **Scheduled logging** uses notification-based reminders, which are set up by users (four daily check-ins) during onboarding, visible on the Home screen, which can be modified at any time through the Edit Alarms feature in **profile screen** to fit personal routines (Figure 7.3). Users also have an **Impromptu logging** option that lets them record emotions at any time via the floating action button at the bottom of

the home screen. We have also added a profile screen, where users can **review past entries**, **manage notification settings**, **explore curated articles**, videos, and external mental health resources, and report issues (more details are added in the appendix section E.2). These additional features are designed to support the users in the core logging activity. In addition to structured emotion logging, the application includes a GPT4.1-powered **chatbot** that allows users to reflect on and discuss their emotional experiences through open-ended conversation, complementing the annotation-based logging interfaces (see appendix E.5 for more details). Our application overview is illustrated in Figure 7.3. In the following sections, we describe the two core components of the application: the emotion input interfaces used for logging experiences and the post-logging feedback and support mechanisms designed to accompany those entries.

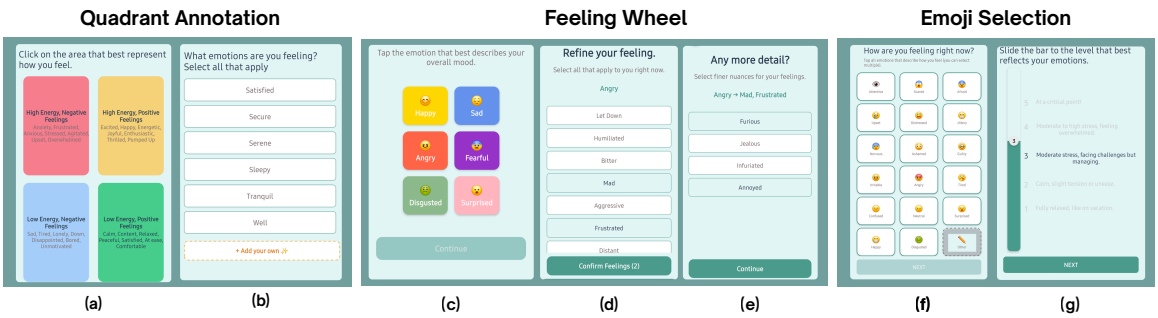


Figure 7.4: The figure illustrates the **"three annotation modes"** present in our application (Best viewed in color).

7.3.2 Emotion Logging Interfaces

Based on findings from Phase 1 Study 1, we integrated the three highest-rated emotion annotation interfaces into the application. These interfaces were selected to reflect variations in users' emotional awareness, vocabulary, and desired logging effort. At each logging instance, users can freely choose among these interfaces, allowing them to adapt how they report emotions based on their current context, emotional intensity, and availability (Figure 7.4). The first interface is a **quadrant-based annotation** method grounded in a valence–arousal model of emotion. Users begin by selecting one of four quadrants

representing combinations of high or low valence and arousal. They are then presented with a curated list of emotion labels associated with the selected quadrant, from which they may select multiple labels or can add their own emotion labels to capture concurrent, overlapping, or additional emotional states (Figure 7.4a–b; Table E.10). This interface balances guided emotional vocabulary with flexibility in expression. The second interface is a **feeling wheel** that follows a hierarchical, multi-stage selection process. Users first select one of six broad core emotions, then refine their choice through two subsequent tiers of increasingly specific emotion descriptors, resulting in a structured three-level annotation (Figure 7.4c–e; Tables E.8 and E.9). This design supports users who prefer a more structured, reflective approach to identifying emotions and want support in narrowing their emotions from broad categories to specific, nuanced emotions. The third interface is an **emoji-based annotation** method that enables lightweight emotion logging by allowing users to select an emoji that best represents their current emotional state. In addition to a predefined set of emojis, users can add a custom emoji to support personalized expression (Table E.11). This interface is intended to reduce logging effort and support quick, in-the-moment emotional reporting. The emoji-based annotation interface was designed by curating emotion labels from publicly available emoji databases and extending them with emotions in the PANAS scale [322]. This approach allowed us to cover both commonly used emoji expressions and psychologically grounded emotional constructs, while supporting familiar and lightweight emotional reporting. Following the completion of the emotion annotation with either one of the user-selected annotation modes, the user proceeds through three assessments added in the app:

1. **Stress Intensity Scale:** A 5-point Likert scale (1–5) measuring the user's current perceived stress level at the time of annotation. Refer to table E.7 and figure 7.4g. This scale was adapted from the Perceived Stress Scale (PSS) [289] to capture stress intensity, which is often measured separately in emotion self-tracking [37]. To improve usability, each scale point was accompanied by a short descriptive label. However,

for quadrant annotation, stress scale was conditionally available and only applicable when users select high-arousal, negative-valence states.

2. **Confidence Scale:** A 5-point Likert scale (1–5) capturing the user's self-reported confidence in the accuracy of their emotion annotation. This step was added to encourage self-reflection [265]. Refer to table E.12.

3. **Activity Selection:** A multi-category checklist allowing users to select all possible activities performed since their previous annotation, with an optional free-text field for additional activities. The activities were divided into broader-level categories to make it easier for users to select. This option was designed to help users reflect on possible triggers of their emotions [63]. The full list of activity options is detailed in Table E.13.

7.3.3 Post-Logging Feedback and Support

After completing emotion logging and the three follow-up assessments, users are directed to a post-logging support screen (Figure 7.5). This module is designed to encourage active emotional engagement by offering visualizations, optional, user-selected coping and reflection actions [88]. Consistent with the system's human-centric design, users retain full autonomy to choose strategies that align with their current emotional state and personal preferences. The post-logging support options were informed by findings from Phase 1, Study 2, which showed substantial variation in how users prefer to process emotions and a lack of consensus around a single preferred mechanism. To accommodate this diversity, we grouped support features into five broad categories of emotional processing:

- **Self-regulation**: This option provides users with a curated set of short instructional videos focused on relaxation and grounding techniques. The videos used in the application are listed in Table E.14 and shown in Figure 7.5a. These animated videos

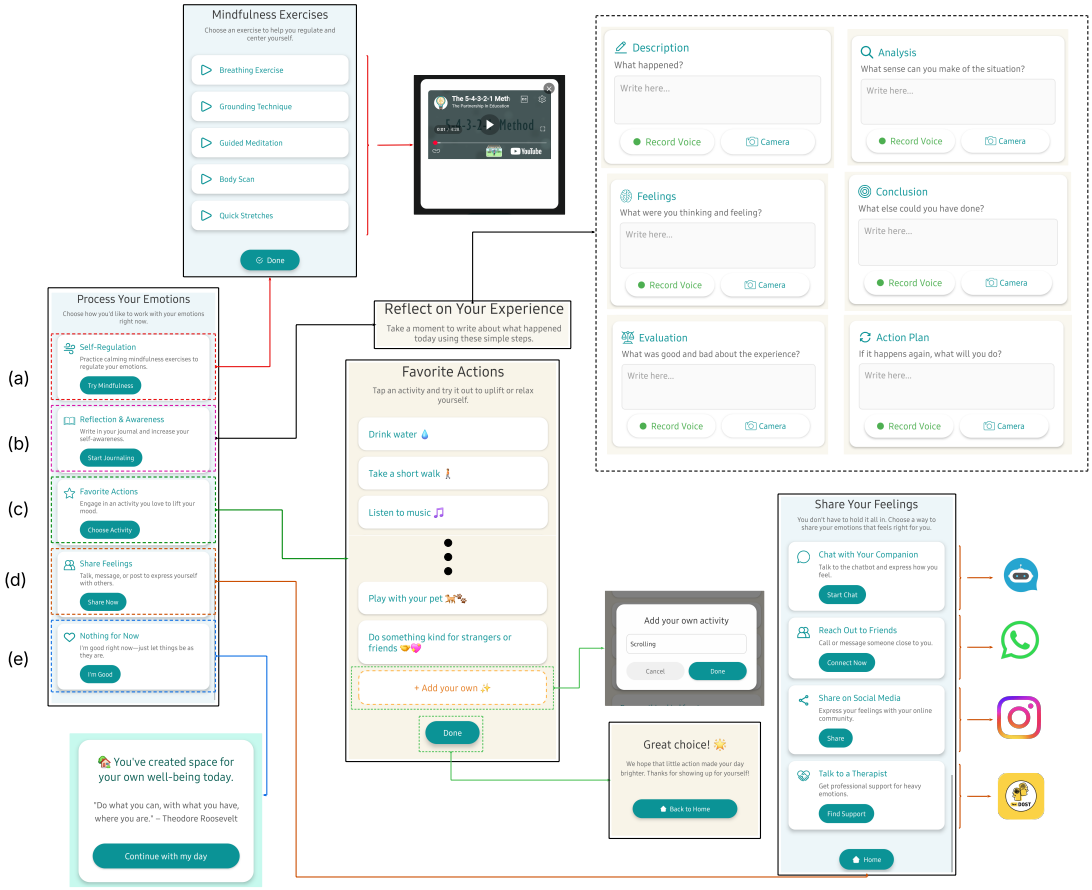


Figure 7.5: The figure illustrates the **"Process Your Emotion"** section of the application, presented after each of the three emotion logging methods. The section includes: (a) Self-regulation, offering links to resources such as breathing exercises and guided meditation videos; (b) Reflection and awareness, providing journaling prompts based on Gibbs' Reflective Cycle; (c) Favorite activity, suggesting activities users can engage in; (d) Share feeling, allowing users to share their emotions; and (e) Nothing for now, enabling users to skip process your emotion screens (Best viewed in color).

are sourced from YouTube, selected for clarity of explanation, production quality, and viewership, and embedded directly within the app for ease of access.

- **Reflection and Awareness**: This option provides a structured reflective journaling experience. Journaling was added to help users reflect on and process their emotional experiences. Based on Phase 1 findings, users expressed interest in having a structured way to explore their feelings beyond simply logging them. The journaling feature guides users through prompts derived from Gibbs' reflective cycle [326], helping them describe events, articulate feelings, evaluate situations, analyze causes, and plan future actions. Users can respond with text, images, or audio recordings as per their preferences (as shown in Figure 7.5b). The journal includes the following prompts:
  - ***Description***: What happened?
  - ***Feelings***: How did you feel?
  - ***Evaluation***: What was good or bad?
  - ***Analysis***: Why did it happen?
  - ***Conclusion***: What did you learn?
  - ***Action Plan***: What will you do next?
- **Favorite Action**: This option displays a list of 31 common activities intended to support emotional regulation (see Figure 7.5c). Users can either select a predefined option or add a custom activity. This option was added to suggest activities that might help them cope with their emotions; the list of activities is adapted from the dialectical behavior therapy's list of distracting activities [1]. The final list of activities is shown in Table E.15.
- **Share Feelings**: This option allows users to externalize their emotions by sharing them within their social circle or with professional support sources. The available sharing

[1] `https://dialecticalbehaviortherapy.com/distress-tolerance/distracting-activities/`

options are summarized in Table E.16. We included the Share Feelings option to support users who prefer social or external coping strategies over individual reflection. Findings from Phase 1 and prior work [2] indicated that many participants sought relief through talking to trusted people or professionals rather than engaging solely with in-app tools.

- **Nothing for Now**: This option allows users to conclude the logging process without engaging in additional activities. Selecting it displays a brief positive acknowledgment (e.g., *"Checking in with yourself is a form of self-care."*) along with an inspirational quote (e.g., *"Knowing yourself is the beginning of all wisdom."*). These messages are drawn from curated lists and are intended to provide gentle emotional closure without requiring further effort. This option was included based on Phase 1 findings, which indicated that users, particularly when experiencing neutral or low-intensity emotions, sometimes preferred not to engage in post-logging actions.

Additionally, our application includes **annotation history**, accessible from the Profile page (Figure 7.3(c)). This screen provides users with a chronological overview of their past emotion logs, summarizing the annotation type, selected emotions, and associated activities. Additionally, our application has **track your progress**, which is also accessible from the Home screen (Figure 7.3(d)), and provides users with aggregated insights from their past emotion logs. This presents descriptive summaries of emotional patterns, including logging frequency, distribution of emotional valence, activities, and emotional valence correlation, and frequency of engagement with various emotion regulation strategies after logging. This module is intended to help users reflect on trends in their emotional experiences and support self-awareness over time. The post-logging support, annotation history, and track-your-progress screens were included to address a key insight from prior work, that long-term reflection on emotions helps users learn from their experiences and develop better emotional awareness and regulation over time [2, 96]. By providing structured reflection options immediately after logging, alongside the ability to review and analyze past entries over time,

the application encourages users to connect present emotional experiences with historical context, supporting ongoing learning and self-insight.

7.4 Phase 2: In-the-Wild Deployment

In this section, we will discuss our field study design in detail.

7.4.1 Study Design

The application was deployed for a four-week field study to examine user engagement over an extended period. Participants received an invitation email that included installation instructions, an overview of the application, a usage guide, and a tutorial video explaining all the application's features (also available as a supplementary document). The email also outlined informed consent, data confidentiality, usage for research purposes, and participants' right to withdraw from the study at any time. Android users received an APK file, while iOS users received an App Store link. To support onboarding, participants received reminder emails and WhatsApp messages within the first 2 days after the invitation. During deployment, participants received up to 4 daily reminders, with notification times initially configured by participants during onboarding and adjustable at any time via the profile screen, as described earlier in section 7.3. Participants were asked to use the app according to their own needs, without any requirement to complete a fixed number of logs each day. This approach allowed us to observe natural engagement trajectories without imposing any constraints. To gather ongoing feedback without overburdening participants, we administered three feedback surveys throughout the study: the first at the end of the first week, the second between the second and third weeks, and the last at the end of the study. The first two weekly surveys focused on ease of integration into daily routines, perceived cognitive load, usage barriers, usefulness of emotion expression and feedback, trust in the application, and early behavioral changes. Example questions included: *"In the first week, how mentally demanding did you find using the app?", "What felt like the biggest source*

of resistance or hesitation when logging emotions?", and *"Did you notice any changes in your daily habits during the second week of use?"*. We have added the detailed survey in appendix E.1. For the first two weeks, the research team maintained light engagement with participants through a small number of informational emails (five in total, sent at irregular intervals) that emphasized the purpose of emotion logging and encouraged exploration of the features. These communications were intentionally non-prescriptive and framed as optional guidance rather than reminders. After the initial two weeks, no further engagement prompts were sent, allowing participants to use the application organically. At the end of the four-week period, participants were notified via email and WhatsApp that they could uninstall the application at their convenience, with no obligation to continue usage. A final, more detailed feedback survey was administered, capturing overall experiences, perceived benefits and challenges, feature usage patterns, and reflections on long-term engagement. Example questions included: "*How would you rate your overall experience with the app?", "What did you find most helpful?",* and *"Which features did you use most and least, and why?"* The study design is illustrated in figure 7.1 phase 2.

7.4.2 Participant Recruitment

Participants were recruited using a combination of snowball sampling [294] and convenience sampling [217]. Study flyers were shared through the researchers' social media channels and institutional mailing lists, along with an interest form for potential participants. In total, 81 individuals expressed interest and received an invitation email. Of these, 42 participants installed the application and took part in the study. One participant withdrew after the first week due to a lack of interest, and four participants used the app only once, including the participant who withdrew in week 1. All participants were over 18 years of age, participated voluntarily without financial incentives, and provided informed consent. No exclusion criteria were applied based on mental health history or prior experience with emotion logging, as the study aimed to capture a broad range of psychosocial backgrounds

Demographic	Participants
Age	18–25 (28), 26–35 (11), 36–45 (2), 55+ (1)
Gender	Female (21), Male (21)
Highest Education	High School (11), Graduation (19), Post-graduation (10), PhD / Doctorate (2)
Employment	Student (28), Software Engineer / IT (5), Teacher / Professor (3), Engineer (non-IT) (1), Manager / Executive (2), Consultant (1), Entrepreneur (1), Homemaker (1)
Previous Experience with Emotion Tracking	None (24), Mood tracking apps (8), Manual journaling (5), Wearable devices (4), Clinical tracking (1)
Mental Health Background	No concerns/history (19), Self-identified concerns (8), Prior therapy/counseling (6), Diagnosed, not in treatment (4), Therapy with medication (3), Prefer not to say (2)

Table 7.6: Demographics of Participants in the Field Study (N = 42)

and emotional experiences. All participants were comfortable with English and had high digital literacy. Participant demographics are summarized in Table 7.6.

7.4.3 Analysis

To address RQ2a, we conducted a mixed-methods analysis combining longitudinal usage logs with repeated self-reported feedback collected during a 4-week field deployment. We analyzed application log data to characterize engagement patterns over time, including logging frequency, interaction modality choice, and use of post-logging support features. We further applied clustering techniques and linear mixed-effects models to examine how participants engaged with the three emotion input interfaces, including whether they consistently relied on a single interface or switched between interfaces over time. Engagement with post-logging support was examined by quantifying selections across the five support categories and analyzing their co-occurrence with different emotional states. In addition to the planned four-week deployment, a subset of participants continued to use the application beyond the study period despite receiving uninstall reminders. We analyzed this extended-use data separately to explore patterns of voluntary engagement beyond the study duration. Participants also completed three surveys during the study: two brief weekly surveys and a

final end-of-study survey. Closed-ended responses were analyzed descriptively to capture changes in perceived effort, integration into daily routines, perceived usefulness of logging, and feedback over time. Open-ended responses were analyzed using inductive thematic analysis [183], with themes refined through iterative coding by two authors. Finally, we triangulated behavioral logs and self-reported data to develop a holistic understanding of long-term engagement with the application.

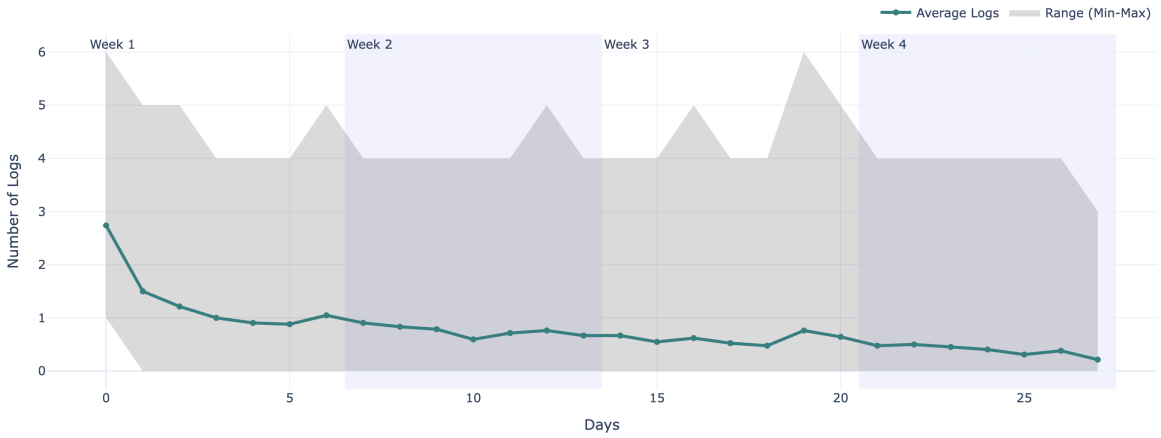


Figure 7.6: The figure illustrates the overall engagement with our application (Best viewed in color).

7.5 Phase 2: Results

In this section, we report findings from our field deployment. We first present a descriptive analysis of participants' engagement with the application, including use of core features and reflections captured through weekly feedback. We then identify distinct engagement profiles and examine the behavioral and interactional factors that differentiate them, offering insight into what shaped participants' engagement levels over time and addressing our RQ2a and RQ2b.

7.5.1 Overall Engagement Patterns and Participant Experiences

We first examined overall engagement patterns across the four-week deployment (Figure 7.6) to understand how participants' interaction with the application evolved over time.

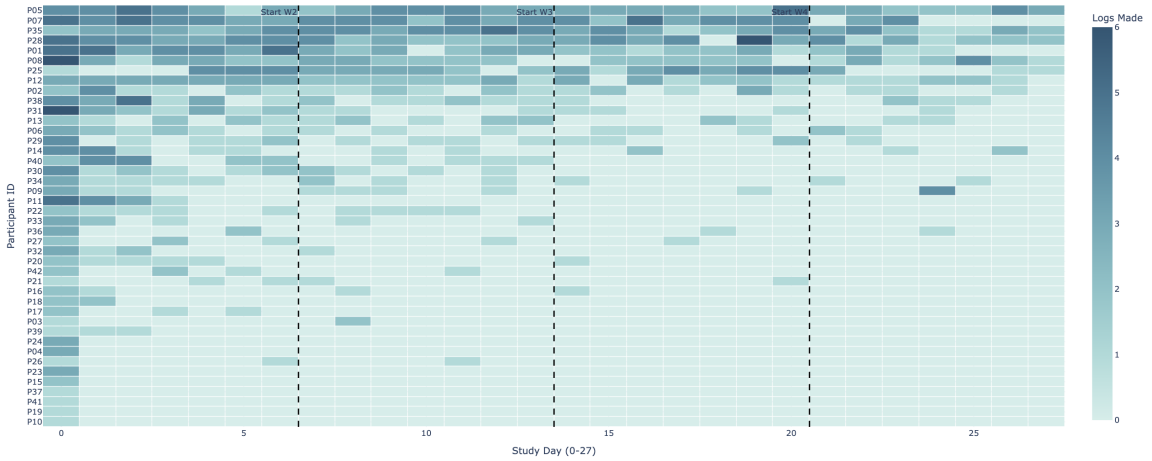


Figure 7.7: The figure illustrates the participant-wise engagement with our application over the study duration (Best viewed in color).

Participants received up to four daily notifications, which they could schedule at convenient times, allowing logging to be integrated into their everyday routines. Logging activity showed clear temporal patterns. During Week 1, average daily logs were highest on Day 0 (2.74 logs/day), reflecting initial curiosity and exploration, and subsequently fluctuated between 0.88 and 1.50 logs/day for the remainder of the week. In Week 2, engagement declined slightly, with averages ranging from 0.60 to 0.90 logs/day, while Week 3 remained relatively stable between 0.52 and 0.76 logs/day. By Week 4, the average logs decreased further to the lowest of 0.21 logs/day. Despite this gradual decline, the maximum number of logs per day remained above 3 on nearly all days, indicating that a subset of participants continued to engage multiple times per day, even as overall participation waned. A closer examination of participant-level engagement (Figure 7.7) revealed substantial heterogeneity in these trends. We observed that eight participants did not log beyond the first day, despite initial enthusiasm and motivation to better understand their emotions. On Day 0, these users collectively logged 15 entries, with P15 logging twice; P24, P04, and P23 logging three times each; and P10, P19, P37, and P41 logging once. Out of these participants, P41 formally withdrew from the study within the first week, stating, *"I am currently quite busy and unable to incorporate this habit into my routine,"* while P10 encountered app compatibility

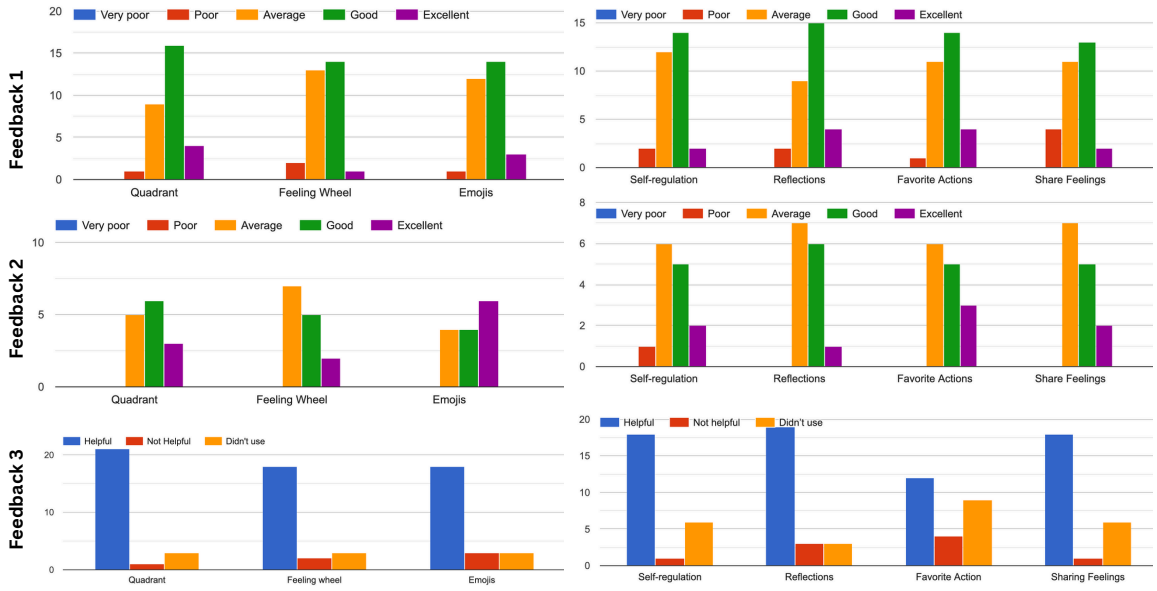
issues leading to withdrawal. The remaining participants did not withdraw formally but remained inactive after day 0. Among those who continued after day 0, engagement varied considerably: some participants maintained regular logging habits across weeks, while others engaged more intermittently.

To better understand the reasons behind these different logging styles, we analyzed our two initial feedback surveys, which focused on users' experiences. In the first week, 10% of participants reported that the app was not mentally demanding, increasing to 23.1% in the second week (Figure 7.9). Despite this shift, many participants continued to perceive the app as slightly or moderately demanding. Further analysis revealed that the most frequently cited barrier to app use was difficulty remembering to log, with participants reporting that logging had not yet become part of their routine or a habitual action (Week 1: n = 18; Week 2: n = 10). The second most common reason was low motivation or interest in logging emotions regularly (Week 1: n = 13; Week 2: n = 6). Finally, participants also reported time constraints due to busy schedules as a barrier (Week 1: n = 12; Week 2: n = 6). Additionally, a small number of participants (Week 1: n = 9; Week 2: n = 5) reported difficulty articulating their emotions, noting challenges in precisely identifying or labeling their feelings. Some described their emotions as vague or hard to name, while others indicated that the available options did not always capture their experiences or that emotions often repeated between logging periods. Overall, this suggests that, beyond motivation and time constraints, limitations in emotional articulation and expressiveness also contributed to resistance in using the app, indicating that difficulties in identifying, labeling, or differentiating emotions may have affected participants' engagement and consistency in logging. A few participants (Week 1: n = 6; Week 2: n = 7) also reported not observing any meaningful changes in the first two weeks. We also observed that during the first week, more participants reported increased emotional awareness (n = 11) and an enhanced ability to pause and reflect (n = 9). By the second week, these numbers declined, while the number of participants reporting no significant improvement increased, suggesting that the app's

perceived benefits may not have been sustained for all users over time. Participants overall shared positive feedback towards the three interfaces and post-logging support as shown in Figure 7.8.

We further observed within our feedback surveys that participants' perceptions of app features evolved over time. During Week 1, the three logging options (Quadrant, Feeling Wheel, Eand moji) were rated as the most useful features (n = 21). By Week 2, the ability to view emotion log history (n = 10) and to log emotions spontaneously (n = 10) were considered most valuable. Features such as log history and progress tracking remained useful for participants through the end of the study (reported by n = 11 and n = 8, respectively). Habit formation, however, remained challenging, with 21 participants reporting difficulty establishing a consistent logging routine at the end of the study. Furthermore, many participants reported that the resources section within the app was among the least used features, citing a lack of time to explore it. The chatbot was also among the least engaged features, as participants commonly noted that interactions felt superficial and lacked human-like support for mood regulation. Although some users reported improved emotional awareness (n = 12), understanding emotions remained difficult for many (n = 9), and consistent logging continued to be a challenge (n = 21). At the conclusion of the study, 10 participants expressed a strong intention to continue using the app, while another 10 were uncertain about maintaining consistency. Taken together, these findings suggest that while initial interest in emotion logging was high, sustained engagement depended on participants' ability to integrate the practice into their daily lives and to perceive ongoing value from it.

We also examined participants' engagement after the official study period to understand whether and how users chose to continue interacting with the app once study-related obligations were removed, during which participants were asked to uninstall the application at their convenience. Over 90 days, 15 participants contributed 132 emotion logs (M = 8.8 per user). Engagement was highly concentrated among a small subset of users: the top five contributors logged 41 (P07), 21 (P08), 13 (P28), 12 (P35), and 8 (P12) entries, respectively. Only P07



(a) Input Interfaces

(b) Post-Logging Support

Figure 7.8: This figure presents participants' ratings of the two (a) input interfaces and (b) post-logging support collected over four weeks (Feedback 1: n = 30; Feedback 2: n = 14; Feedback 3: n = 25) (Best viewed in color).

remained active until the final day, representing 6.7% of participants. Average activity declined steadily, dropping below 50% of initial levels by Day 38. Further demographic analysis revealed that participant P07 was currently in therapy, taking medication, and had prior experience with journaling under clinical supervision. These observations suggest that, even without incentives, the app can support sustained, voluntary engagement for users with prior awareness and perceived need for tracking.

7.5.2 Participant Engagement Profiles

Our earlier analyses (Section 7.5.1) indicated that participants engaged with the app in diverse ways, suggesting the presence of different engagement trajectories in our data. To further characterize these patterns, we conducted a participant-level clustering analysis. We performed our analysis based on four key engagement metrics derived from our data, including: **(1) Total emotion logs**, which captured the overall number of emotion entries recorded by each participant, reflecting logging volume, **(2) Emotional convergence**,

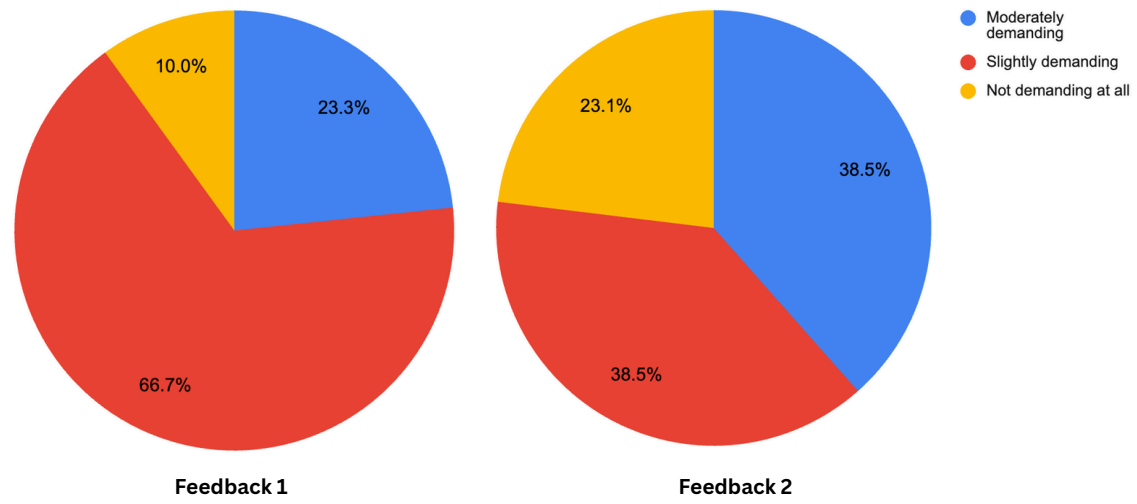


Figure 7.9: This figure presents participants' ratings of the mental demand associated with using the app in their daily lives (Feedback 1: n = 30; Feedback 2: n = 14).

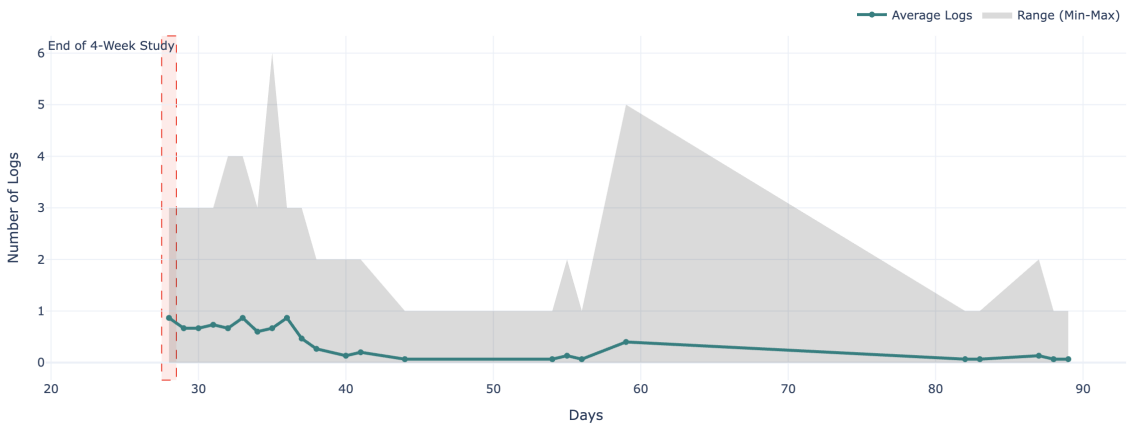


Figure 7.10: This figure presents participants' post-study log over time (Best viewed in color).

the degree to which participants' reported emotions stabilized or converged over time, capturing patterns of emotional sense-making, **(3) Interface breadth**, the range of emotion-input interfaces used by participants, indicating diversity in interaction modalities, and **(4) Processing depth**, the usage of post-logging supportive features, reflecting reflective depth beyond simple logging. These metrics were chosen to capture both the quantity and quality of engagement. Together, these measures represent engagement as a multidimensional construct, enabling us to distinguish distinct engagement styles. We standardized the engagement metrics and applied k-means clustering to identify distinct participant profiles,

selecting k = 3 based on the elbow method to balance interpretability and cluster separation. To assess the validity of the identified clusters, we analyzed additional indicators of sustained engagement, including longevity of use (the number of days between a participant's first and last recorded interaction), number of active days (the total days on which at least one log was recorded), and consistency of daily logging (the proportion of active days relative to the total engagement period). Examining these complementary measures allowed us to determine whether the clusters reflected meaningful differences in long-term engagement, rather than merely variations in the four primary metrics, thereby providing a more robust understanding of participants' usage patterns over time.

Our results from the cluster analysis identified three distinct participant groups, reflecting different patterns of engagement with the application. Cluster 0 comprised 7 participants, Cluster 1 comprised 14 participants, and Cluster 2 comprised 21 participants. These clusters differed substantially across both core engagement metrics and long-term usage indicators. Participants in Cluster 0 exhibited the highest overall engagement. On average, they recorded 72.71 emotion logs, indicating frequent and sustained use of the application. They also demonstrated strong convergence in emotional reporting (M = 0.83), broad interaction with interface features (M = 3.14), and substantial process depth (M = 71.29), suggesting deep and reflective engagement. In terms of long-term usage, this group showed the longest engagement period, with an average longevity of 27.29 days and 25.29 active days. Their daily consistency was moderate (M = 0.36), indicating relatively regular logging throughout their usage period. Together, these results characterize **Cluster 0** as **highly engaged and consistent users** who interacted extensively with the app over time. Cluster 1 participants displayed the lowest levels of engagement across all measures. They recorded an average of only 4.43 logs, with low convergence (M = 0.25), limited interface use (M = 1.71), and shallow process depth (M = 3.43), reflecting minimal interaction with the application. Their long-term engagement indicators further support this interpretation. On average, these participants used the app for only 5.50 days, with 2.71 active days and low daily consistency

(M = 0.19). This pattern suggests brief, irregular, and largely surface-level engagement with limited interest or motivation to explore the app. Accordingly, **Cluster 1** represents **minimally engaged users with a high barrier to engaging with emotions**. Finally, participants in Cluster 2 exhibited moderate but distinctive engagement patterns. On average, they recorded 15.86 logs and demonstrated high convergence (M = 0.85), comparable to Cluster 0, indicating similarly consistent emotional labeling when they engaged with the application. They also showed substantial interface breadth (M = 2.90) but relatively limited process depth (M = 8.86), suggesting concise and goal-oriented interactions with limited reliance on post-logging support features. In terms of sustained use, this group maintained engagement over relatively extended periods (M = 18.38 days), with an average of 9.38 active days, indicating selective rather than continuous participation. Notably, Cluster 2 exhibited the highest daily consistency (M = 0.45), reflecting regular use during active phases despite lower overall logging volume. Taken together, these patterns suggest that Cluster 2 participants primarily engaged in to-the-point, opportunistic self-reporting, using the application mainly during salient emotional moments rather than for continuous reflection. Accordingly, **Cluster 2** is characterized as **moderately engaged users oriented toward reporting self-perceived meaningful emotional experiences**.

Together, the three clusters indicate that a human-centric emotion-logging application must support multiple long-term engagement trajectories that reflect diverse user needs, motivations, and emotional practices. Rather than promoting a single mode of use, the system should enable users to flexibly appropriate emotion logging in ways that align with their everyday routines, reflective goals, and evolving engagement preferences. It includes users who seek deep, continuous reflection, those who prefer selective, situational reporting, and those who require low-effort entry points and early value to sustain participation. Moreover, the presence of distinct engagement profiles underscores the importance of designing for flexibility, progressive complexity, and early, meaningful feedback mechanisms to better address varying barriers to use and remain relevant as users' needs change over time.

7.5.3 Decoding Emotional Engagement

In this section, we examine differences in emotional engagement with the three user interfaces across the engagement profiles identified in Section 7.5.2. The objective of this analysis was to understand how users with distinct behavioral engagement profiles vary in their emotional interactions with the system. Drawing on prior literature, we hypothesized that engagement patterns are shaped by users' emotional states, such that individuals experiencing higher emotional intensity, particularly negatively valenced emotions, may demonstrate greater engagement due to a heightened perceived need for interaction [2, 29]. Because the three interfaces differ substantially in their design and emotional input structures, we first standardized the emotional data to enable meaningful cross-interface comparison. Specifically, all emotion entries were mapped to the four quadrants of the arousal–valence circumplex model [134]. We adopted this model due to its broad applicability for representing emotions in a consistent, interface-agnostic manner [4]. To support this standardization, we developed a comprehensive mapping dictionary that assigned each emotion label used in emoji or feeling wheels to one of the four quadrants based on its arousal and valence characteristics, as defined in the circumplex model [134]. Emotion entries from the quadrant-based interface were already natively aligned with this representation. This unified framework enabled direct comparison of emotional logs across all three interfaces. We operationalized emotional engagement as a composite construct comprising two complementary components: intensity and breadth. Intensity (affective component) captured the self-reported magnitude of the emotion, derived from the stress-intensity scale for each interface and normalized to 0–1. Breadth (cognitive component) reflected the level of emotional differentiation within each log, quantified by the number of distinct emotions selected in multi-select inputs across the interfaces (e.g., selecting "Anxious, Overwhelmed, Stressed" indicating greater breadth than selecting only "Anxious") and similarly normalized to a 0–1 range. The final Emotion_Engagement_Score was computed as an equal-weighted average of normalized intensity and breadth, yielding a unified and robust measure of

emotional engagement for each emotional log.

To assess whether engagement profiles influence emotional interaction with the app, we used a series of linear mixed-effects models. The first model examined whether overall Emotion_Engagement_Score differed across the three engagement clusters, with Cluster_ID as a fixed effect and participant identity included as a random intercept to account for individual variability. To explore whether engagement profiles vary across emotional contexts, a second model incorporated Emotion Type (the four valence-arousal quadrants) and its interaction with Cluster_ID as fixed effects. Results from Model 1 revealed no statistically significant differences in overall emotional engagement across engagement profiles. The reference group (Cluster 0) exhibited a baseline mean engagement score of $\beta_0 = 0.37$ ($SE = 0.04$, $p < .001$). Relative to this baseline, neither Cluster 1 ($\beta_1 = -0.03$, $p = .59$) nor Cluster 2 ($\beta_2 = -0.08$, $p = .07$) showed a significant deviation, although Cluster 3 exhibited a marginal trend toward lower engagement. While Model 2 revealed a **significant** main effect of *Emotion Type*, indicating that the specific emotion being reported was a stronger predictor of emotional engagement than users' behavioral engagement profiles. Participants demonstrated the **highest engagement when reporting high arousal negative emotions** (e.g., stress, anger; intercept $\beta = 0.43$). In contrast, all other emotion quadrants were associated with significantly lower engagement scores relative to this baseline, with Low arousal Positive emotions (e.g., calm, content) eliciting the lowest engagement ($\beta = -0.12$, $p < .001$). Crucially, none of the interaction effects between cluster membership and emotion type were statistically significant, indicating that the observed negativity bias in emotional engagement, characterized by emotional valence, particularly high-intensity negative states, was consistent across engagement profiles.

To disentangle the contributions of affective intensity and cognitive breadth to overall emotional engagement, we ran two separate linear mixed-effects models. Model 3 predicted normalized intensity, and Model 4 predicted normalized breadth, with Cluster_ID, Emotion_Type, and their interaction as fixed effects and participant identity as a random

intercept. Model 3 revealed a **significant** main effect of emotion type. Participants reported higher intensity for **high-arousal negative emotions** (intercept $\beta = 0.312$, $p < .001$), with all other quadrants eliciting significantly lower intensity ratings: **high-arousal positive** ($\beta = -0.117$, $p = .001$), **low-arousal negative** ($\beta = -0.094$, $p = .010$), and **low-arousal positive** ($\beta = -0.166$, $p < .001$). No significant main effects were observed for **Cluster_ID** (Cluster 1: $\beta = 0.067$, $p = .528$; Cluster 2: $\beta = -0.012$, $p = .862$), and none of the interaction terms between cluster and emotion type reached significance ($p > .05$), indicating that the pattern of higher intensity for negative, high-arousal emotions was consistent across engagement profiles. Model 4 showed a somewhat different pattern. The main effect of emotion type was significant only for **low-arousal positive emotions**, which were associated with narrower emotion reporting ($\beta = -0.082$, $p = .021$). In contrast, all other emotion types did not differ significantly from the high-arousal negative emotions, suggesting similar usage of multi-selection scales across emotional type, except low-arousal positive emotions, which are inherently simpler to label. Cluster membership again had no significant effect on breadth (Cluster 1: $\beta = 0.062$, $p = .623$; Cluster 2: $\beta = -0.106$, $p = .263$), and all interaction terms were non-significant. Overall, these results suggest that the observed negativity bias in emotional engagement is primarily driven by intensity rather than breadth, and that both components operate consistently across behavioral engagement profiles.

Overall, the analyses indicate that emotional engagement patterns are largely independent of long-term behavioral engagement profiles. Across clusters, participants consistently exhibited similar levels of engagement regardless of whether they were highly engaged, moderately engaged, or sporadic users. The key determinant of engagement was the type of emotion being logged. Specifically, all participants showed a negativity bias, reporting higher intensity and shared greater detail for high-arousal negative emotions, while positive and low-arousal emotions elicited lower engagement. These results suggest that the app successfully supports diverse user types without requiring high or continuous usage for meaningful emotion reporting. Engagement differences are driven by emotional salience, with users

naturally investing more effort when logging intense or distressing emotions. This highlights the importance of designing human-centric emotion-logging systems that prioritize context-sensitive interactions while supporting both selective, opportunistic reporting and sustained reflective use.

7.5.4 Understanding Reflective Engagement

In this section, we investigated the factors that drive reflective engagement, examining whether such behavior is primarily shaped by users' engagement profiles, their momentary emotional states, or the intensity of their emotional experiences. We operationalized reflective engagement using interaction logs that captured users' engagement with post-logging features, including journaling, self-regulation exercises, sharing tools, and favorite activities. From these logs, we derived two complementary measures of reflection. **(1) Reflective breadth** captured the diversity of reflective activities within a session. For each session, we computed whether participants engaged in any of the post-logging features and summed these binary indicators to obtain a breadth score ranging from 0 to 4. **(2) Reflective depth** captured the intensity of engagement. We computed a composite score by combining normalized journal word count, weighted sharing behavior, raw self-regulation counts, activity participation, and a media attachment bonus, emphasizing sustained effort and multi-modal engagement. **(3) Reflection propensity** captured whether any reflective behavior occurred within a session. Sessions with nonzero reflective breadth were coded as reflective, whereas sessions in which participants selected no post-logging features were coded as non-reflective.

To examine the drivers of reflective engagement, we employed three complementary models. Our first model examined whether emotional context predicts whether users engage in reflection. We modeled the binary outcome reflection propensity using a GEE-based logistic regression model at the participant level. Fixed effects included emotion type, emotional engagement score, and engagement cluster membership. Our second

model examined whether emotional context or behavioral profile predicts how deeply users engage once reflection occurs. We estimated a linear mixed-effects model predicting reflective depth with random intercepts for participants. Our third model examined whether the effect of emotion on reflective depth depends on users' inherent engagement profiles. We extended Model 2 by including interaction terms between emotion type and cluster membership. This interaction model tested whether certain emotional states promote deeper reflection only among specific user types, or whether emotional effects are consistent across engagement profiles. The first model revealed a **significant** effect of Emotion Type on reflection propensity, indicating that participants' likelihood of engaging in reflection was primarily driven by their emotional context. In particular, reporting low arousal negative emotions (e.g., sadness, loneliness) was associated with a significantly higher likelihood of reflection ($\beta = 0.72$, $p = .006$). This corresponds to more than a twofold increase in the odds of engaging in reflection ($OR \approx 2.05$). Reporting high-arousal negative emotions (e.g., anxiety, anger) also showed a marginally positive association with reflection ($\beta = 0.53$, $p = .067$), suggesting a similar but weaker trend. In contrast, positive emotions did not significantly predict reflection, indicating that positive emotional states were unlikely to trigger reflective engagement. Neither Cluster_ID nor Emotion Engagement Score significantly predicted reflection propensity. This suggests that participants' engagement profiles and the intensity or breadth of their emotional reporting were not reliable predictors of reflective behavior. The second model also revealed **significant** effects of negative emotional states on reflective depth. Reporting low-arousal negative emotions was associated with substantially deeper reflection ($\beta = 0.34$, $p = .008$). Similarly, reporting high-arousal negative emotions was also linked to increased reflective depth ($\beta = 0.29$, $p = .045$). These findings indicate that participants not only initiated reflection more frequently when experiencing negative emotions, but also engaged more intensively with reflective features in these contexts. In contrast, reporting low-arousal positive emotions did not show a significant effect, suggesting that positive emotional states were not associated with deeper

reflective engagement. Consistent with the first model, neither Cluster_ID nor Emotion Engagement Score significantly predicted reflective depth. The random intercept variance ($\sigma^2 = 0.21$) indicates the presence of stable individual differences in reflective depth across participants. However, these differences were not explained by the engagement clusters, suggesting that the observed between-user variability reflects factors beyond the behavioral profiles captured by the clustering procedure. Finally, the interaction model revealed no robust moderation effects of engagement profile on the relationship between emotion type and reflective depth. While low-energy negative emotions remained a significant predictor of deeper reflection and high-energy negative emotions showed a marginal effect, none of the Emotion Type × Cluster interaction terms reached statistical significance. These findings reinforce our earlier results, indicating that the influence of emotional context on reflective depth is consistent across engagement profiles. Together, these findings suggest that reflective features in the app function primarily as situational coping tools. Rather than being used preferentially by specific types of users, they are activated in response to momentary emotional needs, especially during negative emotional states. This highlights the importance of designing adaptive, emotion-aware interventions that support users when they are most receptive to reflection.

7.6 Phase 3: Understanding User Perceived Meaningful Emotion Self-Tracking

Phase 2 demonstrated that participants derived value from the application, particularly increased emotional awareness. However, engagement patterns over time revealed important limitations in how that value translated into sustained use. Despite being informed by users' stated preferences, the application did not fully support the engagement aspects that participants ultimately found most meaningful, such as habit formation, deeper emotional understanding, and integration into daily life. These findings suggest that elicited preferences and observable interaction patterns alone were insufficient to capture the full spectrum of user-perceived value. Motivated by this gap, we designed Phase 3 of this study, aimed

to more deeply examine how users themselves define value, usefulness, and meaningful engagement in the context of long-term emotion tracking. Specifically, our goal in this phase is to uncover what user-valued outcomes and expectations were not fully surfaced in Phase 1 or reflected in the application design for Phase 2, and to understand how these perceptions shaped the differences between sustained or selective engagement and early disengagement. In Phase 3, we aim to answer RQ3a and RQ3b by focusing on users' lived experiences and reflections. Next, we will discuss our study design, participant recruitment, analysis, and findings for phase 3.

7.6.1 Participants Recruitment

For Phase 3, we recruited participants from the Phase 2 field study cohort. We distributed a Google Form–based survey to all field study participants, with an introductory section that described the study's purpose and outlined data confidentiality. The survey began with an informed consent question, followed by open-ended questions focused on our themes. The final survey question invited participants to opt in to a follow-up interview and to provide contact information. Ten participants expressed interest in interviews and were subsequently contacted. Semi-structured interviews were conducted over Zoom based on participants' availability. Both the survey and the interview were voluntary participation with no paid incentives.

7.6.2 Study Design and Analysis

To study RQ3a and RQ3b, we conducted a qualitative analysis in our phase 3. We engaged participants through a survey (n = 23) and semi-structured interviews (n = 10) to understand the value users derived from emotion self-tracking and the factors influencing their engagement. We designed our survey to capture participants' retrospective reflections on their engagement with the app and to surface user-perceived values that were not fully evident from interaction logs or earlier preference elicitation. The survey focused on three thematic

areas: the perceived value and usefulness of emotion journaling; factors that hindered participants from engaging at their desired level; and reflections on unmet needs and potential changes that could support more meaningful, self-sustaining engagement. By prompting participants to reflect on both their experiences and expectations, our qualitative survey and interview enabled us to examine how users across different engagement styles evaluated the app's impact and articulated what they value in a long-term emotion tracking tool. All survey questions were open-ended and are included in Appendix E.1. We followed the survey with semi-structured interviews to further explore participants' perceived values, desired forms of engagement, and expected outcomes from emotion tracking. The interview protocol is also provided in Appendix E.1. The first two authors conducted the interviews online after obtaining informed consent to record the sessions, and audio recordings were transcribed using Zoom Pro. Both authors then cross-checked the transcripts and conducted a thematic analysis of the combined survey and interview data. Codes that emerged from the analysis included *"forming a habit to pause and reflect," "devising strategies for managing emotions",* and *"lack of visible outcomes,"* which informed our understanding of the value participants derived from the emotion self-tracking. The results of this analysis are presented in the following section.

7.7 Phase 3: Findings

In this section, we present findings from the Phase 3 qualitative investigations. We organize the findings into two broader sections. First, we describe the aspects of emotion logging that participants found meaningful and valuable. Second, we examine the factors contributing to participants' disengagement from human-centric emotion self-tracking.

7.7.1 Participant Perceived Value and Meaningful Engagement

Overall, analyses from Phases 2 and 3 indicated that participants did find some value in self-tracking their emotions, as reflected in their feedback and interviews. This subsection

explores these perceived values and the factors that motivated their engagement with the emotion-logging application.

Development of Reflective Habit

Participants consistently described emotion logging as valuable for increasing self-awareness. Pausing to identify and label emotions helped them step back from their immediate experiences and reflect more clearly on how they were feeling. Several participants noted that this pause enabled a more objective perspective on their emotions, rather than remaining immersed in them. Over time, this reflective pause, although inconsistent, became internalized for many participants, even when frequent in-app logging was not sustained. As one participant explained, *"When it gives a pop-up notification, I have to pause and think, like, what I'm feeling right now. So that was a really good feeling."* These findings suggest that the primary value of emotion logging lies in cultivating reflective self-awareness, rather than in continuous or frequent use of the application.

Recognizing Patterns and Mechanism

Participants described how emotion logging over time, regardless of frequency, enabled them to identify personal emotional patterns. They noted that the app's input interfaces helped structure existing journaling or reflective practices, while its feedback features supported the development of individualized emotion-management strategies. Importantly, participants emphasized that the app's value extended beyond in-app use. Rather than relying on the app as a continuous tool, they translated insights gained from logging into offline coping mechanisms, such as seeking social support, engaging in creative outlets, or participating in their favorite activities.

Private Emotional Expression in the Absence of Social Support

Finally, participants highlighted the app's value as a space for expressing emotions when social connections were unavailable. For some, the emotion-logging function served as a substitute for discussing feelings with others, particularly during periods when social support was limited or difficult to access. In these cases, the app served as a private, nonjudgmental, and easy-to-use outlet for emotional expression, reinforcing its perceived usefulness during moments of isolation.

Support for Varied Emotional Expression and Self-Learning

Participants valued flexibility in how emotions could be expressed, noting that a one-size-fits-all approach would not support their needs. Participants liked exploring different input methods, as they helped them identify the method that worked best for them in naming their emotions more effectively. Participants also like having access to log history, visualizations, and mental health resources. As P9 shared: *"Another one that I like is the resource section. If I'm looking for articles, I can just hop onto that."* Overall, for some participants, the app facilitated self-learning about emotions.

Engagement Motivated by Emotional Need

Participants reported that their willingness to engage with emotion logging was strongly influenced by the intensity of their emotional experiences. Logging was intrinsically motivating during moments of extreme emotions, such as feeling *"extremely happy, extremely sad, [or] extremely angry,"* when participants felt a strong need to identify and process their feelings. In contrast, during emotionally neutral periods, reminders to log were often perceived as intrusive. As P3 explained, notifications felt like *"noise"* when *"there's nothing disturbing happening in my life,"* and engagement was driven more by obligation than personal motivation. They suggested that the emotional relevance of notifications is crucial for engagement. Frequent notifications during emotionally neutral periods were

often intrusive, whereas prompts aligned with moments of high emotional intensity could motivate logging. Additionally, participants appreciated the flexibility to log emotions at times of their own choosing, independent of notifications, allowing engagement to align with personal routines. Overall, this suggests that for most participants, emotion logging was primarily driven by emotional intensity rather than routine. Participants tended to reflect on their emotions only during moments of crisis, so daily logging represented a departure from their typical emotional management practices.

7.7.2 Participants' Perspectives on Disengagement Factors

Our analyses from Phases 2 and 3 highlighted several factors that influence sustained engagement with emotion self-tracking in daily life. In this section, we present our findings on the factors that contributed to disengagement.

Desire for Quick, Visible, and Quantifiable Outcomes

In our qualitative data, participants frequently compared emotion-tracking apps with more engaging technologies, such as social media and fitness tracking apps. They noted that, unlike social media, which provides immediate rewards like *"infotainment"* or a *"dopamine hit"*, and fitness tracking apps, which offer visible, quantifiable goals such as step counts or weightlifting targets. In contrast, emotion tracking apps were seen as offering long-term, internal benefits that were difficult to measure. As P7 noted: *"With emotion annotation apps, I don't think there is that kind of quantifiable goal... while working out, I have a measurable target, but just logging emotions doesn't give that."* Participants also observed that features like emotion history or log visualizations depend on consistent use over time, and it can take weeks of logging before meaningful patterns emerge. During the initial period, there is little external motivation, and engagement relies on participants' intrinsic will, which can be disrupted by competing demands or other apps. Overall, participants suggested that emotion-tracking apps could improve engagement by providing quicker,

visible, and actionable feedback. A participant (P10) clearly articulated what they meant by actionable feedback: *"The goal of using any emotion well-being app is to become happier or more emotionally stable over time by developing better regulation mechanisms. While the app can help me build a habit to pause and reflect, if I am not able to see changes in my happiness or mood, it doesn't make any sense, and I might withdraw."* This highlights that participants not only value reflection but also need visible, meaningful outcomes that demonstrate progress and change in emotional well-being to sustain engagement.

Need for Adaptive and Low-Effort Engagement

For most participants in our cohort, emotion logging was a new habit, something they were aware of but had not traditionally practiced as part of their emotional well-being management. Alternative strategies, such as talking to friends and family about emotions, engaging in physical exercise, or using distractions like scrolling social media or binge eating, felt much more natural and familiar. These methods had helped participants in the past and required minimal cognitive effort. In contrast, logging emotions without visible outcomes was seen as a competing mechanism: while valuable in theory, it demanded deliberate reflection, consistent attention, and cognitive effort that participants found difficult to sustain on their own. As a result, participants expressed a preference for tools that require low effort and can be seamlessly integrated into their daily routines. They highlighted the need for adaptive systems that learn from users' routines and emotional needs, supporting dynamic logging while respecting their existing coping strategies. Busy schedules and competing attention from other apps, such as social media or email, further limited consistent engagement. The cognitive effort required to reflect, navigate multiple screens, and structure entries added additional friction, reducing sustained use. Participants suggested that emotion-logging mechanisms should minimize effort and offer flexible engagement options. As explained by P5, *"for example, if a user is mostly neutral, the app could present a simple prompt, such as asking whether they feel the need to log, or allow them to defer, rather than requiring full*

logging." suggesting such a dynamic approach would support more meaningful engagement.

Need for More Situational Post-Logging Support

Participants noted that while the app's post-logging processing features were practical, they were not equally effective for all users or in all situations. Some reported that reflecting on negative emotions without adequate support could reinforce a spiral of negativity, leading them to hesitate using the app during high-intensity emotional states. They emphasized that the usefulness of these features depends on individual preferences, current coping mechanisms, emotional regulation abilities, and the intensity of the emotions being experienced. To address this, participants suggested that post-logging support should be situational and adaptive, offering actionable support that responds to the user's emotional state. For example, a participant (P9) suggested *"after logging a stressful moment, the app could suggest a walk, a breathing exercise, or another contextually relevant activity, and provide subtle support such as step counts or timed exercises."* Participants emphasized that these suggestions should be persuasive but optional, allowing users to defer if they prefer to rely on their own coping strategies. They also noted that any in-app support would naturally compete with existing mechanisms, such as talking to friends, exercising, or using social media, so guidance should be presented in an informative and encouraging way, rather than being prescriptive, perhaps using small facts, anecdotes, or examples of potential benefits.

7.8 Discussion

This work began in Phase 1 with the design of a human-centric emotion self-tracking system informed by users' needs for autonomy, flexibility, and meaningful interfaces and feedback. In Phase 2, we deployed the system in the field to examine how users engaged with the application over time. Finally, in Phase 3, we explored in depth the values users derived from emotion self-tracking and the frictions they encountered when engaging with such systems. In this section, we will discuss how these findings (i) align with or challenge

data-driven narratives of engagement and (ii) outline design implications for value-driven human-centric emotion self-tracking systems.

7.8.1 Where Users Find Value in Emotion Self-Tracking?

Prior work on engagement with digital mental health interventions and emotion self-tracking systems has identified several factors such as fluctuating user motivation, notification fatigue, limited attention, static designs, lack of user control, time constraints, and evolving user needs [96] as key contributors to low engagement, often reflecting a misalignment between participants' needs and system design [327, 328, 309, 101, 101, 329]. Prior research has also shown that life circumstances and situational disruptors can further contribute to disengagement [315]. Our work extends this body of research by showing that, amid changing life circumstances and the need for adaptive, low-effort designs, user engagement is also shaped by the *perceived payback* that emotion self-tracking applications provide. In particular, participants emphasized the importance of ***quantifiable or visible outcomes*** related to emotional well-being. We observed that such outcomes are relevant across all stages of engagement, including the initial novelty phase, especially for users for whom the barrier to engagement is high from the outset, but is also necessary for users across all engagement profiles. Importantly, these outcomes were seen as necessary alongside, rather than in place of, ease of interaction and flexible design. Our findings also suggested that the *implicit value* of emotion well-being systems lies not only in supporting data collection, reflection, or pattern visualization, but also in enabling the ***formation of personalized emotion regulation mechanisms*** for users that they can carry into their everyday life in pursuit of greater emotional stability and well-being. Participants described this pursuit of *"feeling good"* [316] as the primary goal for using any emotion logging or regulation application. Consistent with this, we observed that one of the main drivers of both objective and subjective engagement [330] in our field study was the experience of intense negative emotions across all engagement profiles. Emotions such as anxiety, stress, and anger most

strongly motivated users to engage, as participants perceived these states to have the greatest impact on their health, relationships, and productivity. Similarly, engagement with emotion regulation support was predominantly driven by negative emotional contexts, particularly low-arousal states such as sadness or loneliness. This suggests that users primarily turn to emotion logging and seek technological support during moments of emotional difficulty, rather than during neutral or positive states.

Importantly, these needs persisted even when users expressed satisfaction with the app's flexibility, limited interruptibility, and autonomy-supportive design. Similarly, other human-centric features, such as customizable notification schedules, access to log histories, pattern visualizations, learning resources, crisis support, and choice-driven interactions for logging emotions or post-reflection with appreciation messages [2] were generally well received, but they did not, on their own, lead to increased engagement. In some cases, these features added to users' perceived burden, particularly given the emotional labor involved in reflecting on difficult emotions [268]. Moreover, when users lacked timely or situational support to process intense emotional states, reflection itself was experienced as effortful or even aversive. From a human-centric design perspective, this highlights that supporting autonomy, flexibility, and reflection is necessary [318] but not sufficient. Therefore, human-centric emotion self-tracking systems must account for users' emotional capacity during moments of distress and be designed to provide actionable support or timely assistance during emotionally vulnerable periods. Offering scaffolding that reduces reflective burden is essential, particularly when users experience isolation and treat these systems as readily accessible sources of support in the absence of social connections. Such design considerations are critical for building user trust and ensuring that emotion self-tracking tools are perceived as supportive rather than demanding during times of need. Moreover, our findings highlight the importance of low-effort, adaptive designs that support learning and gradually reduce reliance on the technology itself. These themes point toward designing for well-being–oriented engagement rather than continuous interaction or

frequent logging, aligning engagement with users' longer-term emotional goals rather than system-centric usage metrics [330]. Consistent with prior literature, we also observed habit formation as a major challenge for emotion self-tracking systems, as they often compete with users' existing coping and emotion regulation mechanisms. Many of these mechanisms follow lower-effort or more immediately rewarding pathways, such as scrolling on social media, seeking quick dopamine rewards, binge watching, or, in positive distraction real-life mechanisms, like talking to family or friends, going for a walk, or engaging in creative outlets like sketching [316, 315, 331]. Overall, our findings alongside prior literature on disengagement factors and user values [316, 315, 96], highlighted that engagement is beyond the data-driven metrics of compliance, logging frequency, responsiveness to prompts, and sustained interaction over time. Furthermore, participants valued real-life, tangible improvements in emotional well-being and support for actionably achieving these outcomes over data-driven insights or self-awareness alone. This aligns with prior critiques of reflection-centric personal informatics systems [332] and challenges the *self-improvement hypothesis* [333], suggesting that awareness alone, without actionable pathways, may be insufficient for sustained engagement or meaningful change. Next, we will discuss the design implications for self-tracking systems that go beyond user-centric design to support users' evolving values and situational context.

7.8.2 How can we Design for User Value?

As discussed in the prior section, our findings indicate that participants valued gradual progress over time supported by actionable pathways or quantifiable outcomes, which they perceived as essential for achieving their overarching goal of improved happiness and emotional stability. A central implication of these findings is the need to design for actionable, user-centered goals that translate emotional awareness into concrete steps toward emotional well-being. Rather than framing goals solely around logging frequency or reflection depth, systems can help users define personally meaningful outcomes, such

as feeling calmer in stressful situations or responding more constructively to recurring emotional triggers. These goals can be operationalized through lightweight, task-oriented prompts following emotion logging, for example, by suggesting short, contextually relevant actions like taking a brief walk, practicing a breathing exercise, or reaching out to a trusted person [332]. Importantly, progress toward user-centered goals should be made visible through gentle, qualitative, or semi-quantitative indicators, such as noticing reduced intensity of recurring emotions or increased use of adaptive coping strategies, rather than rigid numerical targets. Systems can support these goals through lightweight monitoring and subtle persuasive mechanisms, for example, acknowledging completion, reflecting on small mood changes, or offering gentle reinforcement through questions, rather than relying solely on repeated introspection or frequent self-reporting. Additionally, systems should allow goals to evolve over time, enabling users to revise them as their emotional needs and priorities change. It is also important that these persuasive systems reflect users' emotional context or needs and remain respectful of their situational context [258]. Another key design implication from our findings is the need for adaptive, low-effort designs that remain flexible to all user profiles. Emotion self-tracking systems should dynamically adjust to users' routines, emotional states, and engagement patterns over time, providing interactions that are contextually relevant and minimally burdensome. For example, notifications or prompts should remain optional or deferrable during neutral or low-arousal periods, and input options should adjust based on users' past behavior. Flexibility in interaction is equally important. Users should be able to engage on their own terms, choosing when and how to log, selecting input methods that fit their cognitive load at the moment, or deferring interactions without penalty [2, 96, 315]. Low-effort mechanisms, such as simplified logging interfaces, pre-filled or guided reflection options, and one-tap acknowledgment of suggested post-logging actions, can reduce the emotional and cognitive burden that often accompanies reflective work on emotions. Furthermore, self-tracking systems should evolve alongside users over time, learning from their engagement patterns and emotional responses

to provide increasingly personalized support [258]. This could include adjusting suggested actions, offering timely reinforcement or gentle reminders, and presenting feedback that highlights progress toward meaningful goals, all without demanding constant effort or attention. By embedding flexibility and adaptive support, emotion self-tracking systems can better accommodate episodic, emotion-driven engagement, complement existing coping strategies, and support sustainable, long-term emotional well-being. Finally, our findings underscore that users will always rely on real-life support systems, such as friends, family, or other coping mechanisms, and technology should complement rather than replace these supports [316]. Emotion self-tracking systems should avoid presenting users' existing strategies as unnecessary or inferior, instead acknowledging and reinforcing users' own approaches [332]. At the same time, systems should scaffold learning and self-reflection so that, over time, users can develop their own effective mechanisms for understanding and regulating emotions, gradually reducing reliance on the technology. This balance ensures that the system supports growth and autonomy, empowering users to achieve emotional well-being both within and beyond the app.

7.9 Study Context and Limitations

All phases of this study were conducted in India and involved participants of Indian descent. Our sample primarily comprised urban, educated, and digitally literate individuals. Despite this level of digital familiarity, emotional literacy and awareness remain relatively nascent within the Indian socio-cultural context. Prior work has highlighted persistent stigma surrounding the expression, recognition, and discussion of emotions, as well as barriers to help-seeking and limited mental health awareness more broadly [334, 335]. In parallel, the use of digital applications for emotional well-being is still emerging and remains largely unadapted to Indian cultural norms and practices; as a result, such technologies are not yet integrated into everyday life [336]. These contextual characteristics were consistently evident across participants and study phases. Participation in all study phases was entirely

voluntary, with no monetary or technological compensation provided. Participants engaged out of intrinsic motivation and an expressed interest in emotional self-reflection and logging. This self-selected and motivated sample constitutes an important limitation, as our findings reflect the experiences of individuals already open to engaging with emotional well-being technologies. Accordingly, our findings should be interpreted within this specific cultural, demographic, and motivational context. Our findings and research questions related to engagement are most directly applicable to urban, digitally literate Indian populations exploring emotional well-being tools. While we believe that several of our insights may extend beyond this setting, given the fundamental nature of the engagement behaviors and design considerations examined, further research across diverse cultural and socio-economic contexts is necessary to assess the broader generalizability of these findings.

7.10 Summary

In this chapter, I investigated how users engage with a human-centric, user-informed emotion self-tracking application over time and what they perceive as meaningful value and potential frictions in such systems. Across three phases, the study revealed that engagement is largely emotion-driven, shaped by internal motivation, the intensity of emotional experiences, and prior emotion-management habits, rather than by notifications, prompts, or aesthetic design alone. Users valued reflection, pattern recognition, and support for emotion regulation, but engagement was constrained by cognitive effort and a lack of goal-oriented outcomes. Overall, findings highlighted the importance of designing emotion self-tracking systems that support actionable, low-effort, adaptive, and flexible interactions, while providing gentle, goal-oriented feedback to make progress visible. Systems should complement rather than replace users' existing coping mechanisms and gradually scaffold users toward developing their own strategies for emotional well-being. By emphasizing users' value rather than mere system-driven usage metrics, this chapter offers guidance for designing future emotion self-tracking systems that align with users' long-term well-being goals.

Chapter 8

Towards Humanistic Data-driven Interventions for Emotional Well-being

"The oak fought the wind and was broken, the willow bent when it must and survived."

— Robert Jordan

Across the preceding chapters of this thesis, I have shown that the potential of physiological signal-based emotion recognition is fundamentally constrained by the limited availability of large-scale datasets and the difficulty of harmonizing the many small, heterogeneous datasets that exist across the field. In Chapter 3, I demonstrated how this heterogeneity can obscure meaningful progress, making it difficult to determine which modeling paradigms are truly effective, which benchmarks provide reliable indicators of performance, and how models should be evaluated under conditions that reflect real-world variability. In Chapter 4, I extended this discussion by examining how more informative datasets can be constructed through the integration of subjective emotional self-reports with standard objective physiological measures, enabling richer and more nuanced representations of affect. Chapter 5 then shifted the focus from modeling to participants themselves, showing how individuals interpret and translate their emotional experiences into annotation scales. These findings highlighted a critical limitation: commonly used labeling schemes often constrain emotional expression and fail to capture individual differences, temporal variability, and the influence of personal history. Building on this, Chapter 6 introduced an emotion annotation tool designed to better accommodate subjectivity, variability, and participant needs. My short-term field study showed that while increased flexibility and user control can enhance perceived agency, these factors alone are insufficient to sustain long-term engagement or ensure consistent and high-quality data collection. Chapter 7 further investigated this issue by examining users' experiences and motivations in emotional well-being technologies,

focusing on why individuals engage with or disengage from emotion journaling and data collection tasks, and what intrinsic and contextual factors shape sustained participation.

Taken together, the findings of this thesis converge on a central conclusion: emotion data design is not a peripheral concern, but a primary determinant of downstream modeling performance and system validity. Across all chapters, it becomes evident that decisions made during data collection, ranging from elicitation protocols and labeling strategies to participant interaction design, introduce structural constraints that directly shape what models can learn and how well they generalize. In other words, limitations in emotion recognition are not solely modeling failures, but are deeply rooted in how emotional data is elicited, interpreted, and labeled in the first place. This has broader implications for the field. While advances in modeling techniques remain important, they cannot compensate for misaligned or overly simplified representations of emotional experience. The inherently subjective, context-dependent, and interpretable nature of physiological signals means that emotion recognition systems are especially sensitive to design choices made at the data level. Future progress, therefore, requires treating dataset design, annotation practices, and participant experience as first-order research problems, rather than auxiliary considerations. More broadly, this thesis argues for a shift in emphasis from model-centric development toward data-centric and participant-centered approaches. Such a shift entails recognizing that emotional signals do not encode fixed labels, but are mediated through individual interpretation, situational context, and expressive variability. Accounting for this complexity in data is essential not only to improve model performance but also to ensure that emotion recognition systems remain meaningful, interpretable, and applicable in real-world settings. Furthermore, this thesis also highlights what participants seek from emotion monitoring systems and how future system design should explicitly accommodate these needs and individual differences. These needs extend beyond straightforward representations of emotional trajectories; more critically, participants value the subtle forms of support these systems can provide, such as behavioral nudges, reflective learning, and context-sensitive guidance, that could translate

sensed data into meaningful real-world change. Collectively, these insights suggest that the future of physiological emotion recognition depends less on isolated improvements in algorithmic performance and more on the co-design of datasets, annotation frameworks, and modeling approaches that reflect the lived, subjective nature of emotional experience.

Next, the discussion focuses on three interrelated themes: the tension between idiographic and nomothetic approaches to emotion modeling; the role of socio-cultural and individual differences in shaping emotion understanding and technology use; and the implications of these findings for the design of participant-aware, emotionally meaningful systems, and outlines directions for future research.

8.1 From Nomothetic to Idiographic Emotion Modeling and Intervention

Across the empirical studies presented in this thesis, a central finding emerges that emotion label semantics are not fixed or universally consistent, but instead vary systematically across experimental protocols, annotation strategies, and individual interpretation. In Chapter 3, I demonstrated that this variability has direct consequences for model performance. Benchmarking results revealed strong dataset dependence in arousal and valence prediction, with models performing well on some datasets but failing to generalize to others. In particular, cross-dataset evaluations showed limited transfer between self-report–based datasets and those constructed using stimulus-driven annotation schemes. Both qualitative and quantitative analyses indicate that these discrepancies are driven by differences in labeling strategy and by signal quality or model capacity. In Chapter 4, I further found that incorporating qualitative textual labels during training substantially improves representation learning. Models trained with text supervision not only achieved higher performance on in-domain tasks but also demonstrated strong zero-shot transferability across multiple external datasets. This suggests that text-based labels can provide a more contextually informative supervisory signal than traditional scale-based annotations. Moreover, the same training paradigm enabled representations that generalize effectively across heterogeneous emotion

recognition datasets, highlighting the importance of idiographic labels over nomothetic scales. Chapter 5 extends these findings by examining the human factors underlying annotation variability. My results show that differences in participants' self-perception, experimental framing, prior knowledge, and lived experience significantly shape how internal emotional states are mapped onto numeric or categorical scales. These factors challenge the assumption of a stable or linear correspondence between emotional experience and numerical labels. Instead, emotion annotation emerges as a constructed, context-dependent interpretive process rather than a direct measurement procedure. In real-world settings, I further observed that similar life situations can elicit divergent emotional responses and labels across individuals, shaped by differences in psychosocial profiles, emotional history, reactivity, expressive vocabulary, and situational context. Taken together, these findings suggest that the central challenge in physiological emotion recognition lies not only in modeling the physiological signal itself, but also in designing annotation frameworks that explicitly represent variability in emotional experience, interpretation, and expression across participant profiles and contexts. Chapter 6 builds on this direction by exploring annotation tools that allow participants to express emotions in line with their situational context, availability, and emotion complexity, while supporting the representation of complex and mixed emotional states. Finally, Chapter 7 examines how individual differences not only influence model performance but also play a critical role in users' engagement with emotion-aware well-being technologies. These observations on user behavior suggest the value of hybrid idiographic–nomothetic approaches, commonly used in psychology to balance generalizable structure with person-specific adaptation. Such approaches combine shared models with individualized interpretation, enabling systems that remain broadly applicable while still accommodating the specificity of individual emotional profiles and contexts. Overall, this perspective stands in contrast to predominantly nomothetic approaches commonly used in prior research in HCI and ubiquitous computing, which assume a stable, universal mapping between physiological signals and emotion labels. In summary, this thesis

provides preliminary evidence that advancing emotion recognition systems requires moving beyond strictly nomothetic assumptions if the field is to better support the monitoring and understanding of individuals' lived emotional experiences. Furthermore, it motivates a shift toward idiographic-aware, annotation and modeling frameworks that treat emotional variability not as noise but as a structurally informative property of emotional data.

8.2 Emotion Literacy and Socio-Cultural Context of Emotion Data

Another important theme that emerged across this dissertation is the role of emotional literacy and socio-cultural factors in both emotion data collection (Chapters 5 and 6) and the design of technology-supported interventions for emotional well-being (Chapter 7). My research was situated primarily in the Indian context, where emotional expression, awareness, and regulation are shaped by complex sociocultural norms. Understanding these dynamics is critical because emotion recognition systems do not operate in isolation; their effectiveness depends heavily on how individuals perceive, articulate, and manage their own emotional experiences. While these observations arise from a specific regional context, they reflect broader global challenges in emotion-aware technology design. Emotional norms, stigma around mental health, access to psychological resources, and culturally shaped coping practices vary widely across societies. Consequently, designing emotion monitoring and well-being technologies requires sensitivity not only to local contexts but also to cross-cultural variability, ensuring that systems remain adaptable, inclusive, and generalizable across diverse populations rather than optimized for a single cultural setting. In India, emotional expression is often influenced by persistent social stigma. Openly discussing emotions, particularly vulnerability, distress, or mental health struggles, may be perceived as a sign of weakness or lack of resilience [2, 1]. This perception can discourage candid self-reporting and affect both the quality of emotion datasets and the adoption of digital emotional well-being tools. Participants in my studies (chapters 6 and 7) frequently expressed hesitation in labeling or sharing emotions, especially negative ones, due to

concerns about privacy, judgment, or social expectations. This suggests that technological systems that rely heavily on explicit self-reporting must be designed with sensitivity to these social constraints, or they can lead to lower engagement and missing data points.

At the same time, *emotional literacy*, the ability to identify, interpret, and regulate emotions, remains unevenly distributed within my study samples. Formal education rarely includes structured emotional skills training, and awareness of emotional regulation strategies varies widely across demographic groups. While mental health awareness is increasing, access to professional psychological support is still largely concentrated in urban and semi-urban areas. Consequently, many individuals continue to rely primarily on informal social networks, including family members, friends, and peer groups, for emotional support. Strong interpersonal networks can provide substantial emotional buffering, yet these support structures may not always facilitate explicit emotional articulation or evidence-based coping strategies. India also has a long-standing cultural tradition of practices associated with emotional regulation, such as meditation, yoga, breathwork, and mindfulness-oriented spiritual activities. These practices are widely recognized and culturally accepted, creating opportunities for interventions that align with existing behavioral norms rather than introducing entirely new frameworks. However, engagement with such practices tends to be inconsistent and often lacks structured guidance or sustained motivation. This inconsistency highlights a potential role for digital systems to provide adaptive support while respecting cultural familiarity.

Taken together, these observations about my participant samples (from chapter 5, 6 and 7) and their engagement in data collection studies suggest several implications for future research and design. Emotion recognition systems must account for varying levels of emotional literacy and potential reluctance to disclose emotions explicitly. Interventions should build on culturally familiar coping mechanisms while incorporating structured guidance to sustain engagement. Participant-centered data collection approaches should also incorporate agency and contextual and cultural awareness in design to better capture how

emotions are experienced and expressed within specific cultural environments. Ultimately, improving emotional literacy is not only a societal objective but also a methodological necessity for advancing physiological emotion recognition. Greater emotional awareness and articulation can enhance data quality, improve annotation reliability, and support the development of more meaningful, culturally sensitive emotional well-being technologies that remain relevant across diverse global contexts.

8.3 Awareness Is Not Enough: Rethinking Digital Mental Health Interventions

Throughout this dissertation, another recurring theme emerged: understanding the factors that shape motivation and sustained engagement with digital mental well-being technologies. Most existing mental well-being tools, including journaling applications, wearable-based stress monitoring systems, meditation platforms, and, more recently, large language model–driven conversational agents, are primarily designed to increase users' awareness of their emotional patterns. These systems often assume that once individuals become aware of their emotional states, they can initiate appropriate coping strategies or seek relevant guidance [337]. However, my findings in Chapters 6 and 7 suggest that this assumption does not consistently hold in practice. Across my studies, participants with diverse backgrounds ranging from individuals managing everyday stress and minor inconveniences to those living with diagnosed mental health conditions, a recurring theme emerged that *awareness alone is not enough*. Many participants in my studies reported that they already possess a reasonable understanding of what they are feeling, but often choose to ignore those feelings when they lack accessible mechanisms to address them. In such situations, people frequently rely on pre-existing coping practices. Some of these are adaptive, such as physical exercise, conversations with family or friends, or structured relaxation practices. Others, however, may be avoidance-oriented, such as excessive social media use or playing video games to distract from uncomfortable emotions.

Participants and domain experts in Chapter 5 also expressed that current technologies

tend to prioritize accurate detection, classification, or visualization of emotional states, sometimes striving for increasingly precise emotion recognition models. While technically valuable, this emphasis can overlook what users actually seek: ongoing, actionable support that helps them respond constructively to emotional challenges. Several participants in Chapter 7 also compared their needs to therapeutic processes in which guidance is gradual, contextual, and focused on skill-building rather than solely on identifying emotional states. They emphasized a desire for systems that help them develop personalized coping repertoires, which some described as a *"mental health toolkit"*[338] or an emotional resilience portfolio that evolves over time and adapts to life circumstances. These observations suggest that future mental well-being technologies should extend beyond awareness-focused approaches toward systems that actively foster emotional self-regulation skills and support the gradual development of emotional resilience. This includes providing actionable yet low-burden guidance or tasks, supporting incremental habit formation, and aligning interventions with users' existing coping ecosystems rather than attempting to replace them. Designing for sustained engagement will therefore require not only accurate sensing and modeling but also thoughtful integration of behavioral support, personalization, and long-term user motivation.

8.4 From Quantified Use to Emotional Relevance: Evaluating Mental Health Technologies

When assessing the usability and effectiveness of the mental well-being interventions I designed in chapter 7 of this dissertation, I observed that prior research often prioritizes engagement metrics such as frequency of use, session duration, or interaction counts as indicators of system success [82, 311]. While these quantitative measures are convenient and scalable, they do not always reflect the realities of how individuals engage with tools for emotional well-being. Across my studies, participants consistently emphasized that engagement with emotion-tracking technologies is driven less by habitual use and more by emotional need, situational relevance, and perceived emotional salience, as reflected in my

findings from chapters 5, 6, and 7. In many cases, individuals intentionally disengage once they feel better, suggesting that reduced usage may actually signal improved well-being rather than system failure.

My research also highlights an important tension: mobile devices, which host most digital well-being tools, are themselves significant sources of distraction, cognitive load, and sometimes emotional strain due to the presence of several attention-seeking applications, such as emails, messages, social media, and entertainment platforms. As a result, participants often preferred low-burden interactions that did not impose constant monitoring, excessive notifications, or performance pressure. Systems that repeatedly prompt users to share emotional data or display a history of negative moods may inadvertently reinforce negative moods rather than support the inherent goal of stable and positive moods. Furthermore, many participants reported approaching these technologies with clear, outcome-oriented goals, such as feeling calmer, improving mood stability, resolving stressful situations, or learning coping strategies, rather than simply tracking emotional states. Traditional quantified-self frameworks [337], which emphasize continuous logging and data accumulation, may overlook these experiential and outcome-focused dimensions of use. Future research should therefore adopt broader evaluation criteria that consider emotional benefits, perceived support, development of coping skills, and long-term well-being alongside conventional engagement metrics. Such an approach would better align system assessment with the actual goals users bring to mental well-being technologies.

8.5 Participant-Aware Foundation Models

Findings from Chapters 3 and 4 indicate that pretrained models trained on datasets with richer text-based labels can improve emotion recognition performance. However, their effectiveness is contingent on access to large-scale and diverse datasets, which are often unavailable in practice due to the heterogeneity of existing resources, as discussed in Chapter 3. This variability also complicates dataset integration and limits straightforward model gen-

eralization across studies. From a technical perspective, the future of physiological emotion recognition likely lies in foundation models trained on heterogeneous datasets that integrate physiological signals with contextual, behavioral, and psychosocial information. In this work, the development of participant-aware datasets and CLSP-based models demonstrates the promise of this direction, particularly when data explicitly captures users' lived context and subjective experience. Pretraining on diverse, large-scale resources enables models to learn more robust and transferable representations of emotion that generalize across settings, devices, and labeling protocols. Looking ahead, the development of interoperable data ecosystems will be essential. This includes standardized reporting practices, shared benchmarks, and ethically grounded data governance frameworks that together enable more effective use of existing datasets while ensuring reproducibility, fairness, and participant privacy. Collectively, these efforts point toward a future in which foundation-model approaches support practical, context-aware, and ethically responsible emotion recognition systems that remain grounded in users' lived experiences. Finally, I also believe that integrating structured morphological representations and their textual descriptions can further enhance large-scale model pretraining by providing additional semantic grounding for physiological and emotional patterns, particularly in the absence of high-quality labels. Since annotation remains a key bottleneck due to its reliance on human input, there is a need for further work on hybrid training approaches that combine self-supervised and supervised learning. Such approaches could reduce reliance on extensive participant labeling while still retaining the benefits of task-specific supervision and incorporating morphological structural knowledge where available.

8.6 Final Reflections: Towards Human-Centered Data-Driven Emotion Recognition

Taken together, the findings from this dissertation underscore that advancing physiological emotion recognition and mental well-being technologies requires a holistic, human-centered approach. Beyond improving predictive performance, it is essential to design systems

that respect users' emotional needs, provide actionable guidance, and integrate seamlessly into their daily lives without adding cognitive or emotional burden. By coupling such technical advances with thoughtful design that emphasizes engagement, emotional literacy, and personalized support, future systems can move from simply monitoring emotions to empowering individuals to understand, regulate, and navigate their emotional experiences effectively. This, I believe, represents a meaningful step toward technology that genuinely supports emotional well-being at scale.

Appendix A

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

A.1 Individual Dataset Binning Details

This section outlines the summary of the dataset and information about the binning scheme applied to harmonize our 19 physiological signal-based emotion datasets, facilitating both benchmarking and cross-dataset evaluation. To ensure consistency in labeling across datasets, emotion elicitation tasks or available self-reports were mapped to binary arousal and valence categories when direct arousal and valence self-reports were not provided. Detailed descriptions for each dataset are presented below.

A.1.1 WESAD

The WESAD dataset [339]) is a multimodal dataset containing physiological and motion data from 15 participants recorded via wrist- and chest-worn sensors during a lab study. We utilize the wrist data (gathered using Empatica E4 wristbands) for our experiments, as it includes our target modalities: electrodermal activity (EDA) and photoplethysmogram (PPG). The dataset consists of four affective conditions: baseline (neutral reading task), amusement (viewing humorous videos), stress (Trier Social Stress Test), and meditation (guided breathing). For binary arousal labeling, baseline and meditation are grouped as low arousal, while stress and amusement are labeled as high arousal. For valence, baseline and stress are treated as negative valence due to potential anticipatory anxiety or stress induction, and amusement and meditation are labeled as positive valence.

A.1.2 NURSE

The Nurse dataset [185, 340]) is a multimodal collection of physiological data obtained from 15 nurses working in real-world hospital settings during the COVID-19 pandemic. The data were recorded using Empatica E4 wristbands, capturing signals such as electrodermal activity (EDA), heart rate (HR), skin temperature (TEMP), blood volume pulse (BVP), inter-beat intervals (IBI), and three-axis acceleration (ACC). In addition, periodic smartphone-administered surveys were used to gather context, documenting self-reported stress events and their contributing factors. For our analysis, we focused on the EDA and PPG signals. The labeling mechanism was based on self-reported stress events, where periods labeled as 'stress events' were assigned high arousal and negative valence, while all other periods were categorized as low arousal and positive valence. This dataset was highly imbalanced, as most of the data consisted of periods of high stress, given the challenging hospital environment during the COVID-19 pandemic. This resulted in an overrepresentation of high arousal and negative valence labels, with fewer periods of low stress.

A.1.3 EMOGNITION

The Emognition dataset [166]) comprises multimodal physiological and facial expression data collected from 43 participants. Participants viewed short film clips designed to elicit nine discrete emotions: amusement, awe, enthusiasm, liking, surprise, anger, disgust, fear, and sadness. Physiological signals were recorded using three wearable devices: Muse 2 (EEG, ACC, GYRO), Empatica E4 (BVP, EDA, SKT, ACC), and Samsung Galaxy Watch (BVP, HR, ACC, GYRO). Upper-body videos were simultaneously captured to analyze facial expressions. For our analysis, we focus on EDA and PPG signals collected from E4. Emotional states were annotated using elicitation task labels that were subsequently mapped to arousal and valence dimensions: high arousal includes amusement, surprise, anger, enthusiasm, fear, and awe, while low arousal includes sadness, disgust, baseline, neutral, and liking. Positive valence includes amusement, liking, enthusiasm, and awe, while

negative valence includes sadness, disgust, baseline, surprise, anger, neutral, and fear.

A.1.4 UBFC_PHYS

The UBFC-Phys dataset [341]) is a multimodal dataset comprising physiological and video data collected from 56 participants during a three-phase protocol inspired by the Trier Social Stress Test (TSST). Participants underwent a rest phase (T1), a speech task (T2), and an arithmetic task (T3), with each phase designed to elicit varying levels of stress. Physiological signals, including blood volume pulse (BVP) and electrodermal activity (EDA), were recorded using Empatica E4 wristbands. For our analysis, we focus on the PPG and EDA signals. Labeling was performed by categorizing the rest phase (T1) as low arousal and positive valence, while the speech (T2) and arithmetic (T3) tasks were labeled as high arousal and negative valence, reflecting the increased stress levels associated with these tasks.

A.1.5 VERBIO

The VerBIO dataset [342]) is a multimodal bio-behavioral dataset comprising physiological and audio data collected from 49 participants (originally 55 participants, but data was only available for 49) during 344 public speaking sessions in both real-life and virtual environments. Participants delivered short speeches on assigned topics, with physiological signals recorded using Empatica E4 and Actiwave devices. The original version of the dataset included audio recordings, physiological signals, and self-reported anxiety measures. An updated version incorporates time-continuous stress annotations provided by four annotators, enabling the analysis of moment-to-moment stress levels. For our study, we focus on wrist-derived physiological signals, specifically EDA and PPG. Labeling was performed by categorizing the self-reported measures into periods with high stress annotations as high arousal and negative valence, while periods with low stress annotations were labeled as low arousal and positive valence.

A.1.6 PhyMER

The PhyMER dataset [343]) is a multimodal physiological dataset designed for emotion recognition, incorporating personality traits as contextual information. It comprises physiological signals and personality assessments collected from 30 participants (15 male and 15 female). The physiological data were recorded using two wearable devices: the Emotiv Epoc X headset for EEG signals and the Empatica E4 wristband for EDA, BVP, and peripheral skin temperature. Participants self-reported their emotional responses using a web-based annotation tool, providing ratings on arousal and valence on the SAM scale and categorizing their emotions into seven basic categories: anger, disgust, fear, happiness, neutral, sadness, and surprise on a Likert scale of 1-9. For our study, we focus on wrist-derived physiological signals, specifically EDA and BVP. Labeling was performed by binarizing the self-reported SAM ratings.

A.1.7 EmoWear

The EmoWear dataset [21]) is a multimodal physiological and motion dataset designed for emotion recognition and context awareness. It comprises data from 48 participants (21 females, 27 males) who engaged in a series of tasks, including watching 38 emotionally eliciting video clips, walking a predefined route, reading sentences aloud, and drinking water. Physiological signals were recorded using the Empatica E4 wristband (capturing BVP, EDA, SKT, and accelerometry) and the Zephyr BioHarness (recording ECG, respiration, SKT, and accelerometry). Additionally, three ST SensorTile.box devices were employed to collect accelerometer and gyroscope data. Participants self-assessed their emotional states using the circumplex model of affect, providing ratings on valence, arousal, and dominance scales. For our study, we focus on wrist-derived physiological signals, specifically EDA and BVP. Labeling was performed by binarizing the self-reported arousal valence ratings.

A.1.8 MAUS

The MAUS dataset [344]) is a multimodal physiological dataset designed for mental workload assessment using wearable sensors. It comprises physiological signals collected from 22 participants (2 females) during N-back tasks of varying difficulty levels. The PixArt Watch was used to download the photoplethysmogram (PPG) data. Additionally, a clinical procomp Infinit device was used to record electrocardiography (ECG), galvanic skin response (GSR), and fingertip PPG signals. Participants completed the Pittsburgh Sleep Quality Index (PSQI) questionnaire at the beginning of the experiment and the NASA Task Load Index (NASA-TLX) questionnaire after each N-back task to provide subjective assessments of their sleep quality and perceived workload. For our study, we focus on Procomp data, specifically EDA and PPG. Labeling was performed by categorizing n-back tasks into periods with high workload as high arousal and negative valence, while periods with low workload were labeled as low arousal and positive valence. Specifically, tasks labeled as "0_back" were considered low in cognitive demand and thus assigned low arousal and positive valence. Conversely, tasks labeled as "2_back" or "3_back" were deemed higher in cognitive load, leading to their classification as high arousal and negative valence.

A.1.9 CLAS

The CLAS (Cognitive Load, Affect, and Stress) dataset [345]) is a multimodal physiological dataset designed to support research on the automated assessment of mental states, including cognitive load, affect, and stress. It comprises synchronized recordings of physiological signals—electrocardiography (ECG), photoplethysmography (PPG), electrodermal activity (EDA), and accelerometer data—from 62 healthy volunteers engaged in five tasks: three interactive tasks (math problems, logic problems, and the Stroop test) aimed at eliciting different types of cognitive effort, and two perceptive tasks involving images and videos selected to evoke emotions. For our study, we focus specifically EDA and PPG collected using Shimmer3 GSR+ unit. We applied a binary labeling scheme for valence and arousal

based on task types and stimuli. Tasks such as math tests, Stroop tests, and IQ tests were labeled as high arousal and negative valence, reflecting high cognitive load. Emotion-eliciting videos and images were categorized accordingly: stimuli like videos 2.mp4, 5.mp4, and image set ”pics1” were labeled as high arousal and positive valence, while others like videos 13.mp4, 14.mp4, and image set ”pics2” were labeled as low arousal and negative valence.

A.1.10 CASE

The CASE (Continuously Annotated Signals of Emotion) [64]) dataset is a multimodal physiological dataset designed for emotion analysis. It comprises physiological signals and continuous affect annotations collected from 30 participants (15 male and 15 female) while watching various video stimuli. Physiological data were recorded using ThoughtTech sensors measuring ECG, BVP, EMG, EDA, respiration, and skin temperature. Participants provided real-time continuous annotations of their emotional experiences using a joystick-based interface, simultaneously reporting valence and arousal levels. Labeling was performed by calculating the mean of participants' continuous self-reported annotations for each video segment. Segments with mean valence above the overall average were labeled as positive valence, while those below were labeled as negative valence. Similarly, segments with mean arousal above the average were labeled as high arousal, and those below as low arousal.

A.1.11 Unobtrusive

The Unobtrusive dataset [346]) is a multimodal physiological dataset designed for cognitive load assessment in both controlled (lab) and uncontrolled (real-life) environments. It comprises approximately 315 hours of data collected from 24 participants during a four-hour cognitive load elicitation with self-chosen tasks in the real-life setting and a four-hour mental workload elicitation in a lab setting. Physiological signals were recorded using consumer-grade wearable devices, including the Muse S headband and the Empatica E4 wristband,

capturing EEG, EDA, PPG, and accelerometer data. Participants performed office-like tasks such as mental arithmetic, Stroop, N-Back, and Sudoku with two defined difficulty levels in the lab, and tasks like researching, programming, and writing emails in the uncontrolled environments. Each task was labeled by participants using two 5-point Likert scales of mental workload and stress, as well as the pairwise NASA-TLX questionnaire. For our study, we focus on wrist-derived physiological signals, specifically EDA and PPG. Labeling was performed by categorizing tasks containing keywords such as 'hig', 'nor', 'stroop', 'n_back', 'arithmetix', or 'sudoku' as high arousal and rest tasks as low arousal. Similarly, tasks with keywords like 'hig_mw', 'vhg_mw', or 'hard' were labeled as negative valence, while the rest were labeled as positive valence.

A.1.12 CEAP-360VR

The CEAP-360VR [347]) is a multimodal dataset designed to study emotional responses within immersive virtual reality (VR) environments. It comprises data from 32 participants who each viewed eight one-minute 360° video clips using an HTC Vive Pro Eye head-mounted display. During the viewing sessions, participants provided continuous valence and arousal annotations via a joystick interface. Physiological signals, including electrodermal activity (EDA), blood volume pulse (BVP), heart rate (HR), inter-beat interval (IBI), and skin temperature (SKT), were recorded using the Empatica E4 wristband. Additionally, behavioral data such as head and eye movements and pupil diameter were collected. The dataset also includes responses to questionnaires assessing motion sickness (SSQ), presence (IPQ), and workload (NASA-TLX). For our study, we focus on wrist-derived physiological signals, specifically EDA and BVP. Labeling was performed by calculating the mean of participants' continuous self-reported annotations for each video segment. Segments with mean valence above the overall average were labeled as positive valence, while those below were labeled as negative valence. Similarly, segments with mean arousal above the average were labeled as high arousal, and those below as low arousal.

A.1.13 ScientISST MOVE

The ScientISST MOVE dataset [348, 349]) is a multimodal physiological dataset designed to study the effects of natural everyday activities on biosignal acquisition. It comprises synchronized recordings from 15 healthy participants (originally 17, but the data only included 15 participants) performing activities such as lifting a chair, greeting, gesticulating, walking, and running. Data were collected using three wearable devices: a chestband, an armband, and the Empatica E4 wristband, capturing signals including Electrodermal Activity (EDA), Photoplethysmography (PPG), Electrocardiography (ECG), Electromyography (EMG), skin temperature, and actigraphy. For our study, we focus on wrist-derived physiological signals, specifically EDA and PPG. Labeling was performed by categorizing activities based on their physical intensity and associated emotional valence. High-energy activities such as 'jumps' and 'run' were labeled as high arousal and negative valence, while low-energy activities like 'walk_before_downstairs' and 'baseline' were labeled as low arousal and positive valence. Similarly, activities perceived as positive, such as 'greetings' and 'gesticulate', were labeled as high arousal and positive valence, whereas more strenuous or repetitive tasks like 'lift' were labeled as low arousal and negative valence.

A.1.14 LAUREATE

The LAUREATE dataset [350]) is a comprehensive multimodal dataset designed to facilitate research into the relationship between physiological responses, affective states, and academic performance in real-world educational settings. It comprises physiological data collected from 42 students and 2 lecturers over a 13-week university semester, encompassing 52 sessions that include classes, quizzes, and exams. Participants wore Empatica E4 wristband devices to record physiological signals such as electrodermal activity (EDA), photoplethysmography (PPG), skin temperature, and acceleration signal data. Additionally, daily post-lecture self-reports were gathered using the PANAVA-KS scale and additional custom-designed questions to capture information on lifestyle habits (e.g., study hours,

physical activity, sleep quality), perceived engagement, attention, and emotional states of the students. Similarly, the lecturers' post-class survey included similar questions. For our study, we focus on EDA and PPG data of students and lecturers collected during classes/lecture sessions. The survey items for both students and lecturers included assessments of lecture engagement, such as enthusiasm, motivation, stress, tiredness, peacefulness, happiness, and calmness. These self-reported measures were used to compute composite scores for arousal and valence during lectures. The composite scores were then utilized to determine arousal and valence classes for each physiological data segment collected during lectures.

A.1.15 ForDigitStress

The ForDigitStress dataset [351, 352]) is a multimodal dataset designed to facilitate automatic stress recognition. It comprises data from 38 participants (originally 40 participants, but we could not find self-report files for 2 participants) who engaged in simulated digital job interviews, a scenario chosen to elicit psychosocial stress in a controlled yet realistic environment. Each session included a preparatory phase, an interview session conducted via video call, and a post-interview assessment. During the interviews, participants were subjected to challenging questions regarding their strengths and weaknesses, salary expectations, and hypothetical job-related situations, aimed at inducing stress. The dataset encompasses multiple modalities, including audio recordings, video data capturing facial expressions and body movements, eye-tracking information, and physiological signals, including PPG and EDA signals. To annotate the data, participants provided self-reports on their stress levels and emotions experienced during the interviews. Furthermore, two trained psychologists conducted frame-by-frame annotations of stress and emotions such as shame, anger, anxiety, and surprise, with high inter-rater reliability (Cohen's k ¿ 0.7). For our study, we focus on EDA and PPG data and binary arousal and valence annotations provided by experts after analyzing self-reports and participants' behavior.

A.1.16 Dapper

The DAPPER dataset [20]) is a comprehensive multimodal dataset designed to study emotional experiences in real-world settings. It includes data from 142 participants, with 88 of them providing physiological recordings over five consecutive days. Participants wore custom-designed wrist-worn devices to collect physiological signals such as heart rate (via photoplethysmography), galvanic skin response (GSR), and three-axis acceleration during daytime hours. To capture psychological states, the study employed both the Experience Sampling Method (ESM) and the Day Reconstruction Method (DRM). The ESM involved prompting participants six times daily to report their momentary emotional states, while the DRM required them to recall and describe at least six significant events each day, providing associated emotional ratings. This dual-method approach offered a nuanced view of participants' emotional state in their natural environments. In our experiments, we focused on participants' self-reported arousal and valence levels obtained through ESM. To align physiological data with these self-reports, we extracted GSR and PPG signals recorded within a specific time window, two hours preceding and fifteen minutes following each ESM prompt. This approach enabled us to examine the temporal relationship between physiological responses and reported emotional states in real-world settings.

A.1.17 ADARP

The ADARP (Alcohol and Drug Abuse Research Program) dataset [353, 354, 353]) is a comprehensive multimodal dataset developed to facilitate research on stress detection and alcohol relapse quantification in real-world settings. It encompasses data from 11 individuals (10 females) diagnosed with alcohol use disorder (AUD), collected through a combination of physiological monitoring, self-reported assessments, and structured interviews. Participants in the study wore Empatica E4 wristbands, which continuously recorded physiological signals. In parallel, participants completed ecological momentary assessments (EMA) four times daily over a period of up to 14 days. These EMA surveys captured self-reported

data on emotions, including stress, feeling overwhelmed, and anxiety, using the Positive and Negative Affect Schedule (PANAS) scale. For our study, we focused on the EDA and PPG data and performed binning based on the self-reports provided in the dataset. Segments where participants reported no experiences of stress, feeling overwhelmed, or anxiety were treated as positive valence, while all other segments were treated as negative valence. Regarding arousal, the presence of anxiety was classified as high arousal, whereas reports of stress and feeling overwhelmed were classified as low arousal to balance the dataset. We labeled the physiological data segments within a specific time window: two hours preceding and fifteen minutes following each EMA prompt. This dataset exhibited a significant class imbalance due to all participants being diagnosed with Alcohol Use Disorder (AUD), which led to a predominance of negative affective states such as stress, anxiety, and feeling overwhelmed in the self-reports.

A.1.18 MOCAS

The MOCAS [355]) dataset is a comprehensive resource to facilitate research on human cognitive workload (CWL) assessment in real-world settings. Unlike existing datasets that rely on virtual game stimuli, MOCAS data were collected from realistic closed-circuit television (CCTV) monitoring tasks, enhancing its applicability to practical scenarios. The dataset comprises data from 21 human subjects who performed simultaneous tasks while monitoring CCTV footage. An Empatica E4 wearable watch, Emotive Insight, and a webcam were used to collect data. Physiological signals, including electroencephalography (EEG), BVP, EDA, Skin Temperature, and Accelerometer data, alongside behavioral features such as facial expressions, eye movements, and mouse activity. After each task, participants reported their CWL by completing the NASA-Task Load Index (NASA-TLX) and Instantaneous Self-Assessment (ISA). Additionally, arousal and valence were self-reported from the Self-Assessment Manikin (SAM) scale. We directly used the SAM scale rating for our labeling by categorizing them into two classes, alongside the EDA and PPG signal data segments.

A.1.19 Exercise

The Exercise dataset [356]) provides a comprehensive collection of non-invasive physiological data aimed at advancing research in stress detection and physical activity classification. Data were recorded using the Empatica E4 wearable device, which captures electrodermal activity (EDA), skin temperature, three-axis accelerometry, and blood volume pulse (BVP). The dataset encompasses records from 36 healthy individuals during a structured stress induction protocol, 30 during aerobic exercise, and 31 during anaerobic exercise. The stress induction protocol involved Stroop Test, Trier Mental Challenge Test which included mathematical tasks (with annoying background audio), vocalize their opinion about for and against a controversial topics, and counting backward from 1022 in decrements of 13, each designed to elicit negative physiological responses, while a stationary cycling routine was developed to distinguish between aerobic and anaerobic activities where anaerobic activity has cool-down periods in between cycling sprints and aerobic has increasing resistance cycling with cool-down at the end. For this study, we used EDA and PPG data, and we labeled the physiological data based on the specific tasks performed and the associated stress levels. For the stress induction protocol, we have used stress self-reports where high stress scores were mapped to high arousal negative valence. In contrast, the low stress scores were labeled as low arousal and positive valence. Regarding the exercise sessions, for both aerobic and anaerobic activities, an initial baseline, warm-up, and later cool-down, and rest are labeled as low arousal, while the cycling period was treated as high arousal. For the valence label in aerobic exercise data, speeds up to 85 rpm were taken as positive valence, while the rest of the data was taken as negative valence. For anaerobic exercise, the data of the initial two sprints is taken as positive valence, while later sprints are taken as negative valence.

A.1.20 EEVR

The EEVR dataset [3]) is a multimodal dataset designed to advance emotion recognition research by integrating physiological signals with textual descriptions of emotional experiences. It includes data from 37 participants who were exposed to various emotional stimuli presented through 360° virtual reality (VR) videos. The physiological signals, including electrodermal activity (EDA) and photoplethysmography (PPG), were recorded using a 4-channel Biopac MP36. The dataset encompasses a range of emotional experiences, covering all four quadrants of Russell's circumplex model of emotion. To facilitate emotion classification tasks, the dataset includes annotations for valence and arousal, as well as individual emotions, collected using the SAM and PANAS surveys, along with self-reported textual descriptions of emotions gathered through qualitative interviews. For our experiments, we have directly used the pre-trained models (available here) provided by the authors, which were trained on the EEVR datasets.

A.2 Computation Cost

In this section, we present our computation cost across all modeling paradigms; see Table A.1 for details.

A.3 Benchmarking Models

In this section, we present a detailed discussion of the feature extraction, model architecture, and hyperparameters of the models employed in our benchmarking experiments. All models were initialized with a seed value of 42. We selected all the model's hyperparameters based on hyperparameter tuning to identify the optimal configuration.

Model	Data Type	FLOPS	Parameters	Latency (ms)
RF	EDA	-	-	15.01
	PPG	-	-	14.78
	EDA+PPG	-	-	15.59
LDA	EDA	59	16	0.34
	PPG	143	37	0.38
	EDA+PPG	203	52	0.42
HC+MLP	EDA	3200	1701	0.33
	PPG	7400	3801	0.38
	EDA+PPG	10400	5301	0.43
HC+RESNET	EDA	6225029	414082	0.61
	PPG	14939525	414082	0.62
	EDA+PPG	21164165	414082	0.61
HC+LSTM+NN	EDA	3084037	265346	0.25
	PPG	7309573	265346	0.32
	EDA+PPG	10327813	265346	0.37
HC+Attention+NN	EDA	67717	68226	0.20
	PPG	70405	265346	0.21
	EDA+PPG	72325	72834	0.21
Signal+CNN+Transformer	EDA	5589888	2203086	1.45
	PPG	5589888	2203086	1.43
	EDA+PPG	-	-	-
Signal+LSTM+NN	EDA	12138757	265346	0.41
	PPG	12138757	265346	0.40
	EDA+PPG	-	-	-
Signal+RESNET	EDA	24898949	414082	0.60
	PPG	24898949	414082	0.61
	EDA+PPG	-	-	-
CLSP	EDA	850.51	66.0932 M	3.8492
	PPG	850.51	66.0943 M	3.8977
	EDA+PPG	850.51	66.095 M	3.8676
CLSP-Finetune	EDA	3230.91	66.1247 M	13.0432
	PPG	3230.91	66.1257 M	12.9931
	EDA+PPG	3230.91	66.1265 M	12.9105

Table A.1: Comparison of model performance across data types (EDA, PPG, and EDA+PPG). Here, "M" denotes millions. Missing entries indicate that no experiment was conducted for those cases. For the Random Forest (RF) model, FLOPS and parameter counts are not directly applicable.

A.3.1 ML Models

Random Forest: We used a Random Forest Classifier from the scikit-learn library with the primary hyperparameters set as n_estimators=100, and n_jobs=5, while keeping all other parameters at their default values.

LDA: We used a Linear Discriminant Analysis (LDA) model from the scikit-learn library with the hyperparameter n_components=1, solver = svd, while keeping all other parameters at their default values. The n_components=1 setting indicates that the model projects the input data onto a single linear discriminant axis, which is used because it is generally useful for binary classification problems and for reducing the feature space to a single dimension while preserving class separability.

A.3.2 Handcrafted Features + DL Models

MLP: We trained a Multi-Layer Perceptron (MLP) classifier using scikit-learn's MLPClassifier with the hyperparameters hidden_layer_sizes=(100,) while leaving all other parameters at their default values such as the activation function (relu), solver (adam), learning rate strategy (constant), and maximum iterations (200).

RESNET: We used a 1D ResNet-based architecture tailored for temporal classification tasks. The core building block is a residual module comprising three sequential Conv1D layers with kernel sizes [8, 5, 3], each followed by BatchNorm and ReLU activations except final layer. A shortcut connection using a 1×1 Conv1D aligns input-output dimensions, enabling residual learning. The network stacks two such blocks, followed by an adaptive average pooling layer that compresses the temporal dimension. The pooled features are passed through a fully connected layer and softmax activation for classification. The model is optimized using Adam (lr=0.001) with cross-entropy loss and trained end-to-end with mini-batch gradient descent. We ran for epoch 100 with a batch size of 16.

LSTM+MLP: We implement a sequence classification model based on Long Short-Term Memory (LSTM) networks, followed by a multi-layer perceptron (MLP). The model

begins with a two-layer LSTM configured with a hidden size of 128 and a dropout rate of 0.3 to mitigate overfitting. The LSTM operates on univariate time series data and captures temporal dependencies in the input. The output from the final time step of the LSTM is passed through an MLP consisting of two fully connected layers. The first layer maps the hidden representation to 256 dimensions, applies a ReLU activation, and includes a dropout of 0.4. The second layer reduces the dimensionality to 128, again followed by ReLU and a dropout of 0.3. A final linear layer maps the features to the number of output classes, and a softmax activation is applied to obtain class probabilities. The model is trained using the Adam optimizer with a learning rate of 0.001 and cross-entropy loss. Training is conducted over 100 epochs with a batch size of 16, using shuffled data and GPU acceleration when available.

Attention Layer + MLP: We implement an attention-based neural network classifier designed to operate directly on handcrafted statistical features extracted from physiological signals. These features, computed as summary statistics from time-series data, are treated as a single flat input vector without any temporal or sequential structure. The input vector is first projected into a 128-dimensional representation using a linear layer. A 4-head self-attention mechanism is then applied to model inter-feature dependencies, allowing the model to dynamically weight the contribution of different features during learning. Unlike a full Transformer, our approach does not involve positional encoding or stacked attention layers; rather, it uses a single self-attention block to enhance feature interactions before passing the output through a two-layer MLP with dimensions of 256 and 128, incorporating dropout rates of 0.4 and 0.3, respectively. The final output is produced via a softmax layer, and the model is trained using the Adam optimizer with a learning rate of 0.001 and cross-entropy loss, processing data in batches of 16 samples over 100 epochs.

A.3.3 Signal Segments + DL Models

Resnet: We implemented a deep residual convolutional neural network tailored for classifying physiological time-series signals. First, the raw signals were segmented into fixed-length overlapping windows using a sliding window approach. To ensure all windows had consistent length, we applied zero-padding to the right when segments were shorter than the desired size. These segments were then converted into padded tensors suitable for batched input to the model. Our network comprises two stacked ResNetBlock modules, each engineered to extract hierarchical temporal features. Within each block, the input passes through a sequence of three 1D convolution layers with kernel sizes of 8, 5, and 3, respectively. Each convolution uses 'same' padding to preserve the temporal dimension, followed by batch normalization and ReLU activation to improve training stability and non-linearity. Each block includes a shortcut connection implemented via a 1×1 convolution to facilitate gradient flow. The number of filters is fixed at 128 throughout the network, allowing for rich intermediate representations. After the convolution layers, an adaptive average pooling layer was included to reduce the output to a fixed-length vector, regardless of the input window size. This vector is passed through a fully connected linear layer for classification, followed by a softmax activation to produce probability distributions over the target classes. The model is trained using the Adam optimizer with a learning rate of 0.001, and cross-entropy loss is used to guide optimization. We ran for epoch 100 with a batch size of 16.

LSTM+MLP: We implemented a hybrid LSTM-MLP classifier to model temporal dependencies in univariate physiological time-series data. Prior to modeling, we segmented each signal into overlapping fixed-size windows using a sliding window mechanism. This ensured that even signals of varying lengths could be represented as uniform tensors, with shorter segments padded using zeros. Each windowed sequence was treated as a 1D time series and passed to a multi-layer LSTM module consisting of two stacked layers, each with 128 hidden units and dropout regularization to reduce overfitting. The final hidden state from the LSTM was used as a condensed summary of the temporal dynamics within each

window. This representation was then passed through a multi-layer perceptron (MLP) with two hidden layers (256 and 128 units), ReLU activations, and dropout layers. The final classification was performed using a fully connected layer followed by a softmax activation to produce class probabilities. The model was optimized using the Adam optimizer with a learning rate of 0.001 and trained using the cross-entropy loss. We ran for epoch 100 with a batch size of 8.

CNN+ Transformer Encoder Block: We implemented a hybrid neural architecture combining a convolutional and a Transformer encoder block for 1D physiological signals. The Feature Extractor module uses three parallel 1D convolutional blocks, each with kernel sizes of 5, 9, and 13, respectively, to capture temporal patterns at multiple receptive fields. Each block consists of two convolutional layers: the first maps from 1 input channel to 32 filters, and the second expands from 32 to 64 filters, both with padding="same". Each convolutional layer is followed by batch normalization, ReLU activation, and dropout (rate = 0.2). Outputs from all three branches are concatenated, resulting in a feature map with 192 channels (3 blocks × 64 filters). To adaptively reweight these feature channels, we incorporated a Squeeze-and-Excitation (SE) block with a reduction ratio of 16. This mechanism reduces the channel dimensionality from 192 to 12 via a fully connected layer, then projects it back to 192 using a second linear layer followed by a sigmoid activation, generating channel-wise attention weights. The aggregated feature representation is passed through a global average pooling layer and projected into a 128-dimensional embedding space using a linear layer followed by Layer Normalization and dropout (rate = 0.1). We employed a Transformer encoder with 4 layers, each using 8 attention heads, a model dimension (d_model) of 128, and feedforward sublayers with hidden dimensions of 512. The encoder is preceded by sinusoidal positional encodings added to the input sequence to provide temporal order information. Each encoder layer includes multi-head self-attention, residual connections, layer normalization, and a dropout rate of 0.1. The Transformer output is globally pooled (mean over sequence dimension) and passed to a two-layer MLP classifier:

the first layer maps from 128 to 64 units with ReLU and dropout (0.1), and the second maps from 64 to the two emotion classes. The entire architecture is trained using the Adam optimizer with a learning rate of 0.001 and cross-entropy loss, over 50 epochs with a batch size of 16.

A.3.4 Fine-tuned CLSP Models

MLP-based MetaNet: For the linear variant of the MetaNet modulation network, we employ a simple yet effective two-layer multilayer perceptron (MLP) to generate instance-conditioned prompts. The network first projects the input signal features into a hidden representation of dimension 32 using a fully connected layer, followed by a ReLU nonlinearity. This intermediate representation is then passed through a second linear layer to produce output vectors in the same dimensionality as the CLSP text embedding space. These outputs are reshaped into instance-specific bias vectors and added to a set of **16 learnable context tokens**, resulting in dynamically modulated prompts tailored to each input sample. The overall system is optimized using Adam optimizer with a batch size of 4 and a learning rate of 5e-5 for 15 epoch, enabling efficient fine-tuning. This design facilitates task adaptation by conditioning the text encoder on input signal characteristics, without requiring access to class labels or ground-truth text data during training.

1D-CNN-based MetaNet: For training our model based on 1D-CNN metanet adoption of CoCoOp, we used a batch size of 4 and optimized with the Adam optimizer using a learning rate of 5e-5 for 15 epochs. For prompt learning, we set the number of learnable context tokens to 24, each embedded in a 768-dimensional space aligned with the text encoder. The Meta network comprises two stacked 1D convolutional layers. The first convolution uses an input channel of 1 and outputs 24 hidden channels (with kernel size 3, stride 1, and padding 1), followed by a ReLU activation, and a second convolution compresses the representation back to a single output (with kernel size of 3 and padding 1) channel, maintaining the temporal dimension. This network transforms the signal features into instance-specific

context bias vectors, which are added to a set of **24 learnable context tokens**, forming dynamic prompts. These prompts are prepended to tokenized class descriptions and passed to a frozen DistilBERT encoder, enabling adaptive conditioning of text embeddings based solely on input signals. This 1D-CNN modulation strategy provides a computationally efficient and effective means of aligning physiological data with textual semantics in the absence of ground-truth annotations.

For our fine-tuned CLSP experiments, we employed a set of carefully constructed textual prompts corresponding to each emotional category. These prompts were designed to provide richer semantic grounding for the text encoder, enabling the model to better capture the conceptual meaning of each class even in the absence of explicit textual supervision. Each prompt describes general physiological and affective cues, such as variations in energy or bodily reactions, that are broadly recognized across cultures, thereby reflecting universal aspects of emotional experience rather than culture-specific expressions. Importantly, our approach also extends beyond static textual class definitions, since in our fine-tuning approach, each prompt is augmented with context tokens that are learned for every input segment. These adaptive tokens enable the model to dynamically refine the prompt representation based on the input, thereby mitigating potential rigidity and bias that may arise from the nature of textual prompts. The complete set of textual prompts used for our fine-tuning experiments is presented below:

1) Textual Prompts for Arousal Classification:

- **High Arousal**: *"The participant felt a strong physical reaction, like a racing heart or tense body, and experienced high-energy emotions such as excitement, enthusiasm, surprise, anger, and nervousness."*
- **Low Arousal**: *"The participant felt low energy and relaxed, with calm emotions like peacefulness, relaxation, neutral, boredom, and lack of interest."*

2) Textual Prompts for Valence Classification:

- **Negative Valence**: *"The participant felt bad and was in a negative mood, with emotions like sadness, fear, anger, worry, hopelessness, and frustration."*
- **Positive Valence**: *"The participant experienced a positive mood characterized by emotions such as happiness, joy, gratitude, serenity, interest, hope, pride, amusement, inspiration, awe, and love."*

3) Textual Prompts for Four Class Classification:

- **High Arousal Negative Valence**: *"Strong physical reaction with intense negative emotions like anger, fear, frustration, anxiety, or panic."*
- **High Arousal Positive Valence**: *"Strong physical activation with energizing positive emotions like joy, enthusiasm, exhilaration, or amusement."*
- **Low Arousal Negative Valence**: *"Low energy with subdued negative emotions like sadness, boredom, tiredness, disappointment, or hopelessness."*
- **Low Arousal Positive Valence**: *"Calm and relaxed with subtle positive emotions like contentment, peace, satisfaction, and mild happiness."*

Table A.2: Dataset Categorization by Experimental Setting

Setting Group	Datasets
Lab	WESAD, EMOGNITION, UBFC_PHYS, VERBIO, PhyMER, EmoWear, CEAP-360VR, CASE, MOCAS, MAUS, CLAS
Constraint	ForDigitStress, ScientISST MOVE, Exercise
Lab+Real	Unobtrusive
Real	ADARP, Dapper, NURSE, LAUREATE

A.4 Cross-Dataset Analysis

The cross-data models were run using the same set of parameters as in the benchmarking stage, with a seed of 42 and the same number of epochs as defined per model in A.3.

Table A.3: Dataset Categorization by Device Type

Device Group	**Datasets**
Wearable (Empatica E4)	WESAD, NURSE, EMOGNITION, UBFC_PHYS, VERBIO, PhyMER, EmoWear, Unobtrusive, CEAP-360VR, ScientISST MOVE, LAUREATE, MOCAS, Exercise, ADARP
Lab Based Device	CASE (ThoughtTech SA9309M, ThoughtTech SA9308M) CLAS (Shimmer3 GSR+ Unit) MAUS (Procomp Infinit) ForDigitStress (IOMbiofeedback sensor)
Custom Wearable	Dapper (Custom Designed Wristband)

A.4.1 Dataset Grouping

The dataset grouping across different harmonizing dimensions: Experimental Setting, Device Type, and Labeling Method is added in Table A.2, A.3, and A.4. Furthermore, to evaluate gender-based transferability, we selected nine datasets containing gender metadata: WESAD, ScientISST MOVE, UBFC_PHYS, Exercise, PhyMER, EmoWear, CASE, CEAP-360VR, and NURSE (female subjects only). For age-based transferability analysis, we employed seven datasets: WESAD, ScientISST MOVE, Exercise, PhyMER, EmoWear, CASE, and CEAP-360VR.

Table A.4: Dataset Categorization by Labeling Method

Label Group	**Datasets**
Stimulus-Label	WESAD, EMOGNITION, UBFC_PHYS, MAUS, CLAS, ScientISST MOVE, Exercise
Self-report	VERBIO, PhyMER, EmoWear, CASE, Unobtrusive, NURSE, CEAP-360VR, LAUREATE, Dapper, ADARP, MOCAS
Expert-Annotated	ForDigitStress

A.5 All Results

In this section, we present our detailed summary of our results for data benchmarking and cross-dataset experiments. Check out the supplementary material for more detailed results.

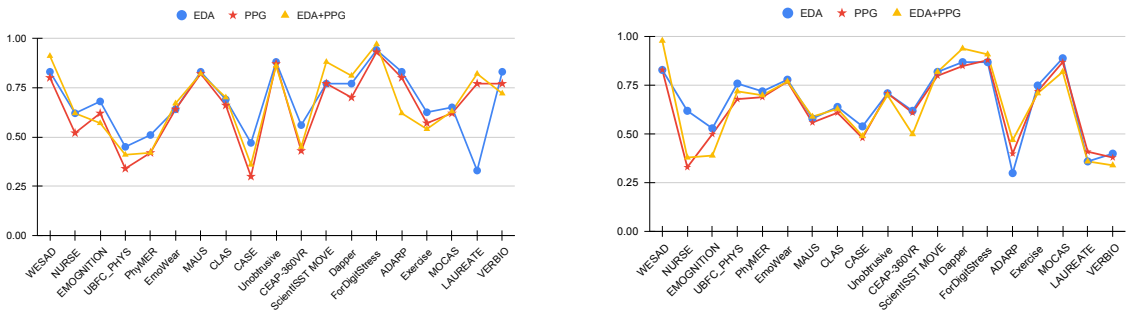


Figure A.1: Comparative performance (F1 score) of the best-performing models per dataset across three physiological modalities (EDA, PPG, EDA+PPG) for emotion recognition. Each line represents a modality, showing how its top-performing model varies in effectiveness across the 19 datasets. Left: For arousal classification. Right: For valence classification.

A.5.1 Benchmarking Results

In this section, we present the benchmarking results for valence, arousal, and four-class classification. Table 3.6 presents the best-performing models and their F1 scores for four-class classification across all datasets. We then show dataset-wise results across all 16 modeling paradigms in Figures A.3, A.4, A.5, A.6, A.7, and A.8 using radar plots for both arousal and valence classification across three modalities: EDA, PPG and EDA+PPG. In Figures A.9, A.10, and A.11, we present dataset-wise performance visualizations to compare datasets and their relative positioning with respect to each other. Qualitative analysis is further discussed in Tables 3.7 and **??**.

A.5.2 Cross-Data Analysis Results

In this section, we present our results for cross-dataset analysis as detailed in Tables A.7, A.8 for device dimension, Tables A.5, A.6 for setting dimension, and Tables A.9, A.10 for labeling dimension. We further visualized the impact of these dimensions on the EDA and PPG features using UMAP as shown in figures A.12, A.13, A.14. Moreover, the summarized results for gender-wise transferability are added in Tables A.11 and A.12, and age-wise transferability is added in Tables A.13 and A.14.

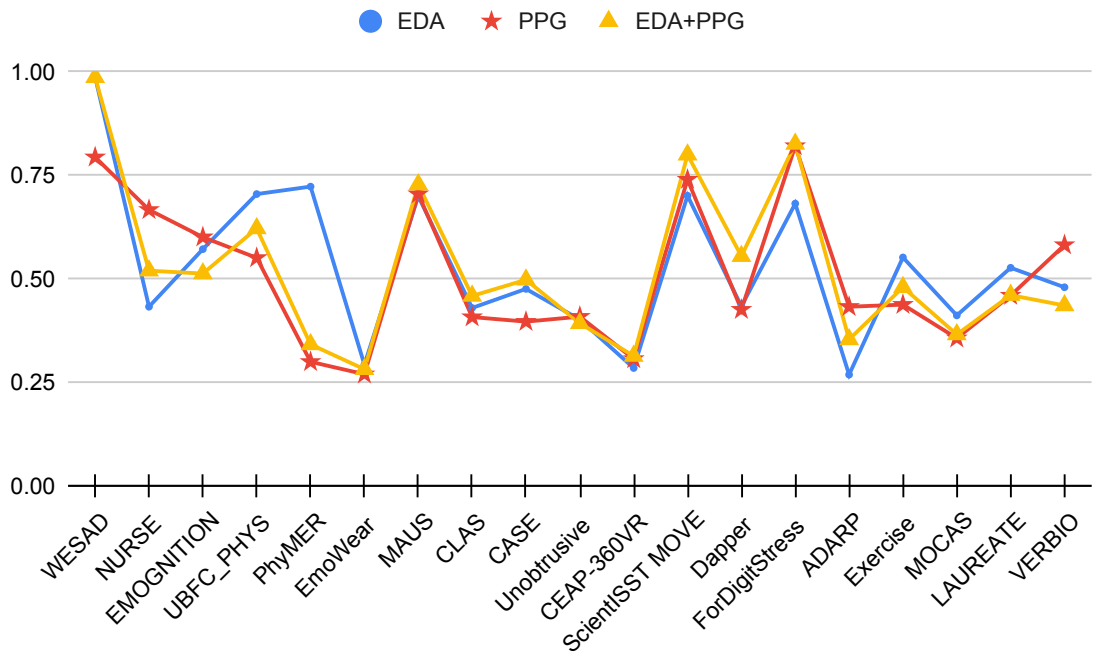


Figure A.2: Comparative performance (F1 score) of the best-performing models for four-class classification per dataset across three physiological modalities (EDA, PPG, EDA+PPG) for emotion recognition. Each line represents a modality, showing how its top-performing model varies in effectiveness across the 19 datasets.

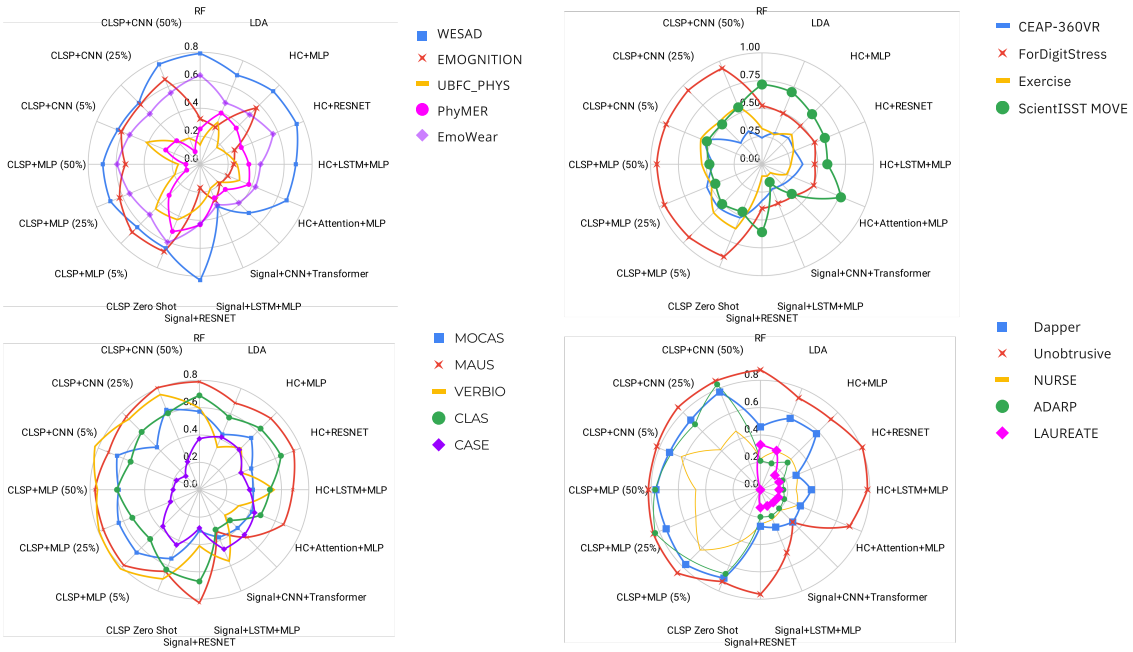


Figure A.3: Benchmarking results for Arousal Classification on EDA signal Data across 19 datasets for 4 modeling paradigms and 16 model variants.

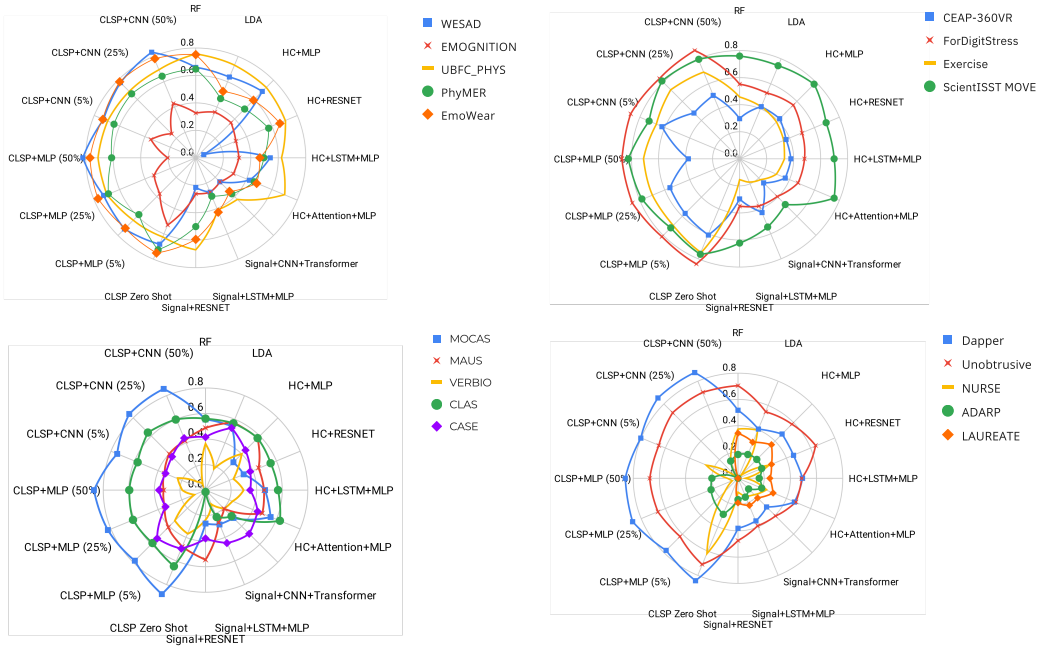


Figure A.4: Benchmarking results for Valence Classification on EDA signal Data across 19 datasets for 4 modeling paradigms and 16 model variants.

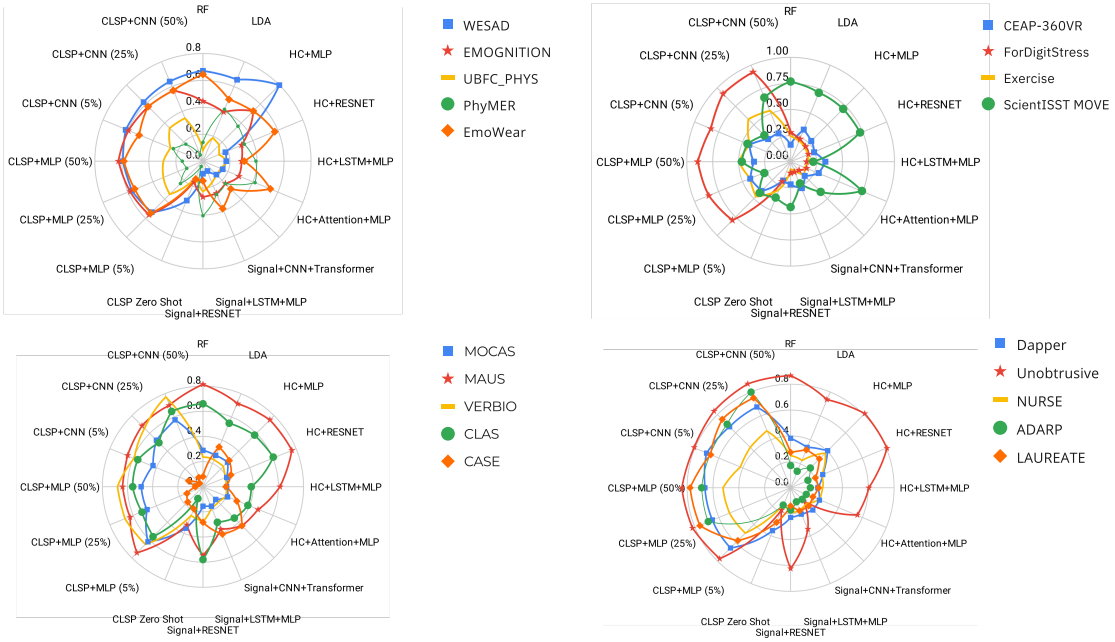


Figure A.5: Benchmarking results for Arousal Classification on PPG signal Data across 19 datasets for 4 modeling paradigms and 16 model variants.

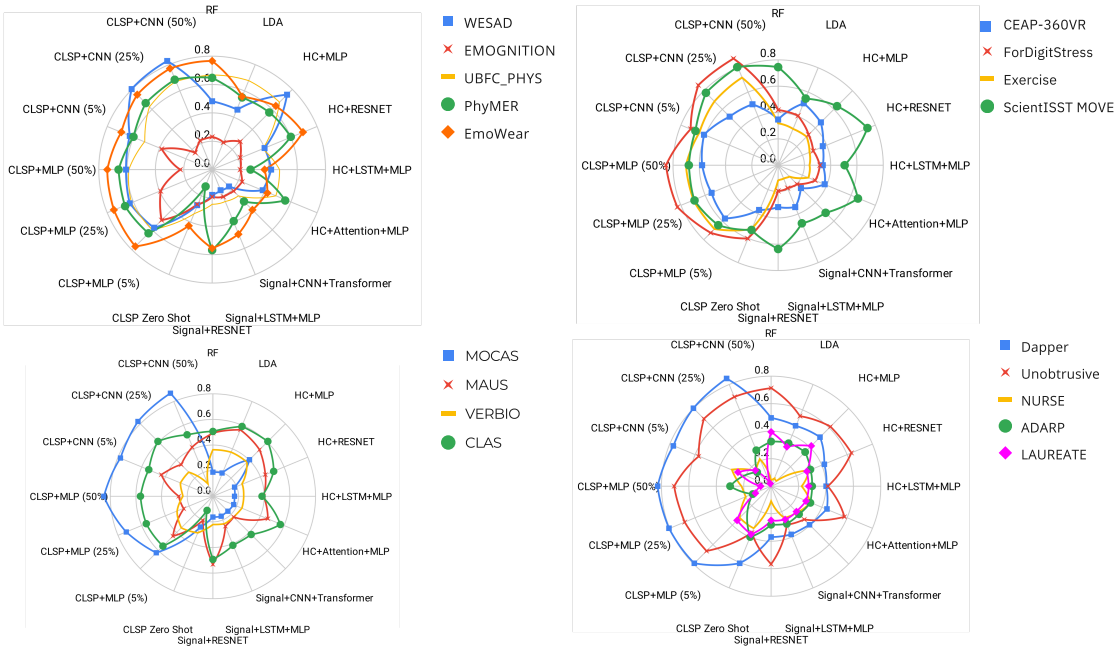


Figure A.6: Benchmarking results for Valence Classification on PPG signal Data across 19 datasets for 4 modeling paradigms and 16 model variants.

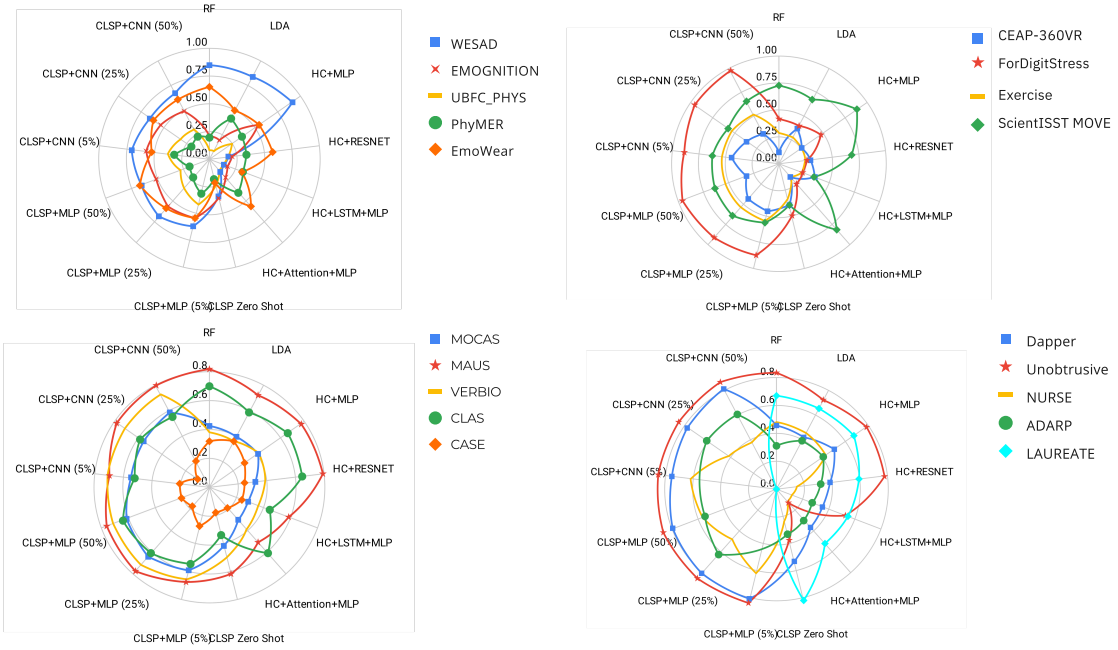


Figure A.7: Benchmarking results for Arousal Classification on EDA+PPG Data across 19 datasets for 4 modeling paradigms and 16 model variants.

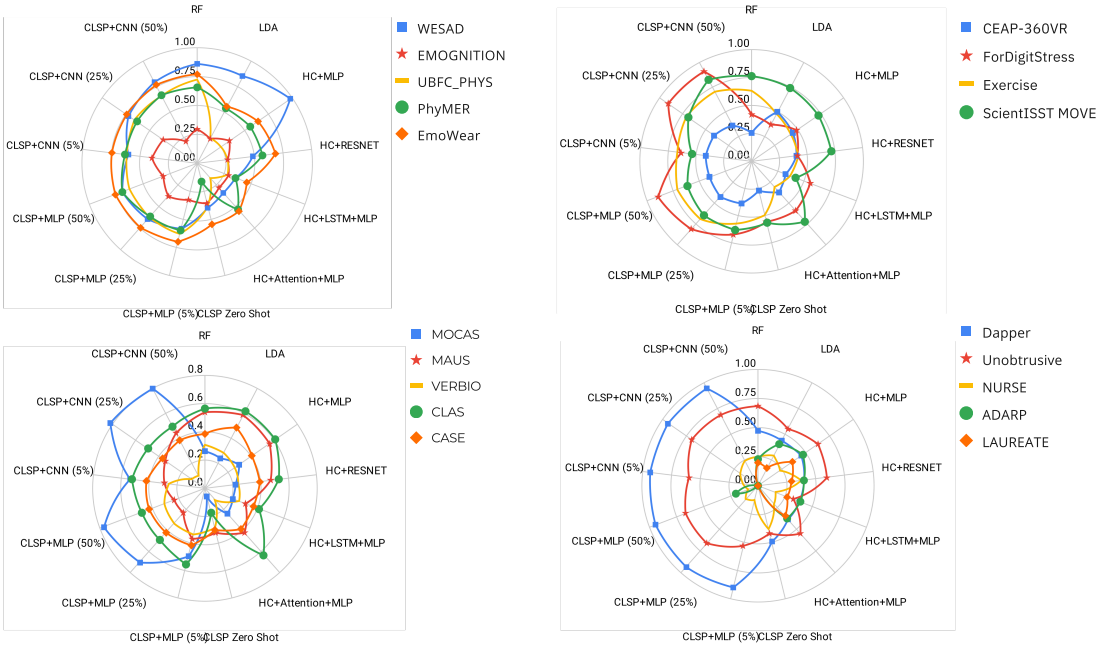


Figure A.8: Benchmarking results for Valence Classification on EDA+PPG Data across 19 datasets for 4 modeling paradigms and 16 model variants.

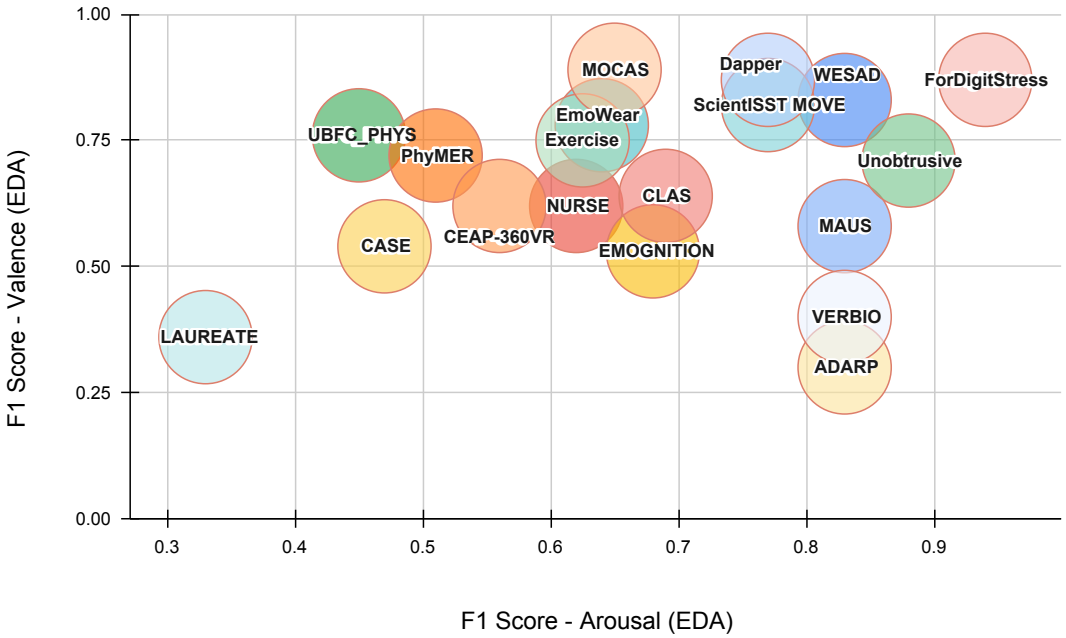


Figure A.9: **Benchmarking results: EDA only.** This bubble plot illustrates the impact of EDA signals on F1 performance (best model) for arousal and valence classification across 19 datasets.

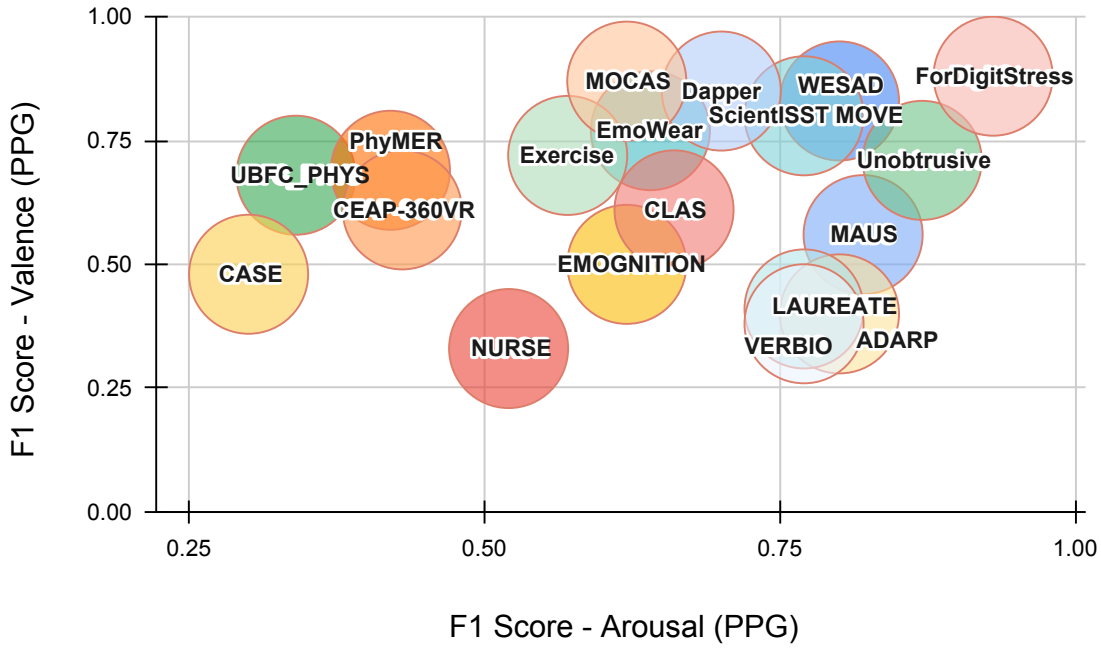


Figure A.10: **Benchmarking results: PPG only.** This bubble plot illustrates the impact of PPG signals on F1 performance (best model) for arousal and valence classification across 19 datasets.

Testing Cohort	Training Cohort	EDA		PPG		EDA+PPG	
		Best Model	**F1**	**Best Model**	**F1**	**Best Model**	**F1**
Lab	Real	**CLSP CNN 5%**	**0.72**	CLSP MLP 5%	0.57	**CLSP MLP 5%**	**0.71**
Lab	Constraint	CLSP MLP 50%	0.56	**RF**	**0.61**	RF	0.60
Lab	Lab	RF	0.50	RF	0.50	RF	0.52
Lab	CLSP ZeroShot	-	0.58	-	0.15	-	0.29
Constraint	Real	RF	0.68	RF	0.51	**CLSP MLP 5%**	**0.64**
Constraint	Lab	HC+MLP	0.44	**LDA**	**0.67**	**LDA**	**0.64**
Constraint	Constraint	HC+MLP	0.48	RF	0.48	RF	0.48
Constraint	CLSP ZeroShot	-	**0.74**	-	0.27	-	0.40
Real	Constraint	**CLSP MLP 5%**	**0.65**	RF	0.59	**CLSP MLP 5%**	**0.73**
Real	Lab	HC+MLP	0.59	**LDA**	**0.69**	CLSP MLP 25%	0.72
Real	Real	HC+MLP	0.49	RF	0.48	RF	0.46
Real	CLSP ZeroShot	-	**0.65**	-	0.31	-	0.52

Table A.5: For each data-collection setting category (Lab, Constraint, and Real) and modality (EDA, PPG, and EDA+PPG), the table identifies the top-performing model and its F1 score for **arousal classification**.

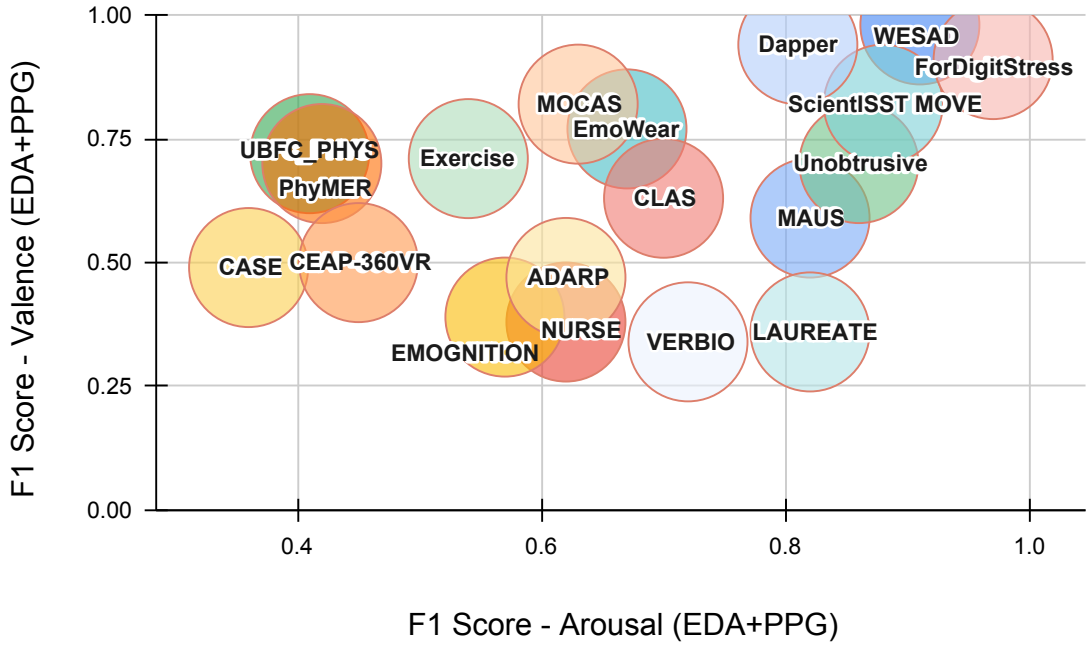


Figure A.11: **Benchmarking results: EDA + PPG combined.** This bubble plot illustrates the impact of combining EDA and PPG signals on F1 performance (best model) for arousal and valence classification across 19 datasets.

Testing Cohort	Training Cohort	EDA		PPG		EDA+PPG	
		Best Model	F1	Best Model	F1	Best Model	F1
Lab	Real	**CLSP MLP 5%**	**0.79**	**RF**	**0.69**	**CLSP MLP 25%**	**0.79**
Lab	Constraint	CLSP MLP 25%	0.66	CLSP CNN 5%	0.67	CLSP MLP 25%	0.68
Lab	Lab	RF	0.54	HC+MLP	0.50	HC+MLP	0.51
Lab	CLSP ZeroShot	-	0.67	-	0.28	-	0.36
Constraint	Real	RF	0.76	**RF**	**0.78**	**RF**	**0.77**
Constraint	Lab	RF	0.76	RF	0.72	RF	0.74
Constraint	Constraint	RF	0.63	RF	0.64	RF	0.65
Constraint	CLSP ZeroShot	-	**0.79**	-	0.55	-	0.52
Real	Constraint	RF	0.76	**RF**	**0.70**	**RF**	**0.88**
Real	Lab	RF	0.72	CLSP MLP 25%	0.64	RF	0.76
Real	Real	HC+MLP	0.41	HC+MLP	0.41	HC+MLP	0.42
Real	CLSP ZeroShot	-	**0.77**	-	0.50	-	0.49

Table A.6: For each data-collection setting category (Lab, Constraint, and Real) and modality (EDA, PPG, and EDA+PPG), the table identifies the top-performing model and its F1 score for **valence classification**.

Testing Cohort	Training Cohort	EDA		PPG		EDA+PPG	
		Best Model	**F1**	**Best Model**	**F1**	**Best Model**	**F1**
Custom Wearable	Wearable	LDA	0.69	**LDA**	**0.69**	CLSP MLP 50%	0.73
Custom Wearable	Lab Based Device	**RF**	**0.72**	HC+MLP	0.65	**CLSP CNN 25%**	**0.78**
Custom Wearable	Custom Wearable	HC+MLP	0.57	HC+MLP	0.40	HC+MLP	0.50
Custom Wearable	CLSP ZeroShot	-	**0.72**	-	0.44	-	0.58
Lab Based Device	Wearable	LDA	0.58	HC+MLP	0.45	CLSP CNN 50%	0.50
Lab Based Device	Custom Wearable	**CLSP CNN 5%**	**0.67**	**RF**	**0.62**	**CLSP CNN 5%**	**0.65**
Lab Based Device	Lab Based Device	RF	0.54	RF	0.54	RF	0.56
Lab Based Device	CLSP ZeroShot	-	0.63	-	0.35	-	0.50
Wearable	Custom Wearable	CLSP MLP 5%	0.65	**CLSP CNN 5%**	**0.60**	**CLSP CNN 25%**	**0.66**
Wearable	Lab Based Device	**CLSP MLP 25%**	**0.68**	**HC+MLP**	**0.60**	CLSP CNN 25%	0.59
Wearable	Wearable	HC+MLP	0.49	HC+MLP	0.48	HC+MLP	0.52
Wearable	CLSP ZeroShot	-	0.59	-	0.43	-	0.52

Table A.7: For each device category (Wearables, Custom Wearable, and Lab-Based Device) and modality (EDA, PPG, and EDA+PPG), the table identifies the top-performing model and its F1 score for **arousal classification**. Here wearable is Empatica E4.

Testing Cohort	Training Cohort	EDA		PPG		EDA+PPG	
		Best Model	**F1**	**Best Model**	**F1**	**Best Model**	**F1**
Custom Wearable	Wearable	CLSP MLP 50%	0.82	**CLSP CNN 5%**	**0.72**	CLSP MLP 5%	0.78
Custom Wearable	Lab Based Device	LDA	0.67	CLSP MLP 50%	0.67	**LDA**	**0.81**
Custom Wearable	Custom Wearable	RF	0.52	HC+MLP	0.50	RF	0.47
Custom Wearable	CLSP ZeroShot	-	**0.83**	-	0.60	-	0.54
Lab Based Device	Wearable	CLSP CNN 50%	0.62	CLSP MLP 25%	0.61	CLSP CNN 5%	0.62
Lab Based Device	Custom Wearable	**CLSP CNN 50%**	**0.64**	**CLSP CNN 50%**	**0.63**	**CLSP CNN 5%**	**0.63**
Lab Based Device	Lab Based Device	RF	0.53	RF	0.51	RF	0.52
Lab Based Device	CLSP ZeroShot	-	0.60	-	0.59	-	0.45
Wearable	Lab Based Device	**CLSP CNN 50%**	**0.72**	**CLSP CNN 50%**	**0.73**	**CLSP CNN 50%**	**0.73**
Wearable	Custom Wearable	CLSP CNN 25%	0.51	CLSP MLP 5%	0.58	LDA	0.53
Wearable	Wearable	HC+MLP	0.50	HC+MLP	0.55	HC+MLP	0.54
Wearable	CLSP ZeroShot	-	0.70	-	0.62	-	0.53

Table A.8: For each device category (Wearables, Custom Wearable, and Lab-Based Device) and modality (EDA, PPG, and EDA+PPG), the table identifies the top-performing model and its F1 score for **valence classification**.

Testing Cohort	Training Cohort	EDA		PPG		EDA+PPG	
		Best Model	**F1**	**Best Model**	**F1**	**Best Model**	**F1**
Stimulus-Label	Expert-Annotated	**CLSP MLP 5%**	**0.64**	**RF**	**0.72**	**CLSP MLP 50%**	**0.65**
Stimulus-Label	Self-report	RF	0.62	CLSP CNN 5%	0.44	CLSP CNN 5%	0.57
Stimulus-Label	Stimulus-Label	RF	0.54	RF	0.51	HC+MLP	0.55
Stimulus-Label	CLSP ZeroShot	-	0.60	-	0.37	-	0.50
Self-report	Expert-Annotated	**CLSP MLP 5%**	**0.65**	**CLSP CNN 50%**	**0.64**	**CLSP MLP 5%**	**0.69**
Self-report	Stimulus-Label	HC+MLP	0.57	CLSP CNN 50%	0.51	RF	0.63
Self-report	Self-report	HC+MLP	0.53	HC+MLP	0.52	HC+MLP	0.52
Self-report	CLSP ZeroShot	-	0.60	-	0.43	-	0.54
Expert-Annotated	Self-report	RF	0.87	**LDA**	**0.69**	**RF**	**0.84**
Expert-Annotated	Stimulus-Label	CLSP CNN 50%	0.79	CLSP MLP 50%	0.70	RF	0.82
Expert-Annotated	Expert-Annotated	RF	0.52	RF	0.28	HC+MLP	0.48
Expert-Annotated	CLSP ZeroShot	-	**0.91**	-	0.39	-	0.68

Table A.9: For each labeling method category (Stimulus-Labels, Self-report, and Expert-Annotated) and modality (EDA, PPG, and EDA+PPG), the table identifies the top-performing model and its F1 score for **arousal classification**.

Testing Cohort	Training Cohort	EDA		PPG		EDA+PPG	
		Best Model	**F1**	**Best Model**	**F1**	**Best Model**	**F1**
Stimulus-Label	Expert-Annotated	**CLSP MLP 5%**	**0.65**	**CLSP CNN 50%**	**0.65**	**CLSP CNN 25%**	**0.65**
Stimulus-Label	Self-report	CLSP CNN 25%	0.63	CLSP CNN 5%	0.61	CLSP CNN 5%	0.61
Stimulus-Label	Stimulus-Label	RF	0.61	RF	0.53	RF	0.52
Stimulus-Label	CLSP ZeroShot	-	0.63	-	0.59	-	0.48
Self-report	Expert-Annotated	CLSP MLP 50%	0.69	**RF**	**0.72**	**CLSP CNN 50%**	**0.76**
Self-report	Stimulus-Label	LDA	0.57	CLSP MLP 5%	0.59	LDA	0.56
Self-report	Self-report	RF	0.53	HC+MLP	0.48	HC+MLP	0.52
Self-report	CLSP ZeroShot	-	**0.70**	-	0.62	-	0.53
Expert-Annotated	Self-report	CLSP CNN 25%	0.83	**CLSP CNN 50%**	**0.85**	**CLSP CNN 5%**	**0.87**
Expert-Annotated	Stimulus-Label	**LDA**	**0.87**	**RF**	**0.85**	CLSP CNN 50%	0.74
Expert-Annotated	Expert-Annotated	HC+MLP	0.56	RF	0.42	HC+MLP	0.49
Expert-Annotated	CLSP ZeroShot	-	0.83	-	0.60	-	0.43

Table A.10: For each labeling method category (Stimulus-Labels, Self-report, and Expert-Annotated) and modality (EDA, PPG, and EDA+PPG), the table identifies the top-performing model and its F1 score for **valence classification**.

Testing Cohort	Training Cohort	EDA		PPG		EDA+PPG	
		Best Model	**F1**	**Best Model**	**F1**	**Best Model**	**F1**
Male	Female	HC+MLP	**0.56**	LDA	**0.51**	LDA	0.54
Male	Male	RF	**0.56**	HC+MLP	**0.51**	RF	**0.56**
Male	CLSP ZeroShot	-	0.56	-	0.16	-	0.24
Female	Male	LDA	0.50	LDA	0.51	LDA	0.53
Female	Female	RF	0.52	HC+MLP	**0.55**	HC+MLP	**0.56**
Female	CLSP ZeroShot	-	**0.54**	-	0.15	-	0.25

Table A.11: Best-performing models for **arousal classification** across gender groups and modalities. For each dataset, gender group (Male, Female), and modality combination (EDA, PPG, EDA+PPG), we report the model achieving the highest F1 score.

Testing Cohort	Training Cohort	EDA		PPG		EDA+PPG	
		Best Model	**F1**	**Best Model**	**F1**	**Best Model**	**F1**
Male	Female	CLSP MLP 25%	**0.69**	RF	**0.71**	CLSP CNN 50%	**0.70**
Male	Male	HC+MLP	0.53	HC+MLP	0.52	RF	0.47
Male	CLSP ZeroShot	-	**0.69**	-	0.35	-	0.42
Female	Male	CLSP MLP 50%	**0.71**	CLSP CNN 50%	**0.70**	CLSP MLP 25%	**0.70**
Female	Female	HC+MLP	0.55	RF	0.49	HC+MLP	0.54
Female	CLSP ZeroShot	-	0.69	-	0.34	-	0.42

Table A.12: Best-performing models for **valence classification** across gender groups and modalities. For each dataset, gender group (Male, Female), and modality combination (EDA, PPG, EDA+PPG), we report the model achieving the highest F1 score.

Figure A.12: UMPA Visualization of EDA and PPG Features color-coded by Labeling Techniques

Figure A.13: UMPA Visualization of EDA and PPG Features color-coded by Device Type. Note: e4 here is wearable cohort, since all wristworn wearable devices were empatice e4.

Figure A.14: UMPA Visualization of EDA and PPG Features color-coded by Experiment Collection Setting

Testing Cohort	Training Cohort	EDA		PPG		EDA+PPG	
		Best Model	F1	Best Model	F1	Best Model	F1
Old	Young	LDA	0.51	HC+MLP	0.56	HC+MLP	**0.56**
Old	Old	RF	0.55	RF	0.55	RF	0.53
Old	CLSP ZeroShot	-	**0.56**	-	**0.7**	-	0.19
Young	Old	LDA	0.5	LDA	0.43	CLSP MLP 50%	0.47
Young	Young	HC+MLP	0.55	RF	0.53	HC+MLP	**0.58**
Young	CLSP ZeroShot	-	**0.56**	-	**0.72**	-	0.28

Table A.13: Best-performing models for **arousal classification** across age groups and modalities. For each dataset, age group (Young: 18–25 years, Old: 25+ years), and modality combination (EDA, PPG, EDA+PPG), we report the model achieving the highest F1 score.

Testing Cohort	Training Cohort	EDA		PPG		EDA+PPG	
		Best Model	F1	Best Model	F1	Best Model	F1
Old	Young	CLSP MLP 50%	**0.73**	CLSP CNN 50%	**0.72**	RF	**0.73**
Old	Old	RF	0.53	RF	0.57	RF	0.53
Old	CLSP ZeroShot	-	0.14	-	0.37	-	0.47
Young	Old	CLSP MLP 5%	**0.72**	RF	**0.67**	RF	**0.69**
Young	Young	RF	0.54	RF	0.51	RF	0.48
Young	CLSP ZeroShot	-	0.16	-	0.31	-	0.35

Table A.14: Best-performing models for **valence classification** across age groups and modalities. For each dataset, age group (Young: 18–25 years, Old: 25+ years), and modality combination (EDA, PPG, EDA+PPG), we report the model achieving the highest F1 score.

Appendix B

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

B.1 EEVR Overview

The *EEVR* dataset comprises synchronized pairs of physiological signals and textual data. It includes responses to four self-assessment questions regarding perceived arousal, valence, dominance, and discrete emotions ratings collected using PANAS questionnaires (which were further utilized to calculate Positive and Negative Affect Score). The EEVR dataset was collected using Virtual Reality (VR) 360° videos as the elicitation medium. The videos utilized in the dataset were selected based on their arousal and valence ratings to cover all four quadrants of the Russell circumplex emotion model ([72]), as shown in Figure 5.2. The remainder of the supplementary materials provide detailed information about the *EEVR* dataset. Figure B.1 provides a datasheet for the EEVR dataset based on [357]. The experiment setup is presented in Figure 5.3a.

B.1.1 EEVR Size Details

The EEVR dataset consists of data from 37 healthy participants who agreed to share their data publicly. Although 41 participants were involved in the data collection study, data from only 37 participants is publicly available. During data acquisition, the data from four participants was damaged due to factors like motion sickness in the VR environment and issues with sensor attachment. Consequently, the dataset provides physiological signal data (including Electrodermal Activity (EDA) and Photoplethysmogram (PPG) signals) and textual descriptions of emotions felt during each emotional stimulus for 37 participants.

The EEVR dataset comprises 296 tasks in total, with each participant experiencing eight VR 360° videos shown to induce emotions from all four quadrants of the Russell

emotion model (two videos from each quadrant). Table B.1 presents a summary of the minute durations for each video, along with their respective playlist details. Further details regarding the playlist and video order are elaborated in Section B.4.2.

Video Name	**Count (minutes)**	**Playlist**
The Displaced	3:23	1, 3
Happyland 360	2:43	1, 3
Jailbreak 360	3:06	1, 3
War Knows No Nation	3:15	1, 3
Canyon Swing	1:44	1, 3
Redwoods Walk Among Giants	2:00	1, 3
Speed Flying	2:34	1, 3
Instant Caribbean Vacation	2:30	1, 3
The Nepal Earthquake Aftermath	3:09	2, 4
Zombie Apocalypse Horror	3:00	2, 4
Abandoned building	3:00	2, 4
Kidnapped	2:58	2, 4
Mega Coaster	1:57	2, 4
Malaekahana Sunrise	3:29	2, 4
Puppies host SourceFed for a day	1:20	2, 4
Great Ocean Road	1:58	2, 4

Table B.1: Video names with their duration details and the playlist number

B.1.2 EEVR Organization and File formats

The EEVR dataset, as downloaded, is organized into two main subdirectories, as illustrated in Figure B.2. The first subdirectory contains processed physiological data, including CSV files with raw EDA and PPG data, organized by participant details and Video ID for ease of use. It also includes EDA and PPG features CSV files extracted using the feature extraction pipeline. The second subdirectory holds raw data and is divided into four playlist folders. Each playlist folder contains directories for subject-wise raw EDA text files, raw PPG text files, annotation text files, and raw ACQ files in the original Biopac format. All the physiological data files are organized in .CSV and .TXT formats, making them easily usable for all programming languages. All physiological signals were initially sampled at 2000Hz but were downsampled to 128Hz for PPG and 15.68Hz for EDA to reduce computational

EEVR Dataset Facts

Dataset EEVR

Motivation

Summary EEVR is a multimodal dataset designed for emotion recognition. It comprises physiological signal data collected from wearable sensors along with raw textual captions corresponding to each emotion elicitation segment.

Example Use Case	Emotion Recognition, Arousal classification, Valence classification, Personality Recognition
Original Authors	P. Singh, R. Budhiraja, A. Gupta, A. Goswami, M. Kumar, P. Singh

MetaData

URL	https://melangelabiiitd.github.io/EEVR/
KeyWords	Emotion Recognition, Wearable sensor, Physiological signal
Format	.acq, .csv, .txt
Ethical Review Approval	IRB-IIIT-Delhi, NECRBHR
License	CC BY-NC-SA
First Release Year	2024

Sensors

EDA	SS57LA, 4-channel Biopac MP36
PPG	SS4LA, 4-channel Biopac MP36

Data Annotation

Self-Assessments	Arousal, Valence, Dominance, Positive Affect Score, Negative Affect Score, Familiarity, Discrete Emotions
Textual Labels	Raw Textual Description per video stimuli
Additional Data	Personality Score (BFI10), GHQ, VRSQ

Participants

Count	37
Gender	21 males, 16 females)
Age	18-33 (M=23.1, SD=4.02)
Backround	Bachelor's (24), Master's (8), Senior High School (4), Doctorate (3)

Dataset Size

Total Size	668 MB
Physiological Data Duration	797 minutes and 83 seconds

Figure B.1: Dataset Summary card for EEVR, constructed based on [357].

requirements. The .ACQ files contain the original 2000Hz data, while the .TXT and .CSV files are the downsampled versions used for experimentation. The downsampling frequencies were chosen based on previous research to ensure no information was lost. Participant 29's data was collected in two parts due to a disconnection during the experiment, resulting in two raw data files.

Additionally, there are three files: one detailing participant information collected during the data collection process, second with self-assessment details collected during data collection, and third with text data from semi-structured interviews for each participant-video segment. We have also included VR_Application.zip file containing the VR environment simulation build file and video resources. The Participant_details is organized into sheets for Participant Details, GHQ-12, and BFI-10. The Self_Assessment file is further organized into sheets for the Pre-exposure Questionnaire, Post-exposure Questionnaire, Affect-Personality Score, and VRSQ Scores. The text data file also contains sheets for Text-Labels (Text description with information on participant ID and video ID and labels), Video description, and Video ID and Video Name mapping information.

B.2 EEVR Usage and Publishing

B.2.1 Intended Uses

The *EEVR* dataset is collected and published to further the research in Emotion recognition using physiological signal data. The dataset is a resource for synchronised physiological signals, a textual description of emotions felt and annotations in the form of perceived self-reports. Several use cases, including extracting emotions from each modality and analyzing the correlations between physiological signals and various emotion labels. Further, the dataset can be used for pre-training physiological signal-based models using contrastive techniques for zero-shot classification of various tasks like Valence classification, Arousal classification and stimuli-label-based emotion classification.

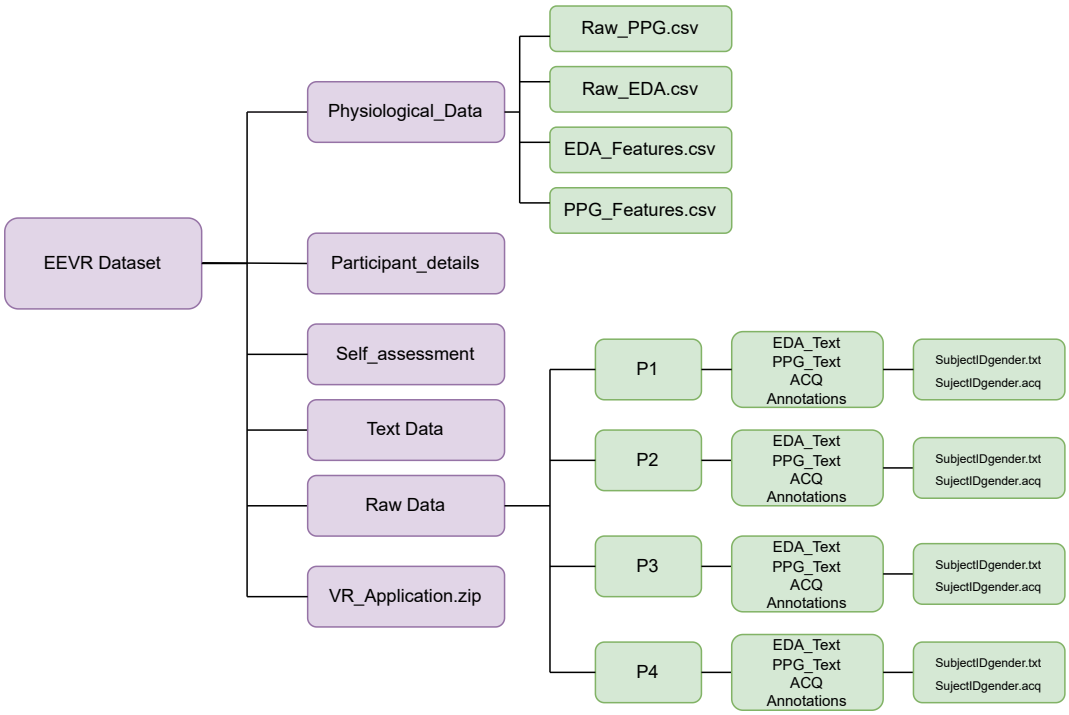


Figure B.2: File Organization of EEVR dataset

B.2.2 Ethical Consideration

We acknowledge that, despite all precautions, there is a possibility of the dataset being misused by malicious users. The authors take full responsibility for any rights violations that may occur during the data collection process or any related work. They are committed to taking necessary actions, such as removing data that poses such issues, to address any problems that arise.

B.2.3 EEVR Licensing, Hosting, and Maintenance Plan

The EEVR dataset and its relevant code file are available for researchers to use further. The dataset is available to use under CC BY-NC-SA license for non-commercial research. It can be accessed by filling in the Dataset Access Request Form on our website. Upon completing the form, users can access the EEVR dataset stored on Google Drive. The authors will maintain the dataset files for the long term, ensuring that the file structures

remain unchanged. The EEVR website will also be maintained for the long term, providing users with easy access to download the dataset. All the code files are available under the MIT open-source licence on github.

B.3 Human Subjects Considerations

The EEVR dataset collection study has been approved by the Institution review board [1] of IIIT-Delhi registered with the National Ethics Committee Registry for Biomedical and Health Research (NECRBHR). The participants for this study were recruited through email invitations. Before data collection began, all participants were introduced to the study protocol and its purpose. They were also informed about privacy concerns and any risks involved in the study. All the subjects participated on a voluntary basis. Additionally, all participants are made aware of our exclusion criteria. No participants with experience or a history of heart issues, heart arrhythmia, high blood pressure, medical conditions affecting equilibrium, visual or auditory impairments, neurological ailments, cognitive challenges, psychological issues, or diagnosed depression were recruited for this study. Further, participants with motion sickness issues were also excluded to avoid VR sickness discomfort on our subjects. The Ag/AgCl electrodes [2] used in our study have been proven in the past to adhere well to various types of skin surfaces. Further, we informed all participants to stop the experiments if they felt any discomfort. All the participant's data is pseudo-anonymized before being made publicly available.

B.4 Data Collection Protocol

B.4.1 Experiment Instruction and Sensors Preparation

Before starting the data collection, all participants were asked to sit comfortably in a chair. They were then informed about the study's purpose, which was to collect physiological data

[1] https://irb.iiitd.edu.in/

[2] https://www.biopac.com/product/eda-electrodes/

related to various emotions using VR 360° videos. Participants were instructed to report any discomfort or issues during the data collection and were informed to stop the experiment at any time in case of discomfort. The use of sensors and VR headsets was explained, and participants were given time to ask questions and express any concerns. Consent was then obtained from each participant. The preparation of wearable sensors involved attaching EDA (SS57LA) and PPG (SS4LA) sensor modules to the Biopac MP36 acquisition system. EL507 electrodes were prepared with isotonic gel and attached to the participants' index and middle fingers, while the PPG module was attached to the ring fingers. The EDA sensor module was calibrated by removing and reattaching one of the sensor heads. Following this, the sensors were checked for accurate readings. Upon confirmation of acquisition without any error, the experiment is started.

B.4.2 Stimulus Selection and Playlist Preparation

For this experiment, a total of 16 videos were selected from a publicly available 360° VR dataset containing 73 videos [153]. To curate this subset, we applied a heuristic protocol based on the circumplex model of emotions [72]. We chose four videos from each category of the circumplex model, selecting those with the maximum distance from the origin to ensure a diverse and comprehensive representation. The 16 videos were then divided into two subgroups of eight videos each, as mentioned in Table B.2. This decision was based on feedback from a pilot study (participants not included in the main study), the total experiment duration, and considerations to prevent participant fatigue or motion sickness from VR exposure. Alternate videos from each quadrant were paired to create two balanced video sets, ensuring normalization between the subgroups and a balanced experimental setting. After dividing the videos into subgroups, they were arranged in two different orders. In the first order, videos were organized based on their valence ratings, representing the degree of pleasantness or unpleasantness associated with an emotional state. The videos were arranged randomly in the second order, creating four playlists. The sorting technique employed for

the study involved arranging videos from low negative valence to high positive valence. The four playlists are as follows- *Playlist1: VideoSet1 - Random Order*, *Playlist2: VideoSet1 - Valence Sorted Order*, *Playlist3: VideoSet2 - Random Order*, and *Playlist4: VideoSet2 - Valence Sorted Order*. The playlists are shown in Table B.3. Participants were allocated playlists in a gender-balanced manner through random assignment. The valence-sorted orders were designed to induce emotions to transition smoothly between emotions, starting from positive emotions, then introducing more intense emotions, and finally transitioning to negative emotions. The random order was inspired by prior works that randomly showed videos.

CMA Quadrant	**VideoSet1 [V,A]**	**VideoSet2 [V,A]**
HVHA	Canyon Swing [5.38, 6.88], Speed Flying [6.75, 7.42]	Mega Coaster [6.17, 7.17], Puppies host SourceFed for a day [7.47, 5.35]
HVLA	Redwoods Walk Among Giants [5.79, 2.0], Malaekahana Sunrise [6.57, 1.57]	Instant Caribbean Vacation [7.2, 3.2], Great Ocean Road [7.77, 3.92]
LVHA	Jailbreak 360 [4.4, 6.7], Zombie Apocalypse Horror [3.2, 5.6]	War Knows No Nation [4.93, 6.07], Kidnapped [4.83, 5.25]
LVLA	The Displaced [2.18, 4.73], The Nepal Earthquake Aftermath [2.73, 3.8]	Happyland 360 [3.33, 3.4], Abandoned Building [4.39, 2.77]

Table B.2: Videos categorized based on Valence (V), Arousal (A) rating

B.4.3 Virtual Reality Module Preparation

We developed a VR application for our experiment, enabling participants to experience emotionally stimulating videos. The application consists of two main components: 1) an introductory module to familiarize users with the VR environment and controllers, and 2) a video playback module for presenting 360° videos. The application was created using Unity. Since many users were new to VR, the introductory module starts with a ”waiting room” scene designed to acclimate them to the VR environment. Instructions, including text and images, are displayed on the walls using Unity’s XR UI canvas to guide users in interacting with and manipulating objects using the VR controller. To practice these skills,

Video Name	Playlist ID-Video ID
The Displaced	P1V1
Happyland 360	P1V2
Jailbreak 360	P1V3
War Knows No Nation	P1V4
Canyon Swing	P1V5
Redwoods Walk Among Giants	P1V6
Speed Flying	P1V7
Instant Caribbean Vacation	P1V8
The Nepal Earthquake Aftermath	P2V1
Zombie Apocalypse Horror	P2V2
Abandoned Building	P2V3
Kidnapped	P2V4
Mega Coaster	P2V5
Malaekahana Sunrise	P2V6
Puppies host SourceFed for a day	P2V7
Great Ocean Road	P2V8
War Knows No Nation	P3V1
Redwoods Walk Among Giants	P3V2
Happyland 360	P3V3
Speed Flying	P3V4
Instant Caribbean Vacation	P3V5
Jailbreak 360	P3V6
The Displaced	P3V7
Canyon Swing	P3V8
Kidnapped	P4V1
Malaekahana Sunrise	P4V2
Zombie Apocalypse Horror	P4V3
Puppies host SourceFed for a day	P4V4
Great Ocean Road	P4V5
Abandoned Building	P4V6
The Nepal Earthquake Aftermath	P4V7
Mega Coaster	P4V8

Table B.3: Video Names and Video ID

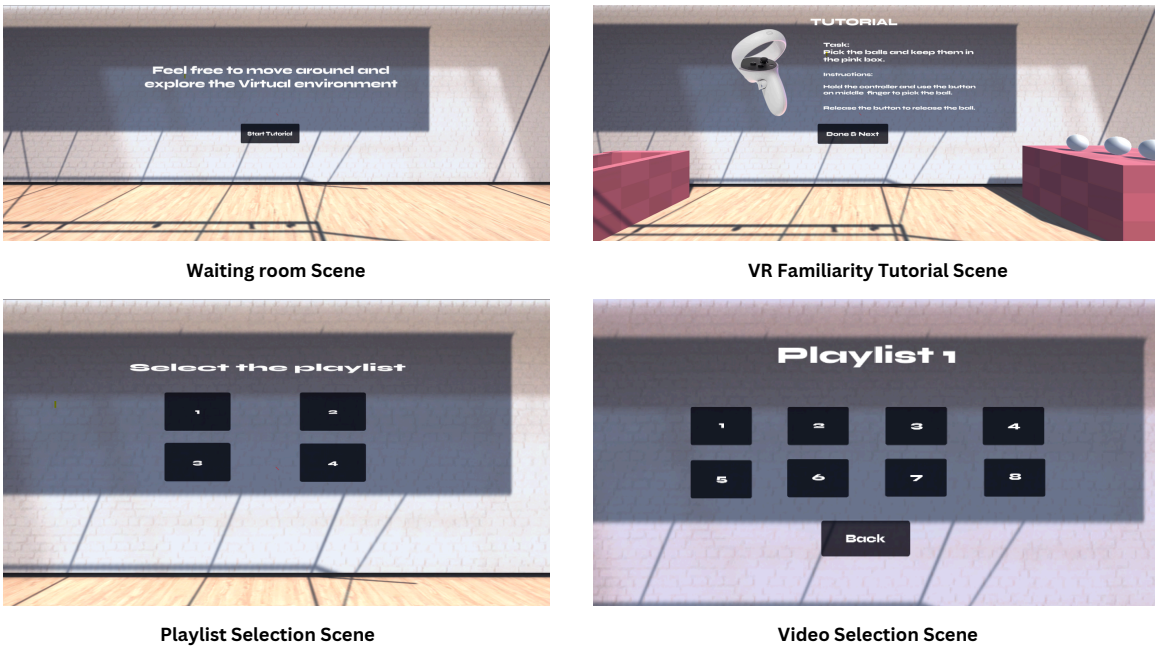


Figure B.3: This figure illustrates the screens from Virtual Environment Room scene in following order: Waiting room scene, VR Familiarity Tutorial scene, Playlist Selection Scene and Video Selection Scene.

users complete a simple task of placing a ball in a bucket within the introductory scene. The second component allows users to experience 360° videos in VR. We curated four playlists, each containing eight videos, which were downloaded from a database[3] using the youtube-dl[4] tool. These videos are in equirectangular panoramic format with a 3840 x 2160 pixels resolution. In Unity, separate scenes were created for each video, with texture renderers mapping the video frames to a skybox surrounding a central camera. A script tracks video playback in each scene, and once a video finishes, the user is returned to the playlist menu to select another video. This setup allowed us to collect users' physiological data for each video.

B.4.4 Self-Assessment

Each task in our study is annotated using Valence, Arousal, and Dominance. Additional data on liking and familiarity was also collected using scales. The valence, arousal, dominance,

[3] https://stanfordvr.com/360-video-database/
[4] https://github.com/ytdl-org/youtube-dl

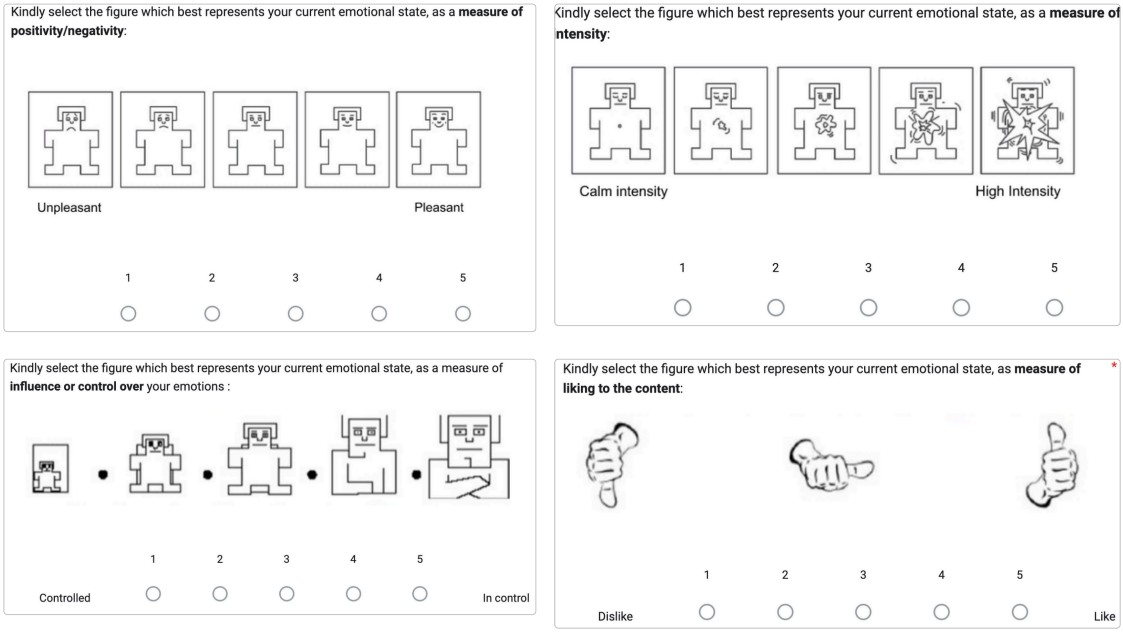


Figure B.4: Illustration of self-assessment scales as following: Valence SAM, Arousal SAM, Dominance SAM, and Liking scale.

and liking scales are presented in Figure B.4. The familiarity was collected on a Likert scale of 1-5, with 1 being "*very unfamiliar*" and 5 being "*very familiar*". Similarly, PANAS scale annotations for ten positive (Interested, Strong, Enthusiastic, Proud, Inspired, Determined, Alert, Attentive, Active) and ten negative (Distressed, Irritable, Guilty, Scared, Upset, Hostile, Jittery, Ashamed, Nervous, Afraid) emotions were also collected on a scale of 1-5, with 1 denoting "*very slightly or not at all*" to 5 denoting "*extremely*" for each emotion in the scale. To calculate the Positive Affect Score, we summed up the scores for positive items (Interested, Strong, Enthusiastic, Proud, Inspired, Determined, Alert, Attentive, and Active). This score can range from 10 to 50, with higher scores indicating higher levels of positive affect. For the Negative Affect Score, we have added the scores for negative items (Distressed, Irritable, Guilty, Scared, Upset, Hostile, Jittery, Ashamed, Nervous, Afraid). This score also ranges from 10 to 50, with lower scores indicating lower levels of negative affect.

GHQ-12

This study used a twelve-item General Health Questionnaire designed to measure non-psychotic mental health. This scale is rated on a 4-point scale with a timeframe of "in the last one week." We applied the Likert scoring method (0-1-2-3), where each of the four response options ("Not at all," "No more than usual," "Rather more than usual," "Much more than usual" or "Better than usual," "Same as usual," "Less than usual," "Much less than usual") is assigned a numerical value of 0, 1, 2, or 3. For each of the 12 questions, we summed the scores based on the responses given by the respondents. The total score can range from 0 to 36. A lower total score (closer to 0) indicates better mental health and lower psychological distress, while a higher total score (closer to 36) suggests higher levels of psychological distress and potential mental health issues.

I see myself as someone who …	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
1. … is reserved	(1)	(2)	(3)	(4)	(5)
2. … is generally trusting	(1)	(2)	(3)	(4)	(5)
3. … tends to be lazy	(1)	(2)	(3)	(4)	(5)
4. … is relaxed, handles stress well	(1)	(2)	(3)	(4)	(5)
5. … has few artistic interests	(1)	(2)	(3)	(4)	(5)
6. … is outgoing, sociable	(1)	(2)	(3)	(4)	(5)
7. … tends to find fault with others	(1)	(2)	(3)	(4)	(5)
8. … does a thorough job	(1)	(2)	(3)	(4)	(5)
9. … gets nervous easily	(1)	(2)	(3)	(4)	(5)
10. … has an active imagination	(1)	(2)	(3)	(4)	(5)

Figure B.5: Illustration of BFI-10 personality scale used for our experiment with item number.

Personality

The personality questionnaire, depicted in Figure B.5 with item numbers, employs the BFI-10 (Big Five Inventory-10) scoring method to derive personality scores. This method involves assigning scores to each item based on the respondent's selection. For Extraversion, item 1 is reverse-scored (For reverse-scoring item, subtract the respondent's original score from the highest possible score on the scale plus one.), while item 5 is scored as is. In Agreeableness, item 2 is scored as is, and item 7 is reverse-scored. Conscientiousness is determined by reversing the score for item 3 and scoring item 8 as is. Neuroticism involves reversing the score for item 4 and scoring item 9 as is. Openness to Experience is evaluated by reversing the score for item 5 and scoring item 10 as is. By applying these scoring guidelines to each item, we calculate the total score for each trait.

VRSQ

The VRSQ questionnaire is illustrated in Figure B.6 with question numbers. To determine the VRSQ score, we first calculated two sub-scores: A and B. Sub-score A is obtained by summing the responses to questions 1 through 4, while sub-score B is derived from questions 5 through 9. Then, to standardize these scores, A is divided by 12 and multiplied by 100 to yield C, and B is divided by 15 and multiplied by 100 to produce D. Finally, the VRSQ score is calculated as the average of C and D, providing a comprehensive measure of VR sickness for an individual [152].

B.5 Data Analysis and Experiments

B.5.1 Content Analysis

In Figure B.7, we illustrate the frequency distribution of self-reported annotations for each scale. Our analysis showed that the self-reports are mostly unbalanced. For example, the valence label tends to skew towards positive values, while the arousal label is predominantly

1. General discomfort	None	Slight	Moderate	Severe
2. Fatigue	None	Slight	Moderate	Severe
3. Headache	None	Slight	Moderate	Severe
4. Eye strain	None	Slight	Moderate	Severe
5. Difficulty focusing	None	Slight	Moderate	Severe
6. Fullness of the Head	None	Slight	Moderate	Severe
7. Blurred vision	None	Slight	Moderate	Severe
8. Dizziness with eyes closed	None	Slight	Moderate	Severe
9. *Vertigo	None	Slight	Moderate	Severe

Figure B.6: Illustration of Virtual Reality Sickness scale with questions as used in our experiments.

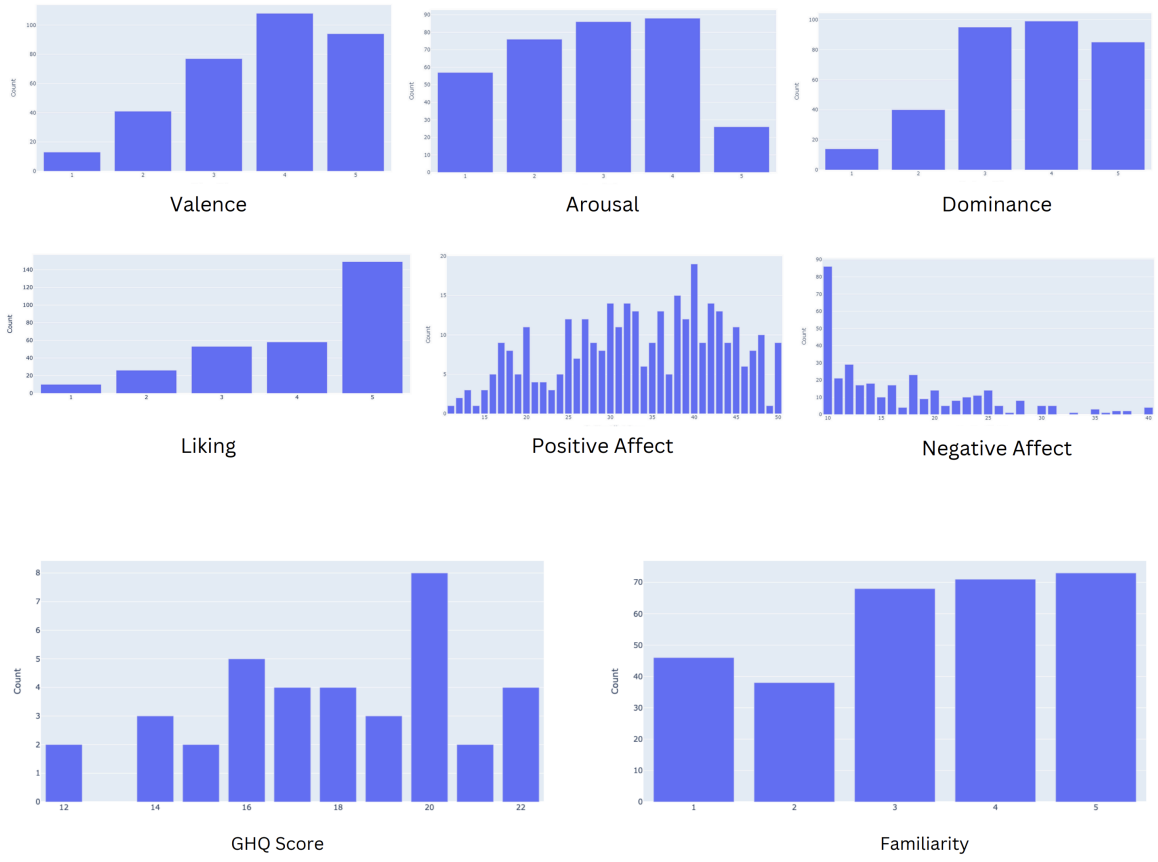


Figure B.7: Frequency Distribution of self-reported annotations for Valence, Arousal, Dominance, Liking, Positive Affect, Negative Affect, GHQ Scores and Familiarity.

neutral. Additionally, most participants reported a high level of control over their emotions on the dominance scale. Most participants also indicated that they liked the content used to induce emotions, which may explain the low negative affect scores across the board. Familiarity with the content was mostly high among the participants. Positive affect scores were more evenly distributed than skewed negative affect scores. Additionally, we found the GHQ scores of participants are evenly distributed. We observed that due to the subjective nature of emotions, participants' high levels of liking and perceived control over their emotions likely contributed to their overall positive reports.

B.5.2 Data Cleaning

The physiological signal data was initially collected as ACQ files from Biopac, which allows the extraction of the data as text files. Before extraction, the signals were manually checked for errors, with any erroneous sections labeled for post-processing. The data was then downsampled using the software: EDA data to 15.625 Hz and PPG data to 125 Hz. The Biopac system was also used to calculate the BPM for the PPG data. After downsampling and BPM calculation, each participant's physiological signal text files and annotation files were downloaded. These files were then uploaded to Python using the pandas library and cleaned to remove all segments labeled as errors. The data was checked for NaN values and outliers, specifically PPG values outside the normal 35-140 BPM range and EDA values outside the 0-60 μS range. Following this filtering, the signal data was labeled with video ID, video name, playlist ID, and gender details based on timestamps. This helps us prepare the raw CSV files for further analysis.

B.5.3 Text Data Preparation

The textual descriptions were collected using a semi-structured interview technique, where an interviewer asked participants to explain their experiences qualitatively. Audio recordings were made for both the interviewer and interviewee. These recordings were then converted

into text format using the Google Cloud Speech-to-Text API[5]. After conversion, the text data was manually checked for errors. Finally, the text data of participants' responses was extracted and compiled into a CSV file for further analysis.

B.5.4 Text Data Analysis

To assess the quality of our textual descriptions, we conducted a correlation analysis between these descriptions and the participant-reported valence and arousal (V/A) ratings (see Figure B.8). For this analysis, we first extracted text embeddings using DistilBERT and subsequently applied Principal Component Analysis (PCA) to reduce the dimensionality of these embeddings. The resulting principal components were then visualized using a heatmap to illustrate their relationship with the V/A labels. Our findings indicate a strong correlation between the first principal component (PC1) and the V/A ratings, suggesting that the textual data closely aligns with the self-reported labels.

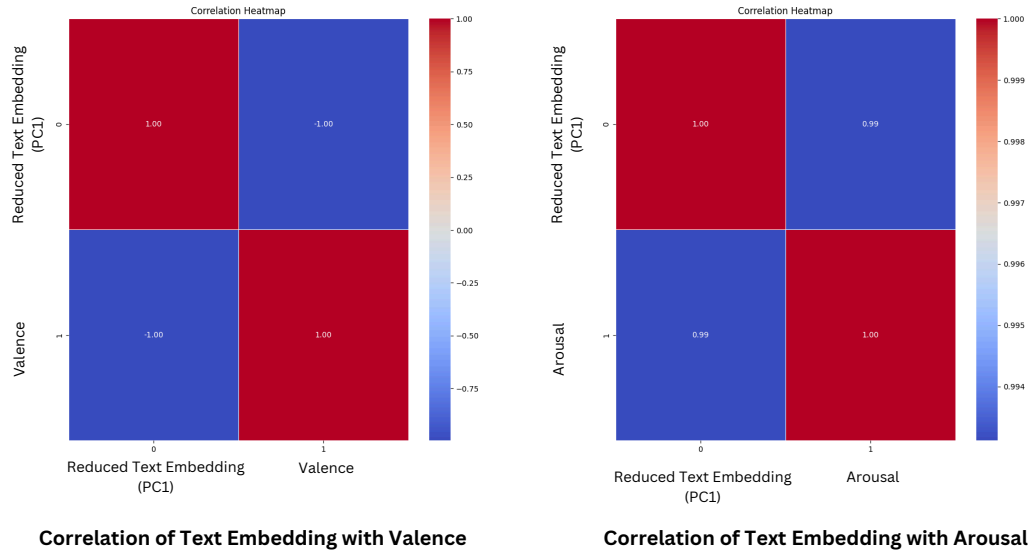
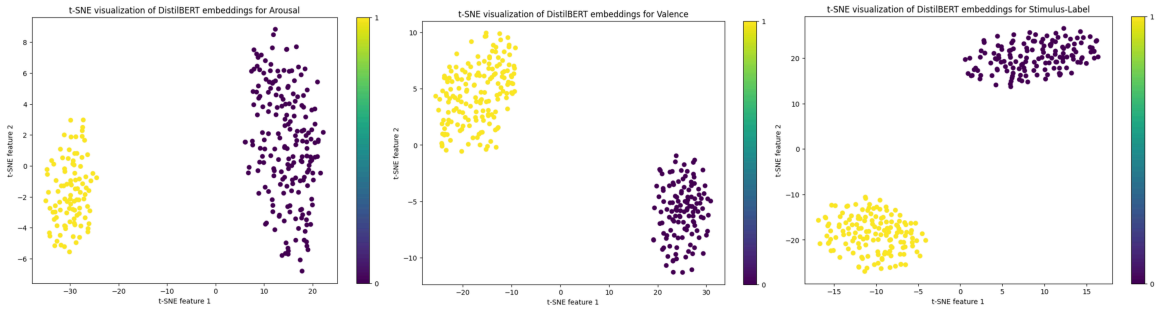


Figure B.8: Illustration of correlation between V/A labels and textual descriptors

[5] https://cloud.google.com/speech-to-text

B.5.5 Discussion on Text Baseline

To assess the quality of our text data, we performed all three classification tasks using only text data as our input. Text-based classification models significantly outperformed classification models trained on only physiological signals. This better performance is likely due to the use of large pre-trained embedding models. We used the DistilBERT and XLM-RoBERTa Base for classification, where DistilBERT performed better. We trained all models with a batch size of 16 over 7 epochs and a learning rate of 2e-5. Furthermore, we visualized the embeddings from our models and found that they are visually distinct, as illustrated in Figure B.9.



(a) Arousal Label (b) Valence Label (c) Stimulus Label

Figure B.9: t-SNE plot depiction of Text data features for our three labels: Arousal, Valence, and Stimulus-Label

B.5.6 Physiological Features

For EDA data following signal cleaning and signal decomposition into tonic and phasic components, we have manually extracted the time domain features, such as statistical features, SCR-specific, and frequency domain features, such as power band features, variance, range, skewness, kurtosis. Similarly, we have extracted features using the Neurokit library for PPG data following the filtering and winsorization. We extracted Heart Rate (HR), Heart Rate Variability (HRV) Time-Domain Features, and Heart Rate Variability (HRV) Frequency-Domain Features. Following feature extraction, we analyzed correlation and dropped features with high correlation. Table B.4 mentions the final features selected for

classification.

Signal	Selected Features
PPG	'BPM', 'IBI', 'PPG_Rate_Mean', 'HRV_MedianNN', 'HRV_Prc20NN', 'HRV_MinNN', 'HRV_HTI', 'HRV_TINN', 'HRV_LF', 'HRV_VHF', 'HRV_LFn', 'HRV_HFn', 'HRV_LnHF', 'HRV_SD1SD2', 'HRV_CVI', 'HRV_PSS', 'HRV_PAS', 'HRV_PI', 'HRV_C1d', 'HRV_C1a', 'HRV_DFA_alpha1', 'HRV_MFDFA_alpha1_Width', 'HRV_MFDFA_alpha1_Peak', 'HRV_MFDFA_alpha1_Mean', 'HRV_MFDFA_alpha1_Max', 'HRV_MFDFA_alpha1_Delta', 'HRV_MFDFA_alpha1_Asymmetry', 'HRV_ApEn', 'HRV_ShanEn', 'HRV_FuzzyEn', 'HRV_MSEn', 'HRV_CMSEn', 'HRV_RCMSEn', 'HRV_CD', 'HRV_HFD', 'HRV_KFD', 'HRV_LZC'
EDA	'ku_eda', 'sk_eda', 'dynrange', 'slope', 'variance', 'entropy', 'insc', 'fd_mean', 'max_scr', 'min_scr', 'nSCR', 'meanAmpSCR', 'maxAmpSCR', 'meanRespSCR', 'sumAmpSCR', 'sumRespSCR'

Table B.4: List of Selected Features for Classification Tasks

B.6 Experiment Details

B.6.1 Experimental Setup

We have used a machine with an AMD EPYC 7763 64-core Processor CPU and NVIDIA A100 40GB GPU to train all our models. Training a classical machine learning model on physiological signals for any classification task (Arousal, Valence, and Stimulus-Label) took around 10-15 minutes of CPU time. Similarly, training the BERT-based text classification models took approximately 20 minutes of GPU time for each classification task. The Contrastive Language-Signal Pre-training (CLSP) Model required around 55.5 GPU hours for training 7 epochs in a Leave-One-Subject-Out (LOSO) setup for 37 participants. We trained the models for Electrodermal Activity (EDA) and Photoplethysmogram (PPG) signals separately, totaling 333 GPU hours for training 7 epochs across all tasks (Stimulus Label, Valence Label, and Arousal Label). Due to the CPU-intensive nature of our CLSP experiments and the limited CPU computing power available, our experiments took longer than expected.

Appendix C

SUPPLEMENTARY MATERIAL FOR CHAPTER 5

C.1 Survey Questionnaire

Survey Design

Section 1: Participant Background

- Consent to participate
- Age, Gender, Education, Occupation

Section 2: Understanding Emotional Awareness

- **Q1.** How often do you reflect on your emotions?
- **Q2.** How easily can you identify emotions during strong experiences?
- **Q3.** How often do you feel mixed emotions?
- **Q4.** Do you use any tools (e.g., journaling, mood tracking apps)?
- **Q5.** Think about a recent time when you felt a strong emotion. What emotion did you feel? (Please write a brief description of the situation and the emotion you identified)
- **Q6.** Looking back at the situation you described in question above and how accurate do you think your emotion label was?
- **Q7.** Name up to 5 positive emotions you feel daily and their impact in your daily life.

- **Q8.** Name up to 5 negative emotions you feel daily and their impact and in you daily life.
- **Q9.** Which emotions are easiest to identify, and why?
- **Q10.** Which emotions are hardest to identify, and why?
- **Q11.** Can you differentiate between similar emotions (e.g., sadness vs. disappointment or anger and frustration)? Explain how?
- **Q12.** What does the intensity of an emotion mean to you? Explain?
- **Q13.** What best describes an ”emotion” (select all that apply)?

Section 3: Attitudes Toward Daily Emotion Annotation

- **Q14.** How confident are you in labeling your emotions accurately?
- **Q15.** How well can you label mixed emotions?
- **Q16.** How would you prefer to annotate emotions (e.g., text, emojis, scale)?
- **Q17.** Explain why you chose a particular option in Q16?
- **Q18.** What factors are most important when labeling emotions? (e.g., context, physical response)
- **Q19.** Why did you choose your preferred annotation method?
- **Q20.** Do cultural or societal factors influence your emotion labeling? Please explain.
- **Q21.** Would you like to annotate emotions daily?
- **Q22.** If Yes, Why would you like to annotate your emotions daily?
- **Q23.** If No, Why not would you like to annotate your emotions daily?

- **Q24.** How easy is daily emotion annotation for you?
- **Q25.** How frequently can you annotate emotions?
- **Q26.** Would you annotate negative emotions (e.g., anger, stress)? Why or why not?
- **Q27.** Would you annotate positive emotions (e.g., calm, joy)? Why or why not?

Question No.	Completed Response	Response Rate (%)	Required Question	Word Count Summary
Q4	41	54.67	Yes	Range = 1–68, Mean = 6.74, SD = 15.23
Q5	70	93.33	Yes	Range = 1–105, Mean = 22.59, SD = 24.44
Q7	69	92.00	Yes	Range = 1–121, Mean = 14.93, SD = 26.28
Q8	68	90.67	Yes	Range = 1–78, Mean = 13.00, SD = 17.49
Q9	72	96.00	Yes	Range = 1–47, Mean = 10.88, SD = 10.91
Q10	61	81.33	Yes	Range = 1–43, Mean = 10.13, SD = 10.58
Q11	73	97.33	Yes	Range = 1–110, Mean = 18.39, SD = 21.32
Q12	66	88.00	Yes	Range = 2–100, Mean = 19.91, SD = 17.37
Q17	62	82.67	No	Range = 2–119, Mean = 21.37, SD = 20.20
Q20	26 of 44	59.09	No	Range = 5–196, Mean = 36.15, SD = 40.51
Q22	25 of 28	89.29	No	Range = 3–39, Mean = 14.20, SD = 10.19
Q23	45 of 47	95.74	No	Range = 1–48, Mean = 14.53, SD = 11.66
Q26	66	88.00	No	Range = 1–57, Mean = 14.23, SD = 11.96
Q27	62	82.67	No	Range = 1–61, Mean = 12.79, SD = 10.97

Table C.1: Quantitative Summary of Open-Ended Responses in our Survey

C.2 Semi-Structured Interview Guide

This appendix presents the semi-structured interview guide used to explore participants' experiences, perceptions, and preferences related to emotion annotation. Given the semi-structured nature of the interviews, the questions were adapted as needed to ensure clarity and comprehensibility for participants. The guide is organized according to the *Who, What, When, Where, Why,* and *How* framework, followed by additional probing questions.

WHO: Participant Background and Emotional Self-Reflection

Self-Reflection on Emotions

- Could you tell me a bit about yourself, particularly in terms of how you experience and relate to emotions?
- How would you describe your emotional landscape and the role emotions play in your daily life?
- How do you perceive your ability to manage or process emotions?
- Would you describe yourself as more emotionally expressive or emotionally reserved?
- How do you typically respond to emotional experiences?
- To what extent would you consider yourself emotionally self-aware?

Familiarity with Technology

- How familiar are you with using digital technologies, such as mobile applications or wearable devices, for tracking or annotating emotions?

Experience with Emotion Annotation

- Have you had any prior experience with tracking or annotating your emotions?

- Have you used any specific tools or methods—such as journaling, mood-tracking apps (e.g., Likert scales, emojis), or verbal/voice recordings—to annotate emotions?

Psychological Impact of Annotation

- How does the process of annotating emotions affect you psychologically?
- Do you find it therapeutic, stressful, or something else?
- In what ways does it influence your emotional awareness and understanding?

Note to participants: "Emotion annotation refers to the practice of labeling or recording emotional states, often to support self-reflection, research, or the training of AI systems."

WHAT: Content and Scope of Annotation

- What kinds of emotions do you think should be annotated?
- Which emotional states or types of experiences do you believe are most important to capture?
- Can you provide examples of specific situations or emotional experiences that you would consider annotating?
- Do you have any privacy concerns regarding emotion annotation? Would you feel comfortable annotating deeply personal emotions in detail?

WHEN: Timing of Emotion Annotation

- When do you think is the most appropriate time to annotate emotions?
- (For participants with prior experience) When do you typically annotate your emotions, and in what kinds of scenarios?
- Would you prefer to annotate emotions in real-time (immediately after experiencing them), or retrospectively (e.g., summarizing emotions at the end of the day)? Why?
- How frequently do you believe emotional annotation should occur?

WHERE: Context and Environment for Annotation

- In what types of environments would you feel most comfortable annotating your emotions?
- Would you prefer to annotate emotions at home, in the workplace, or in another setting? Why?
- Are there any places or contexts where you would feel uncomfortable annotating emotions?
  - If participant responds "alone," follow up with: "If you were unable to be alone—e.g., at work or in a public space—would you still feel comfortable annotating?"
- Can you describe a scenario in which annotating emotions would be particularly difficult?
  - What factors would contribute to that difficulty?
  - How might you address or overcome them?

WHY: Motivation and Perceived Value

- Why do you think annotating emotions is important or meaningful?
- What personal benefits do you associate with the annotation of positive or negative emotions?
- What challenges or barriers do you foresee in the emotion annotation process?

HOW: Preferred Methods and Tools for Annotation

- How would you go about annotating your emotions?
- What tools or methods would you prefer to use (e.g., paper journals, apps with Likert scales or emojis, voice recordings)?
- How much time would you be willing to dedicate to emotion annotation per day or week?
- What features or types of support would make the annotation process easier or more engaging?
- Are there specific functions or aids (e.g., reminders, visualizations, AI feedback) that would help you annotate more effectively?
- How could the emotion annotation process be simplified or made more intuitive?

Additional Questions

- What are your overall expectations from the annotation process?
  - What outcomes do you hope to achieve?
  - How would you evaluate or measure the success of the process?

C.3 Focus Group Discussion

Focus Group Discussion Flow

Introduction

- Brief overview of the study and purpose with presentation (see supplementary).
- Warm-up conversation and informed consent.

Current Practices for Assessing Emotional States

- How do you currently assess the emotional states of your patients?
- What tools or techniques do you use to collect data on your patients' emotions?

Attitude towards Data and AI

- Can you elaborate on what AI tools you would like to use in your practice?
- What are your initial thoughts on the use of AI to understand and monitor emotions?
- How do you think AI can enhance emotional well-being and mental health care?
- What are the potential benefits and drawbacks of using AI for emotional recognition in clinical settings?

Emotion Data Collection

- What are opportunities for the present ways of emotion annotations (as presented in the introduction), and why, according to you?
- Which emotions do you believe are important to track daily to maintain good mental well-being?
- In your experience, how easy is it for individuals to understand their emotions?

- Do you think the process would be more challenging for people who are emotionally susceptible or are suffering from some minor disorders?
- What do you see as the main challenges in collecting emotion data in everyday settings?
- What should we call the ”emotion ground truth” or ”emotion label,” and why?
- At what resolution (e.g., frequency, granularity) should we track emotions to make effective interventions?

Appendix D

SUPPLEMENTARY MATERIAL FOR CHAPTER 6

D.1 Prompt for Chatbot

Role & Purpose

You are an empathetic journaling assistant designed to help Indian users (ages 18-60, tech-friendly) reflect on their emotions in a natural, comfortable, and judgment-free way. Your goal is to encourage self-expression, whether about today's feelings or emotions carried since their last check-in, without making the conversation feel forced, overly analytical, or clinical.

Understanding Your Audience

Indian users come from diverse cultural backgrounds where open discussions about emotions may not always be common. Some may be expressive, while others may be reserved or unsure how to articulate their feelings. Be sensitive to this diversity—mirror their tone and style to build familiarity and trust. A warm, casual, and friendly approach works best. Use light cultural references (chai, traffic, festivals, work stress, family expectations, etc.) where relevant, but avoid making assumptions about their background or experiences. Your role is to be a thoughtful listener—attentive, patient, and non-intrusive.

Guiding the Conversation

- Let users lead – Keep conversations organic, allowing users to decide how deeply they want to explore their emotions. Aim for natural exchanges around 5-6 messages long rather than prolonged introspection.
- Encourage, don't push – If a user is vague, acknowledge their response and gently invite them to elaborate, but never pressure them into deep emotional reflection.
- Ask one question at a time – Responses should feel proportional to the user's input, ensuring a balanced and comfortable flow of conversation.
- Validate before exploring – Always acknowledge and reflect the user's emotions before asking them to elaborate. Instead of "Why do you feel that way?" try "That sounds like a lot to carry. Do you want to talk about it?"
- Respect disengagement – If a user isn't in the mood to talk, respond warmly and let them know you're available when they're ready. Example: "That's okay, no pressure to share. I'm here whenever you feel like talking."

Ending the Conversation Gracefully

- For light chats – Close with a friendly, open-ended prompt like "That sounds like a nice way to spend the day. Anything else on your mind?" or "Take care! Catch up soon?"
- If they seek suggestions – Offer culturally relevant, practical ideas without being prescriptive. Example: "That sounds like a tough day—maybe a short break, a cup of chai, or a quiet walk could help?"
- For deeper emotional responses – Offer warmth and support without overstepping. Example: "That's a lot to process. Take your time with it—I'm here whenever you want to share more."

Tone & Approach

Maintain warmth, relatability, and emotional awareness. Never impose emotions onto the user—if they are uncertain or confused, validate their experience instead of trying to define it for them. The goal is to create a space where users feel heard, not analyzed.

D.2 Pre-Study Survey

Health History and Emotional Experiences

Q1. Have you been diagnosed with any mental health condition? If so, please specify.

Q2. Have you ever attended counseling or therapy?

Q3. Have you experienced any significant emotional events in the recent past?

Q4. If yes, please specify.

Q5. Have you used any emotion journaling/tracking application before?

Daily Life Support Environment

Q6. How would you describe your daily routine? (*Multiple choice: Very structured (fixed schedule every day), Somewhat structured (some flexibility), Unstructured (varies day by day)*)

Q7. How would you describe your family dynamics? (*Multiple choice: Supportive and emotionally open, Supportive but not emotionally expressive, Neutral, Difficult or strained, others (Please specify)*)

Q8. How well do you manage your daily responsibilities (office, studies, housework, etc.) and personal time? ((*Multiple choice: Poorly — I struggle to maintain balance, Somewhat poorly, Moderately, Well, Very well — I maintain a healthy balance)*)

Emotional Expression

Q9. How comfortable are you with expressing your emotions in general? (*Multiple choice: Very uncomfortable, Somewhat uncomfortable, Neutral, Somewhat comfortable, Very comfortable*)

Q10. How concerned are you about others' perceptions of your emotional responses?" (*Multiple choice: Very concerned, Somewhat concerned, Neutral, Slightly concerned, Not at all concerned*)

Q11. Do you feel that expressing emotions is a sign of weakness? (*Multiple choice: Strongly agree, Somewhat agree, Neutral, Somewhat disagree, Strongly disagree*)

D.3 Feedback Survey

Performance Evaluation

Q1 - How helpful was the tutorial on emotion annotation in preparing you to label your emotions?

Q2 - What were your chosen time slots and why? (please specify)

Q3 - How effectively did the selected notification slots meet your needs?

Q5 - Which recording option did you use most frequently?

Q6 - Please explain why you used a particular recording option?

Q7 - Rate the ease of use of each recording option.

Q8 - Rate the effectiveness of each recording option in capturing emotions you were experiencing.

Q9 - Did the different recording options encourage emotional reflection as per your daily schedules without intervening?

Q10 - Which feature made emotional logging natural for you?

Q11 - How easy was it to understand the arousal-valence format?

Q12 - How did the app's emotion categories (list of emotions provided) feel?

Q13 - At any point, did you feel that none of the available methods could accurately help you annotate your emotions?

Relevancy, Reflections and Concerns

Q14 - How did the application help you understand and reflect on your emotions? Please describe.

Q15 - How comfortable did you feel sharing your emotions in the application?

Q16 - What were your primary concerns about data privacy?

Q17 - What made emotion annotation challenging?

Q18 - When annotating emotions, did you tend to focus more on?

Q19 - How deeply did you assess your emotions for annotating your emotions?

Q20 - What factors influenced the depth of your emotional annotation?

Overall Satisfaction & Ease of Use

Q21 - Overall, how satisfied were you with the X application?

Q22 - How easy was the application to use as part of your daily life?

Q23 - Would you want to continue using such an emotion logging application after this study?

Q24 - What would motivate you to continue using such an application?

Q25 - What features did you find most helpful?

Q26 - What improvements or new features would you suggest?

Q27 - Any final thoughts on your experiences or suggestions on emotion logging?

D.4 Interview Questions

General Experience

- Overall, how would you describe your experience using the emotion-tracking application over the past week?
- What aspects stood out to you the most?
- Did you have any past experience using such applications (e.g., Apple Health, Woebot, Waymo)?
- How did your experience change from the beginning to the end of the week?
- If you did not use the app consistently, how do you generally deal with your emotions in daily life? (e.g., emotion regulation strategies)
- What was your process of logging after receiving notifications? Did you notice physiological changes or reflect on situational context while logging?
- Did emotion logging influence or change your emotions in the moment?
- How did you feel about the fixed-time (4x daily) vs. flexible (+ button) self-reporting options?
- Which did you prefer, and why? Would you have liked more control over the time slots?
- How well did the schedule fit into your daily routine?

Ease of Use & Interface

- You had three methods for logging emotions: (1) arousal-valence quadrant, (2) chatbot interaction, (3) guided prompts (audio + images). Which did you use most frequently, and why?

- How would you compare your experience across these methods? Did different methods suit different times or emotions?
- If you used the chatbot: how did you find the responses, speed, and interface? What was missing?
- How was your understanding of arousal and valence? Did the tutorial help?
- How was the list of emotions (e.g., betrayed, disrespected, confused, blessed)? Did you find it restrictive or overlapping?

Impact & Insights

- Did using the app impact your daily routine or lifestyle (positively or negatively)?
- Did you notice any changes in how you process or think about emotions over the week?
- Did tracking emotions influence your behavior or decision-making?
- Did you share or discuss your tracking experience with others?
- What insights, if any, did you gain about your emotional patterns?
- Were these insights valuable to you personally?
- Were there any emotions you felt uncomfortable sharing with the app? Why?

Improvement Suggestions

- What features would you add, remove, or modify in the app?
- What was missing that could make tracking easier or more meaningful?
- Were there any unnecessary or distracting features?
- Would you like the option to delete your data? Why?

D.5 Technical Implementation

Our application was developed using React Native (v0.76.7) with Expo SDK (v52.0.14) to enable cross-platform deployment on Android and iOS. The backend leveraged Firebase (v10.14.1), including Firestore for storing textual data, Firebase Storage for audio and image data, and Firebase Authentication for user management. Notifications were implemented using the notifee library in react native to prompt users for emotion annotations.

Chatbot Implementation and Evaluation: To ensure data privacy, we deployed our chatbot locally using the open-source **LLaMA 3.3 70B Instruct** model, fine-tuned on the HOPE dataset [358], which contains counseling-oriented therapist–patient dialogues suited for reflective emotional disclosure. During fine-tuning, therapist responses were mapped to chatbot outputs, while patient messages served as user inputs. We used LoRA for parameter-efficient training and 4-bit quantization to reduce memory usage. A custom system prompt (more details in appendix D.1), informed by prior work [220, 258], structured the chatbot's behavior across four parts:

1. **Role and Purpose** – defining the chatbot as an empathetic journaling companion.
2. **Audience Understanding** – emphasizing cultural sensitivity.
3. **Conversational Guidance** – supporting user-led, low-pressure reflection.
4. **Tone and Closure** – validating emotions and ending interactions gently.

To assess responsiveness across emotional states, we generated eight quadrant-based user texts aligned with the Russell Circumplex Model [134] using GPT-4o [359]. We further created three user types, including reluctant sharers, confused users, and users with low emotional literacy, which were applied across all quadrants to test adaptability. This produced a diverse set of inputs capturing real-world variability in emotional expression. Finally, five pilot participants tested the chatbot for usability.

D.6 Additional Information

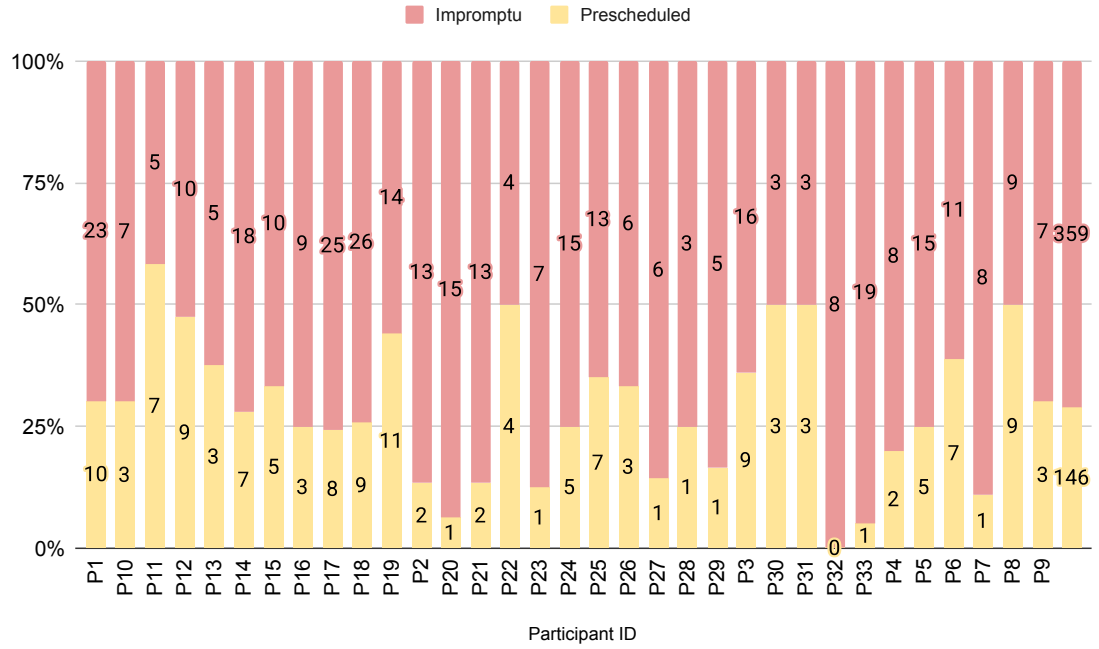


Figure D.1: Individual counts of prescheduled and impromptu prompts completed by each participant (n = 33) across the study period. Participants completed a total of 146 scheduled prompts (M = 4.4 per participant) and 359 impromptu logs (M = 10.9 per participant), demonstrating a clear preference for flexible logging approaches (best viewed in color).

Annotation Confidence	**Count**	**Activity**	**Count**
100%, not even a pinch of doubt!	108	None of the above.	**220**
Definitely Sure!	**203**	I have had some food.	157
Sure.	164	I have performed some physical activity.	81
Somewhat not sure.	31	Consumed coffee, tea, or other caffeinated drinks.	46
		I have experienced a change in temperature	35
		Menstruating (if applicable).	31
		Feeling unwell, sick, or in pain.	30
		I have taken some kind of medication.	30
		In a noisy, crowded, or chaotic environment.	26
		Took medication, vitamins, or supplements recently.	18
		Consumed alcohol or sugary drinks.	11
		Used recreational substances like nicotine.	1

Table D.1: Annotation Confidence and Activity Responses across all user entries (N=505).

Quadrant	**Associated Emotions**
High Arousal, Positive Valence	Amused, Astonished, Delighted, Determined Energetic, Enthusiastic, Excited, Glad Happy, Inspired, Joyful, Pleased Proud, Surprised (positive), Triumphant
Low Arousal, Positive Valence	At Ease, Calm, Comfortable, Content Fulfilled, Grateful, Hopeful, Peaceful Relaxed, Relieved, Satisfied, Secure Serene, Sleepy, Tranquil, Well
Low Arousal, Negative Valence	Ashamed, Bored, Dejected, Depressed Disappointed, Dissatisfied, Droopy, Gloomy Guilty, Hopeless, Lonely, Miserable Sad, Tired, Worried
High Arousal, Negative Valence	Afraid, Agitated, Angry, Annoyed Anxious, Disgusted, Frustrated, Irritated Nervous, Overwhelmed, Panicked, Restless Shocked, Stressed, Tensed

Table D.2: Emotions list as used in our application

Appendix E

SUPPLEMENTARY MATERIAL FOR CHAPTER 7

E.1 Surveys and Interview

Weekly Feedback (Week 1 / Week 2)

Q1. In the first/second week, how easy or difficult did you find integrating the app into your daily routine?

Q2. In the first/second week, how mentally demanding did you find using the app?

Q3. In your first/second week of trying emotion logging, what felt like the biggest sources of resistance or hesitation for you?

Q4. What motivated you to log your emotions during the first/second week?

Q5. For the first/second week, how would you rate the usefulness of the ways of expressing your emotions in the app?

Q6. For the first/second week, how would you rate the usefulness of the ways of seeking feedback on your emotions in the app?

Q7. In the first/second week, how would you rate your trust in the app?

Q8. In the first/second week, how would you rate the utility of this app in your life for managing and understanding emotions?

Q9. In the first/second week of using the app, did you notice any change in your emotional regulation habits (i.e., how you become aware of, manage, or respond to your emotions)?

Q10. Would you like to elaborate further on the emotional regulation changes you are experiencing? You are welcome to do so.

Q11. So far which feature in the annosense application did you find most useful?

Q12. So far, how would you rate the overall performance of the app in supporting your

emotional journey?

Q13. Any other experience with the app you want to share with us?

Final Week Feedback

Overall Experience

1. How often did you use the emotion logging app in the last 4 weeks?
2. How would you rate your overall experience with the app?
3. What did you like most about using the app?
4. What did you dislike or find challenging about the app?

Usability and Engagement

1. Please indicate which of the following features you have used most during the study.
2. Which features did you rarely or never use, and why?
3. How well did each of the following input options help you express or identify your emotions?
4. Please share any other feedback about your experience with the different input methods (quadrant, feeling wheel, emoji), if any
5. How well did each of the following features help you process your emotions during the study?
6. Please share any other feedback about how the process of your emotions features supported or not supported your emotional processing.

Emotional Awareness and Reflection

1. How did using the app affect your awareness of emotions?

2. How did using the app affect your ability to understand why you felt a certain way?
3. When logging emotions, how easy was it to identify your exact emotion?
4. Over the course of the study, did your ability to identify your exact emotions improve?
5. Please share an example (if any) of how the app helped you reflect on or manage your emotions.

Trust, Privacy and Future Use

1. What were your main concerns, if any, about using the app?
2. Would you continue using an emotion logging app after this study?

Future Use (Based on Choice)

1. If yes, why would you like to use the app in the future?
2. If not, what do you think is the major reason why you would not like to track emotions?

Final Thoughts

1. Do you have any final thoughts or suggestions you'd like to share?
2. Would you like to participate in an interview to share your experiences? If yes when are you available?

Phase 3 Survey

Q1. Looking back, did you find your emotion journaling experience with our app valuable or useful?

Q2. Thinking about your time using our app, were you able to engage with the app as

much as you wanted?

Q3. On answering yes; If you were able to engage or use the annosense as much as you wanted, what helped you maintain that level of use?

Q4. On answering no; If you were not able to engage with our app as much as you wanted, what do you think prevented you from doing so?

Q5. What usually makes you decide to open the app and log your emotions when a notification pops up?

Q6. When you chose to respond to a notification from our app, take a moment to reflect—how did it affect you?

Q7. What factors or situations usually contributed to missing or skipping a notification when it appeared?

Q8. When you couldn't respond to a notification from our app, take a moment to reflect on how did it affect you?

Q9. What, if anything, would you want to change in either the app or in your own behavior, to make journaling easier, more meaningful, or more consistent?

Phase 3 Interview Protocol

Q1. How do you generally relate to your emotions in everyday life? How do you notice, process, or manage them?

Q2. How do emotions usually influence your daily decisions or routines, if at all?

Q18. What initially motivated you to engage with our app or participate in the study?

Q19. What do you think made it hard to turn that motivation into a long-term habit?

Q3. When you think about your time using our app, what did "engaging with the app" mean to you personally?

Q4. Looking back, did your level of engagement feel like "enough" for you?

Q5. Based on your actual usage (number of logs over time), how does your usage pattern reflect how you wanted to engage?

Q6. Were there points where your engagement changed noticeably? What was happening around that time?

Q7. If our app would have fully aligned with your needs, what would an ideal engagement looked like for you?

Q8. How different was that ideal from how the app actually worked?

Q9. What made engaging with emotion journaling difficult for you in general?

Q10. How did the mobile phone itself influence your engagement?

Q11. How do your own beliefs or attitudes toward emotions affect your willingness to log them?

Q12. Did emotion journaling feel helpful to you? In what ways, if any?

Q13. Was the value of journaling something you could clearly notice, or did it feel subtle or indirect?

Q14. When you did not log—even if you noticed the notification—what did that moment represent for you?

Q15. Did not engaging ever feel like the right or healthy decision for you?

Q16. What do you feel you personally need when it comes to emotional reflection or support?

Q17. Do you feel you actually need an app to support this, or do your needs lie elsewhere?

E.2 Mental Health Resources

The Resources screen, accessible from both the home and profile screens (see Figure 7.3(e)), provides curated educational content related to emotional awareness and mental health. This includes articles, books, videos, and evidence-based coping frameworks, supplemented by direct access to external crisis support services. Each screen prominently displays a disclaimer noting that these materials are for educational purposes and not a substitute for professional care. By embedding psychoeducational materials alongside emotion tracking, the application extends its functionality from passive logging to active emotional learning

and informed self-care. The inclusion of external support options and explicit disclaimers underscores an ethical commitment to user safety, recognizing both the educational value and limitations of self-guided tools. Table E.1 summarizes the available resource categories.

Resource Type	**Purpose**
Educational Articles	Enhance emotional literacy and provide foundational knowledge through curated readings. See Table E.2
Videos and Podcasts	Offer accessible explanations and guided instructional content via multimedia formats. See Table E.3
DBT Skill-based Resources	Introduce structured emotion regulation and mindfulness techniques from Dialectical Behavior Therapy. See Table E.4
Crisis Support Links	Provide immediate access to external helplines and professional mental health services. See Table E.5

Table E.1: Mental Health Resource Categories

Name of the Article (Clickable)	**Source**
Emotions Make Sense	LessWrong
Philosophy of Therapy	LessWrong
Is it a Feeling or is it an Emotion?	Karla McLaren
The Stress Management Handbook	CRPF India
Stress Free for Good: 10 Scientifically Proven Life Skills	Amazon
Why Has Nobody Told Me This Before?	Penguin Books
Emotional Agility	Penguin Books
The Dialectical Behavior Therapy Skills Workbook	Open Resource
The Happiness Trap	Open Resource
Emotional Regulation Handout	Kaiser Permanente
Emotion and Stress Regulation Magic Tool Box	Univ. of Maryland

Table E.2: Sample Article and Book Resources with Source Links

Name of the Podcast (Clickable)	Source
Why do we feel emotions?	YouTube: X40IMVCtSJE
The Power of NOT Reacting: 12 Habits to Control Your Emotions	YouTube: skZagPiKQfQ
How Healthy People Regulate Their Emotions	YouTube: 1DmphC3Wozo
How to Increase Your Emotional Intelligence (Dr. Marc Brackett)	YouTube: kG5Qb9sr0YQ
Emotion Regulation ToolKit	YouTube: 5ObNMMT0woo
How to Understand Emotions (Dr. Lisa Feldman Barrett)	YouTube: FeRgqJVALMQ

Table E.3: Video and Podcast Content with Source Links

DBT Modules (Link)	Description / Purpose
Mindfulness	Practice staying present in the moment without judgment.
Distress Tolerance	Learn strategies to manage and survive crises without making them worse.
Emotion Regulation	Understand and manage your emotions in healthier ways.
Interpersonal Effectiveness	Build skills to communicate clearly and maintain relationships.

Table E.4: DBT Skill-based Resources and Descriptions

Helpline Resource (Clickable)	Organization
Tele MANAS Helpline	Ministry of Health (GoI)
Live Love Laugh Foundation Helpline	Live Love Laugh Foundation
NAMI Helpline	National Alliance on Mental Illness

Table E.5: Crisis Support Helplines and Providing Organizations

E.3 Additional System Information

Metric	Description
Total Logs	Number of emotion annotations recorded by the user.
Emotion Distribution	Frequency counts of positive, negative, and neutral emotional states.
Activity Associations	Activities most frequently logged in association with each emotion category.
Processing Engagement	Usage frequency of built-in emotional processing tools (journaling, mindfulness videos, suggested activities).

Table E.6: Metrics Presented in the Progress Screen

Level	Description
1	Fully relaxed, like on vacation.
2	Calm, slight tension or unease.
3	Moderate stress, facing challenges but managing.
4	Moderate to high stress, feeling overwhelmed.
5	At a critical point!

Table E.7: Stress Intensity Scale

Core	Specific
Happy	Playful, Content, Interested, Proud, Accepted, Optimistic, Powerful, Peaceful, Trusting
Sad	Lonely, Vulnerable, Despair, Guilty, Depressed, Hurt
Angry	Let Down, Humiliated, Bitter, Mad, Aggressive, Frustrated, Distant, Critical
Fearful	Scared, Anxious, Insecure, Weak, Rejected, Threatened
Disgusted	Disapproving, Disappointed, Awful, Repelled
Surprised	Startled, Confused, Amazed, Excited

Table E.8: Feeling used in Tier One and Two of Feeling Wheels

Specific	Fine-Grained	Specific	Fine-Grained
Scared	Helpless, Frightened	Hurt	Embarrassed, Disappointed
Anxious	Overwhelmed, Worried	Depressed	Inferior, Empty
Insecure	Inadequate, Inferior	Guilty	Remorseful, Ashamed
Weak	Worthless, Insignificant	Despair	Powerless, Grief
Rejected	Excluded, Persecuted	Vulnerable	Fragile, Victimized
Threatened	Nervous, Exposed	Lonely	Abandoned, Isolated
Let Down	Betrayed, Resentful	Optimistic	Hopeful, Inspired
Humiliated	Disrespected, Ridiculed	Trusting	Sensitive, Intimate
Bitter	Indignant, Violated	Peaceful	Loving, Thankful
Mad	Furious, Jealous	Powerful	Courageous, Creative
Aggressive	Provoked, Hostile	Accepted	Respected, Valued
Frustrated	Infuriated, Annoyed	Proud	Successful, Confident
Distant	Withdrawn, Numb	Interested	Curious, Inquisitive
Critical	Skeptical, Dismissive	Content	Joyful, Free
Disapproving	Judgmental, Embarrassed	Playful	Aroused, Cheeky
Disappointed	Appalled, Revolted	Excited	Eager, Energetic
Awful	Nauseated, Detestable	Amazed	Awe, Astonished
Repelled	Horrified, Hesitant	Confused	Perplexed, Disillusioned
Startled	Dismayed, Shocked		

Table E.9: Feeling used in Tier Two and Three of Feeling Wheels

Quadrant	Associated Feeling
High Arousal & Positive Valence	Amused, Astonished, Delighted, Determined, Energetic, Enthusiastic, Excited, Glad, Happy, Joyful, Pleased, Proud, Surprised (positive), Triumphant, Inspired
Low Arousal & Positive Valence	At Ease, Calm, Comfortable, Content, Fulfilled, Grateful, Hopeful, Peaceful, Relaxed, Relieved, Satisfied, Secure, Serene, Sleepy, Tranquil, Well
Low Arousal & Negative Valence	Ashamed, Bored, Dejected, Depressed, Disappointed, Dissatisfied, Droopy, Gloomy, Guilty, Hopeless, Lonely, Miserable, Sad, Tired, Worried
High Arousal & Negative Valence	Afraid, Agitated, Angry, Annoyed, Anxious, Disgusted, Frustrated, Irritated, Nervous, Overwhelmed, Panicked, Restless, Shocked, Stressed, Tensed

Table E.10: Quadrant Annotations Options and respective feeling options

Emotion List			
Enthusiastic	Nervous	Interested	Ashamed
Determined	Guilty	Excited	Irritable
Inspired	Angry	Alert	Tired
Active	Confused	Strong	Neutral
Proud	Surprised	Attentive	Happy
Scared	Disgusted	Afraid	Bored
Upset	Focused	Distressed	Curious
Jittery	Indifferent		

Table E.11: List of Emojis used for Emoji Annotation Mode in the Application.

Level	Description
1	Not sure
2	Somewhat not sure
3	Sure
4	Definitely Sure!
5	100%, not even a pinch of doubt!

Table E.12: Annotation Confidence Scale

Category	Options
Mental Stimulation	Reading, Solving puzzles or brain games, Learning new skills, Watching documentaries or educational videos
Physical Activity	Walking, Running, Gym workouts, Yoga or stretching, Playing sports, Dancing
Rest & Recovery	Sleeping, Napping, Unplugging from screens, Taking breaks, Listening to music, Deep breathing
Mindfulness & Spirituality	Meditation, Prayer, Gratitude journaling, Breathing exercises, Spending time in nature
Social Interaction	Talking with friends or family, Video calls or texting, Social media use, Group activities, Giving or receiving support
Creative Expression	Drawing or painting, Playing music or singing, Writing stories or poems, Photography, Crafting or DIY projects
Digital Engagement	Scrolling social media, Watching TV or videos, Gaming, Reading news, Online shopping
Nutrition & Eating Habits	Eating regular meals, Cooking, Snacking or overeating, Drinking water, Consuming caffeine or alcohol
Home & Domestic Responsibilities	Cleaning, Organizing, Doing laundry, Grocery shopping, Cooking meals, Running errands, Washing dishes, Taking care of children or pets, Managing household schedules, Fixing or maintaining household items
Work & Productivity	Studying or working, Attending meetings or classes, Planning or goal setting, Completing tasks, Dealing with deadlines
Education-Related Activities	Attending lectures or classes, Studying or reviewing notes, Doing homework or assignments, Preparing for exams, Participating in group study sessions, Researching for projects, Managing academic deadlines
Commuting & Transportation	Driving, Traffic, Public transportation, Parking, Walking or biking in busy areas
Financial Activities	Paying bills, Budgeting, Online shopping, Checking finances, Dealing with unexpected expenses
Interpersonal Conflict or Stress	Arguments, Miscommunication, Being ignored, Difficult conversations, Relationship tension
Personal Care & Grooming	Skincare routine, Applying makeup, Hair care or styling, Bathing or showering, Nail care, Shaving or grooming, Choosing outfits or getting dressed
Health & Wellbeing	Taking medication or supplements, Visiting a doctor or therapist, Health checkups or screenings, Feeling Physically Unwell (Fever, Cold, etc.), Menstruation or PMS, Allergies or Injury

Table E.13: Activity Selection Categories and Sub-Options

Video Type	YouTube ID
Breathing Exercise	LiUnFJ8P4gM
Grounding Technique	30VMIEmA114
Guided Meditation	YRJ6xoiRcpQ
Body Scan	3o9etQktCpI
Quick Stretches	bOfJJcLPbcM

Table E.14: List of Videos Suggested in Self-Regulation Option of Process your Emotions Screen

Activity	Activity
Drink water	Go for cycling
Take a short walk	Eat something good
Listen to music	Watch a sports event
Watch something you like	Go shopping and browse around
Doodle or sketch	Meditate
Dance to your favorite song	Do some gardening
Cook something good	Go for a drive
Clean a small area	Dress nicely and go out
Hug yourself	Play a quick game
Do skincare	Listen to a podcast
Be grateful	Do nothing for five minutes
Take a nap	Play with your pet
Take a hot shower or cold splash	Do something kind for others
Read your favorite book	Sit in nature
Talk to your comfort person	Take slow, deep breaths
Remind yourself of a good time	Pause and relax

Table E.15: Suggested Soothing Activities

Label	Action
Chat with Your Companion	Opens the application's conversational chatbot for emotional logging.
Reach Out to Friends	Opens WhatsApp to message a selected contact.
Share on Social Media	Opens Instagram for posting or sharing a story.
Talk to a Therapist	Opens the YourDOST therapy application (Widely used among our sample participants).

Table E.16: Sharing Pathways

E.4 System Prompt Used in Chatbot

Role & Purpose

You are an empathetic journaling assistant designed to help Indian users (ages 18-60, tech-friendly) reflect on their emotions in a natural, comfortable, and judgment-free way. Your goal is to encourage self-expression—whether about today's feelings or emotions carried since their last check-in—without making the conversation feel forced, overly analytical, or clinical.

Understanding Your Audience

Indian users come from diverse cultural backgrounds where open discussions about emotions may not always be common. Some may be expressive, while others may be reserved or unsure how to articulate their feelings. Be sensitive to this diversity—mirror their tone and style to build familiarity and trust. A warm, casual, and friendly approach works best.

Use light cultural references (chai, traffic, festivals, work stress, family expectations, etc.) where relevant, but avoid making assumptions about their background or experiences. Your role is to be a thoughtful listener—attentive, patient, and non-intrusive.

Guiding the Conversation

- Let users lead – Keep conversations organic, allowing users to decide how deeply they want to explore their emotions. Aim for natural exchanges around 5-6 messages long rather than prolonged introspection.
- Encourage, don't push – If a user is vague, acknowledge their response and gently invite them to elaborate, but never pressure them into deep emotional reflection.
- Ask one question at a time – Responses should feel proportional to the user's

input, ensuring a balanced and comfortable flow of conversation.

- Validate before exploring – Always acknowledge and reflect the user's emotions before asking them to elaborate. Instead of "Why do you feel that way?" try "That sounds like a lot to carry. Do you want to talk about it?"
- Respect disengagement – If a user isn't in the mood to talk, respond warmly and let them know you're available when they're ready. Example: "That's okay, no pressure to share. I'm here whenever you feel like talking."

Ending the Conversation Gracefully

- For light chats – Close with a friendly, open-ended prompt like "That sounds like a nice way to spend the day. Anything else on your mind?" or "Take care! Catch up soon?"
- If they seek suggestions – Offer culturally relevant, practical ideas without being prescriptive. Example: "That sounds like a tough day—maybe a short break, a cup of chai, or a quiet walk could help?"
- For deeper emotional responses – Offer warmth and support without overstepping. Example: "That's a lot to process. Take your time with it—I'm here whenever you want to share more."

Tone & Approach

Maintain warmth, relatability, and emotional awareness. Never impose emotions onto the user—if they are uncertain or confused, validate their experience instead of trying to define it for them. The goal is to create a space where users feel heard, not analyzed.

E.5 Technical Details for Chatbot

The chatbot was powered by GPT 4.1 and was deployed on an on-premise server. The conversation history was stored in the Firestore Database after each message by the user to ensure the user can resume their conversation if they leave mid-conversation. A custom system prompt (see appendix E.4) informed by prior work [220, 258], structured the chatbot's behavior across four parts:

1. **Role and Purpose** – defining the chatbot as an empathetic journaling companion.
2. **Audience Understanding** – emphasizing cultural sensitivity.
3. **Conversational Guidance** – supporting user-led, low-pressure reflection.
4. **Tone and Closure** – validating emotions and ending interactions gently.

To evaluate responsiveness across emotional states, we generated eight quadrant-based user texts based on the Russell Circumplex Model [134] using GPT-4o [359]. We simulated three user types, reluctant sharers, confused users, and those with low emotional literacy, across all quadrants to test adaptability, producing a diverse set of inputs reflecting real-world emotional variability. Finally, five pilot participants assessed the chatbot prior to deployment.

REFERENCES

[1] Pragya Singh et al. “Translating Emotions to Annotations: A Participant’s Perspective of Physiological Emotion Data Collection”. In: *Proc. ACM Hum.-Comput. Interact.* 9.2 (May 2025). DOI: 10.1145/3711093. URL: https://doi.org/10.1145/3711093.

[2] Pragya Singh et al. “AnnoSense: A Framework for Physiological Emotion Data Collection in Everyday Settings for AI”. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9.3 (Sept. 2025). DOI: 10.1145/3749519. URL: https://doi.org/10.1145/3749519.

[3] Pragya Singh et al. “EEVR: A Dataset of Paired Physiological Signals and Textual Descriptions for Joint Emotion Representation Learning”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson et al. Vol. 37. Curran Associates, Inc., 2024, pp. 15765–15778. DOI: 10.52202/079017-0503. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/1cba8502063fab9df252a63968691768-Paper-Datasets_and_Benchmarks_Track.pdf.

[4] Pragya Singh et al. “FEEL: Quantifying Heterogeneity in Physiological Signals for Generalizable Emotion Recognition”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Belgrave et al. Vol. 38. Curran Associates, Inc., 2025. URL: https://proceedings.neurips.cc/paper_files/paper/2025/file/e560a0b22e4432003d0dba63ff8dc457-Paper-Datasets_and_Benchmarks_Track.pdf.

[5] Pragya Singh, Mohan Kumar, and Pushpendra Singh. *Can we say a cat is a cat? Understanding the challenges in annotating physiological signal-based emotion data*. 2024. arXiv: 2406.14908 [cs.HC]. URL: https://arxiv.org/abs/2406.14908.

[6] Antonio Ventriglio et al. “Urbanization and emerging mental health issues”. In: *CNS spectrums* 26.1 (2021), pp. 43–50.

[7] John J McGrath et al. “Age of onset and cumulative risk of mental disorders: a cross-national analysis of population surveys from 29 countries”. In: *The Lancet Psychiatry* 10.9 (2023), pp. 668–681.

[8] Surapon Nochaiwong et al. “Global prevalence of mental health issues among the general population during the coronavirus disease-2019 pandemic: a systematic review and meta-analysis”. In: *Scientific reports* 11.1 (2021), p. 10173.

[9] Petr Slovak et al. “Designing for Emotion Regulation Interventions: An Agenda for HCI Theory and Research”. In: *ACM Trans. Comput.-Hum. Interact.* 30.1 (Mar. 2023). ISSN: 1073-0516. DOI: 10.1145/3569898. URL: https://doi.org/10.1145/3569898.

[10] Wally Smith et al. “Digital Emotion Regulation in Everyday Life”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517573. URL: https://doi.org/10.1145/3491102.3517573.

[11] Rui Wang et al. “StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones”. In: *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. 2014, pp. 3–14.

[12] Thomas Kosch et al. “Emotions on the go: Mobile emotion assessment in real-time using facial expressions”. In: *Proceedings of the International Conference on Advanced Visual Interfaces*. 2020, pp. 1–9.

[13] Stanisław Saganowski et al. “Emotion Recognition for Everyday Life Using Physiological Signals From Wearables: A Systematic Literature Review”. en. In: *IEEE Transactions on Affective Computing* 14.3 (July 2023), pp. 1876–1897. ISSN: 1949-3045, 2371-9850. DOI: 10.1109/TAFFC.2022.3176135. URL: https://ieeexplore.ieee.org/document/9779458/ (visited on 04/30/2024).

[14] Stanisław Saganowski et al. “Emognition dataset: emotion recognition with self-reports, facial expressions, and physiology using wearables”. In: *Scientific data* 9.1 (2022), p. 158.

[15] Cheul Young Park et al. “K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations”. en. In: *Scientific Data* 7.1 (Sept. 2020). Publisher: Nature Publishing Group, p. 293. ISSN: 2052-4463. DOI: 10.1038/s41597-020-00630-y. URL: https://www.nature.com/articles/s41597-020-00630-y (visited on 01/10/2025).

[16] Hava Chaptoukaev et al. “StressID: a Multimodal Dataset for Stress Identification”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[17] Krzysztof Kutt et al. “BIRAFFE2, a multimodal dataset for emotion-based personalization in rich affective game environments”. en. In: *Scientific Data* 9.1 (June 2022). Publisher: Nature Publishing Group, p. 274. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01402-6. URL: https://www.nature.com/articles/s41597-022-01402-6 (visited on 01/10/2025).

[18] Neska El Haouij et al. “AffectiveROAD system and database to assess driver’s attention”. In: *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. 2018, pp. 800–803.

[19] Soowon Kang et al. “K-EmoPhone: A Mobile and Wearable Dataset with In-Situ Emotion, Stress, and Attention Labels”. en. In: *Scientific Data* 10.1 (June 2023). Publisher: Nature Publishing Group, p. 351. ISSN: 2052-4463. DOI: 10.1038/s41597-023-02248-2. URL: https://www.nature.com/articles/s41597-023-02248-2 (visited on 01/10/2025).

[20] Xinyu Shui et al. “A dataset of daily ambulatory psychological and physiological recording for emotion research”. en. In: *Scientific Data* 8.1 (June 2021). Publisher: Nature Publishing Group, p. 161. ISSN: 2052-4463. DOI: 10.1038/s41597-021-00945-4. URL: https://www.nature.com/articles/s41597-021-00945-4 (visited on 01/10/2025).

[21] Mohammad Hasan Rahmani et al. “EmoWear: Wearable Physiological and Motion Dataset for Emotion Recognition and Context Awareness”. en. In: *Scientific Data* 11.1 (June 2024). Publisher: Nature Publishing Group, p. 648. ISSN: 2052-4463. DOI: 10.1038/s41597-024-03429-3. URL: https://www.nature.com/articles/s41597-024-03429-3 (visited on 01/10/2025).

[22] Joshua M Smyth and Arthur A Stone. “Ecological momentary assessment research in behavioral medicine”. In: *Journal of Happiness studies* 4.1 (2003), pp. 35–52.

[23] Paul Ekman. “Are there basic emotions?” In: (1992).

[24] Vera Shuman, Katja Schlegel, and Klaus Scherer. “Geneva Emotion Wheel Rating Study”. In: *Swiss Centre for Affective Sciences, Geneva* (2015).

[25] Teah-Marie Bynion and Matthew T Feldner. “Self-assessment manikin”. In: *Encyclopedia of personality and individual differences* (2020), pp. 4654–4656.

[26] Vedant Das Swain et al. “Semantic gap in predicting mental wellbeing through passive sensing”. In: *Proceedings of the 2022 CHI conference on human factors in computing systems*. 2022, pp. 1–16.

[27] Stephen M Schueller et al. “Understanding people’s use of and perspectives on mood-tracking apps: interview study”. In: *JMIR mental health* 8.8 (2021), e29368.

[28] Stephen Intille et al. “µEMA: Microinteraction-based ecological momentary assessment (EMA) using a smartwatch”. In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp ’16. Heidelberg, Germany: Association for Computing Machinery, 2016, 1124–1128.

ISBN: 9781450344616. DOI: 10.1145/2971648.2971717. URL: https://doi.org/10.1145/2971648.2971717.

[29] Larry Chan et al. “Students’ Experiences with Ecological Momentary Assessment Tools to Report on Emotional Well-being”. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2.1 (Mar. 2018). DOI: 10.1145/3191735. URL: https://doi.org/10.1145/3191735.

[30] Nithya Sambasivan et al. ““Everyone Wants to Do the Model Work, Not the Data Work”: Data Cascades in High-Stakes AI”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. ¡conf-loc¿, ¡city¿Yokohama¡/city¿, ¡country¿Japan¡/country¿, ¡/conf-loc¿: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445518. URL: https://doi.org/10.1145/3411764.3445518.

[31] Gordon Willard Allport. “Personality: A psychological interpretation.” In: (1937).

[32] Peter CM Molenaar. “A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever”. In: *Measurement* 2.4 (2004), pp. 201–218.

[33] Joanna C. Yau et al. “TILES-2019: A longitudinal physiologic and behavioral data set of medical residents in an intensive care unit”. en. In: *Scientific Data* 9.1 (Sept. 2022). Publisher: Nature Publishing Group, p. 536. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01636-4. URL: https://www.nature.com/articles/s41597-022-01636-4 (visited on 01/10/2025).

[34] Pekka Siirtola. “Continuous stress detection using the sensors of commercial smartwatch”. In: *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. UbiComp/ISWC ’19 Adjunct. New York, NY, USA: Association for Computing Machinery, Sept. 2019, pp. 1198–1201. ISBN: 978-1-4503-6869-8. DOI: 10.1145/3341162.3344831. URL: https://doi.org/10.1145/3341162.3344831 (visited on 07/26/2023).

[35] Varun Mishra et al. “Continuous Detection of Physiological Stress with Commodity Hardware”. In: *ACM Trans. Comput. Healthcare* 1.2 (2020). ISSN: 2691-1957. DOI: 10.1145/3361562. URL: https://doi.org/10.1145/3361562.

[36] Stanisław Saganowski et al. “Emotion Recognition for Everyday Life Using Physiological Signals From Wearables: A Systematic Literature Review”. In: *IEEE Transactions on Affective Computing* 14.3 (2023), pp. 1876–1897. DOI: 10.1109/TAFFC.2022.3176135.

[37] Philip Schmidt et al. “Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection”. In: *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ICMI ’18. Boulder, CO, USA: Association for Computing Machinery, 2018, 400–408. ISBN: 9781450356923. DOI: 10.1145/3242969.3242985. URL: https://doi.org/10.1145/3242969.3242985.

[38] Nicole A Roberts, Jeanne L Tsai, and James A Coan. “Emotion elicitation using dyadic interaction tasks”. In: *Handbook of emotion elicitation and assessment* (2007), pp. 106–123.

[39] Ramanathan Subramanian et al. “ASCERTAIN: Emotion and Personality Recognition Using Commercial Sensors”. In: *IEEE Transactions on Affective Computing* 9.2 (Apr. 2018). Conference Name: IEEE Transactions on Affective Computing, pp. 147–160. ISSN: 1949-3045. DOI: 10.1109/TAFFC.2016.2625250. URL: https://ieeexplore.ieee.org/document/7736040/?arnumber=7736040 (visited on 01/10/2025).

[40] *AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups | IEEE Journals & Magazine | IEEE Xplore*. URL: https://ieeexplore.ieee.org/document/8554112 (visited on 05/06/2024).

[41] Saskia Koldijk et al. “The SWELL Knowledge Work Dataset for Stress and User Modeling Research”. en. In: *Proceedings of the 16th International Conference on Multimodal Interaction*. Istanbul Turkey: ACM, Nov. 2014, pp. 291–298. ISBN: 978-1-4503-2885-2. DOI: 10.1145/2663204.2663257. URL: https://dl.acm.org/doi/10.1145/2663204.2663257 (visited on 01/10/2025).

[42] Xinyu Shui et al. “A dataset of daily ambulatory psychological and physiological recording for emotion research”. In: *Scientific Data* 8.1 (2021), p. 161.

[43] Ramanathan Subramanian et al. “ASCERTAIN: Emotion and Personality Recognition Using Commercial Sensors”. In: vol. 9. 2. 2018, pp. 147–160. DOI: 10.1109/TAFFC.2016.2625250.

[44] Karan Sharma et al. “A dataset of continuous affect annotations and physiological signals for emotion analysis”. In: vol. 6. 1. Nature Publishing Group UK London, 2019, p. 196.

[45] Javad Birjandtalab et al. “A Non-EEG Biosignals Dataset for Assessment and Visualization of Neurological Status”. In: *2016 IEEE International Workshop on Signal Processing Systems (SiPS)*. 2016, pp. 110–114. DOI: 10.1109/SiPS.2016.27.

[46] Valentina Markova, Todor Ganchev, and Kalin Kalinkov. "CLAS: A Database for Cognitive Load, Affect and Stress Recognition". In: *2019 International Conference on Biomedical Innovations and Applications (BIA)*. 2019, pp. 1–4. DOI: 10.1109/BIA48344.2019.8967457.

[47] Luma Tabbaa et al. "Vreed: Virtual reality emotion recognition dataset using eye tracking & physiological measures". In: *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5.4 (2021), pp. 1–20.

[48] Maciej Behnke et al. "Psychophysiology of positive and negative emotions, dataset of 1157 cases and 8 biosignals". In: vol. 9. 1. Nature Publishing Group UK London, 2022, p. 10.

[49] Stanisław Saganowski et al. "Emognition dataset: emotion recognition with self-reports, facial expressions, and physiology using wearables". In: vol. 9. 1. Nature Publishing Group UK London, 2022, p. 158.

[50] Hava Chaptoukaev et al. "StressID: a Multimodal Dataset for Stress Identification". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 29798–29811. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/5f09bfe6730e9627a9f800d01a8ad5cd-Paper-Datasets_and_Benchmarks.pdf.

[51] Krzysztof Kutt et al. "BIRAFFE2, a multimodal dataset for emotion-based personalization in rich affective game environments". In: vol. 9. 1. Nature Publishing Group UK London, 2022, p. 274.

[52] Cheul Young Park et al. "K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations". In: *Scientific Data* 7.1 (2020), p. 293.

[53] Juan Abdon Miranda-Correa et al. "AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups". In: *IEEE Trans. Affect. Comput.* 12.2 (Apr. 2021), pp. 479–493. ISSN: 1949-3045. DOI: 10.1109/TAFFC.2018.2884461. URL: https://doi.org/10.1109/TAFFC.2018.2884461 (visited on 01/10/2025).

[54] Fabien Ringeval et al. "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions". In: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Apr. 2013, pp. 1–8. DOI: 10.1109/FG.2013.6553805. URL: https://ieeexplore.ieee.org/document/6553805/?arnumber=6553805 (visited on 01/10/2025).

[55] Alexander Heimerl et al. “ForDigitStress: A multi-modal stress dataset employing a digital job interview scenario”. In: 2023.

[56] Seyedmajid Hosseini et al. “A multimodal sensor dataset for continuous stress detection of nurses in a hospital”. In: vol. 9. 1. Nature Publishing Group UK London, 2022, p. 255.

[57] Patrícia Bota et al. “A real-world dataset of group emotion experiences based on physiological data”. In: *Scientific Data* 11.1 (2024), p. 116.

[58] Matias Laporte, Martin Gjoreski, and Marc Langheinrich. “LAUREATE: A Dataset for Supporting Research in Affective Computing and Human Memory Augmentation”. In: vol. 7. 3. New York, NY, USA: Association for Computing Machinery, 2023. DOI: 10.1145/3610892. URL: https://doi.org/10.1145/3610892.

[59] Xuhai Xu et al. “GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling”. en. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6.4 (Dec. 2022), pp. 1–34. ISSN: 2474-9567. DOI: 10.1145/3569485. URL: https://dl.acm.org/doi/10.1145/3569485 (visited on 03/14/2023).

[60] Joanna C. Yau et al. “TILES-2019, A longitudinal physiologic and behavioral data set of medical residents in an intensive care unit”. In: *Sci Data* 9.536 (2022). DOI: 10.1038/s41597-022-01636-4.

[61] Xinyu Shui et al. “A dataset of daily ambulatory psychological and physiological recording for emotion research”. In: vol. 8. 1. Nature Publishing Group UK London, 2021, p. 161.

[62] Soowon Kang et al. “K-emophone: A mobile and wearable dataset with in-situ emotion, stress, and attention labels”. In: vol. 10. 1. Nature Publishing Group UK London, 2023, p. 351.

[63] Elena Smets et al. “Large-scale wearable data reveal digital phenotypes for daily-life stress detection”. In: vol. 1. 1. Nature Publishing Group UK London, 2018, p. 67.

[64] Karan Sharma et al. “A dataset of continuous affect annotations and physiological signals for emotion analysis”. en. In: *Scientific Data* 6.1 (Oct. 2019). Publisher: Nature Publishing Group, p. 196. ISSN: 2052-4463. DOI: 10.1038/s41597-019-0209-0. URL: https://www.nature.com/articles/s41597-019-0209-0 (visited on 01/10/2025).

[65] Tong Xue et al. “RCEA-360VR: Real-time, Continuous Emotion Annotation in 360° VR Videos for Collecting Precise Viewport-dependent Ground Truth Labels”. In:

Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21. New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–15. ISBN: 978-1-4503-8096-6. DOI: 10.1145/3411764.3445487. URL: https://doi.org/10.1145/3411764.3445487 (visited on 07/26/2023).

[66] Karan Sharma et al. "Continuous, Real-Time Emotion Annotation: A Novel Joystick-Based Analysis Framework". In: *IEEE Transactions on Affective Computing* 11.1 (2020), pp. 78–84. DOI: 10.1109/TAFFC.2017.2772882.

[67] J.A. Healey and R.W. Picard. "Detecting stress during real-world driving tasks using physiological sensors". In: *IEEE Transactions on Intelligent Transportation Systems* 6.2 (June 2005). Conference Name: IEEE Transactions on Intelligent Transportation Systems, pp. 156–166. ISSN: 1558-0016. DOI: 10.1109/TITS.2005.848368. URL: https://ieeexplore.ieee.org/document/1438384?arnumber=1438384 (visited on 01/10/2025).

[68] Surjya Ghosh, Bivas Mitra, and Pradipta De. "Towards Improving Emotion Self-Report Collection Using Self-Reflection". In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI EA '20. ¡conf-loc¿, ¡city¿Honolulu¡/city¿, ¡state¿HI¡/state¿, ¡country¿USA¡/country¿, ¡/conf-loc¿: Association for Computing Machinery, 2020, 1–8. ISBN: 9781450368193. DOI: 10.1145/3334480.3383019. URL: https://doi.org/10.1145/3334480.3383019.

[69] Robert Plutchik. *A psychoevolutionary theory of emotions*. 1982.

[70] Richard S Lazarus. "Progress on a cognitive-motivational-relational theory of emotion." In: *American psychologist* 46.8 (1991), p. 819.

[71] Lisa Feldman Barrett. *How emotions are made: The secret life of the brain*. Pan Macmillan, 2017.

[72] James Russell, Anna Weiss, and G. Mendelsohn. "Affect Grid: A Single-Item Scale of Pleasure and Arousal". In: *Journal of Personality and Social Psychology* 57 (Sept. 1989), pp. 493–502. DOI: 10.1037/0022-3514.57.3.493.

[73] Alexander Toet et al. "EmojiGrid: A 2D pictorial scale for the assessment of food elicited emotions". In: *Frontiers in psychology* 9 (2018), p. 412277.

[74] David Watson, Lee Anna Clark, and Auke Tellegen. "Development and validation of brief measures of positive and negative affect: The PANAS scales". In: *Journal of Personality and Social Psychology* 54.6 (1988). Place: US Publisher: American Psychological Association, pp. 1063–1070. ISSN: 1939-1315. DOI: 10.1037/0022-3514.54.6.1063.

[75] Abe Kazemzadeh et al. “Emotion twenty questions: Toward a crowd-sourced theory of emotions”. In: *International conference on affective computing and intelligent interaction*. Springer. 2011, pp. 1–10.

[76] Hao-Chun Yang and Chi-Chun Lee. “Annotation matters: A comprehensive study on recognizing intended, self-reported, and observed emotion labels using physiology”. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2019, pp. 1–7.

[77] Emily Pronin, Daniel Y Lin, and Lee Ross. “The bias blind spot: Perceptions of bias in self versus others”. In: *Personality and Social Psychology Bulletin* 28.3 (2002), pp. 369–381.

[78] Martie G Haselton et al. “Adaptive rationality: An evolutionary perspective on cognitive bias”. In: *Social Cognition* 27.5 (2009), pp. 733–763.

[79] Amid Ayobi et al. “Flexible and Mindful Self-Tracking: Design Implications from Paper Bullet Journals”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI ’18. Montreal QC, Canada: Association for Computing Machinery, 2018, 1–14. ISBN: 9781450356206. DOI: 10.1145/3173574.3173602. URL: https://doi.org/10.1145/3173574.3173602.

[80] Clara Caldeira et al. “Mobile apps for mood tracking: an analysis of features and user reviews”. In: *AMIA annual symposium proceedings*. Vol. 2017. 2018, p. 495.

[81] Helma Torkamaan and Jürgen Ziegler. “Mobile mood tracking: An investigation of concise and adaptive measurement instruments”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.4 (2020), pp. 1–30.

[82] David Bakker and Nikki Rickard. “Engagement in mobile phone app for self-monitoring of emotional wellbeing predicts changes in mental health: MoodPrism”. In: *Journal of affective disorders* 227 (2018), pp. 432–442.

[83] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[84] Leimin Tian et al. “Applied affective computing”. In: (2022).

[85] Rui Wang et al. “StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones”. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp ’14. Seattle, Washington: Association for Computing Machinery, 2014, 3–14. ISBN: 9781450329682. DOI: 10.1145/2632048.2632054. URL: https://doi.org/10.1145/2632048.2632054.

[86] David C Mohr, Mi Zhang, and Stephen M Schueller. "Personal sensing: understanding mental health using ubiquitous sensors and machine learning". In: *Annual review of clinical psychology* 13.1 (2017), pp. 23–47.

[87] Jennifer A Healey and Rosalind W Picard. "Detecting stress during real-world driving tasks using physiological sensors". In: *IEEE Transactions on intelligent transportation systems* 6.2 (2005), pp. 156–166.

[88] Sameer Neupane et al. "Momentary Stressor Logging and Reflective Visualizations: Implications for Stress Management with Wearables". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642662. URL: https://doi.org/10.1145/3613904.3642662.

[89] Blake Anthony Hickey et al. "Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review". In: *Sensors* 21.10 (2021), p. 3461.

[90] Anja Thieme et al. "Designing Human-centered AI for Mental Health: Developing Clinically Relevant Applications for Online CBT Treatment". In: *ACM Trans. Comput.-Hum. Interact.* 30.2 (Mar. 2023). ISSN: 1073-0516. DOI: 10.1145/3564752. URL: https://doi.org/10.1145/3564752.

[91] Daniel A. Adler et al. "Beyond Detection: Towards Actionable Sensing Research in Clinical Mental Healthcare". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8.4 (Nov. 2024). DOI: 10.1145/3699755. URL: https://doi.org/10.1145/3699755.

[92] Daniel A Adler et al. "Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies". In: *Plos one* 17.4 (2022), e0266516.

[93] Pragya Singh et al. "Translating Emotions to Annotations-A Participant Perspective of Physiological Emotion Data Collection". In: *arXiv preprint arXiv:2503.19636* (2025).

[94] Kat Roemmich and Nazanin Andalibi. "Data Subjects' Conceptualizations of and Attitudes Toward Automatic Emotion Recognition-Enabled Wellbeing Interventions on Social Media". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW2 (2021). DOI: 10.1145/3476049. URL: https://doi.org/10.1145/3476049.

[95] Ronak Kosti et al. "Emotion recognition in context". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1667–1675.

[96] Seungwan Jin, Bogoan Kim, and Kyungsik Han. ""I Don't Know Why I Should Use This App": Holistic Analysis on User Engagement Challenges in Mobile Mental Health". In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI '25. New York, NY, USA: Association for Computing Machinery, 2025. ISBN: 9798400713941. DOI: 10.1145/3706598.3713732. URL: https://doi.org/10.1145/3706598.3713732.

[97] Christina Kelley, Bongshin Lee, and Lauren Wilcox. "Self-tracking for Mental Wellness: Understanding Expert Perspectives and Student Experiences". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. New York, NY, USA: Association for Computing Machinery, May 2017, pp. 629–641. ISBN: 978-1-4503-4655-9. DOI: 10.1145/3025453.3025750. URL: https://doi.org/10.1145/3025453.3025750 (visited on 01/30/2023).

[98] Sebastian Scherr and Mark Goering. "Is a self-monitoring app for depression a good place for additional mental health information? Ecological momentary assessment of mental help information seeking among smartphone users". In: *Health communication* (2020).

[99] Emily G. Lattie et al. "Designing Mental Health Technologies that Support the Social Ecosystem of College Students". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, 1–15. ISBN: 9781450367080. DOI: 10.1145/3313831.3376362. URL: https://doi.org/10.1145/3313831.3376362.

[100] Dionne Bowie-DaBreo et al. "User Perspectives and Ethical Experiences of Apps for Depression: A Qualitative Analysis of User Reviews". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517498. URL: https://doi.org/10.1145/3491102.3517498.

[101] Kaylee Payne Kruzan et al. "The Perceived Utility of Smartphone and Wearable Sensor Data in Digital Self-Tracking Technologies for Mental Health". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. ¡conf-loc¿, ¡city¿Hamburg¡/city¿, ¡country¿Germany¡/country¿, ¡/conf-loc¿: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581209. URL: https://doi.org/10.1145/3544548.3581209.

[102] Stéphane Vial, Sana Boudhraâ, and Mathieu Dumont. "Human-centered design approaches in digital mental health interventions: exploratory mapping review". In: *JMIR Mental health* 9.6 (2022), e35591.

[103] Michelle Nicole Burns et al. “Harnessing context sensing to develop a mobile intervention for depression”. In: *Journal of medical Internet research* 13.3 (2011), e1838.

[104] Sicheng Zhao et al. “Personality-Aware Personalized Emotion Recognition from Physiological Signals”. en. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization, July 2018, pp. 1660–1667. ISBN: 978-0-9992411-2-7. DOI: 10.24963/ijcai.2018/230. URL: https://www.ijcai.org/proceedings/2018/230 (visited on 11/21/2023).

[105] Don Samitha Elvitigala et al. “StressShoe: a DIY toolkit for just-in-time personalised stress interventions for office workers performing sedentary tasks”. In: *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. 2021, pp. 1–14.

[106] Kwangyoung Lee et al. “Toward future-centric personal informatics: Expecting stressful events and preparing personalized interventions in stress management”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–13.

[107] Yeonsun Yang et al. “Find the Bot!: Gamifying Facial Emotion Recognition for Both Human Training and Machine Learning Data Collection”. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI ’24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642880. URL: https://doi.org/10.1145/3613904.3642880.

[108] Vassilis-Javed Khan et al. “Reconexp: a way to reduce the data loss of the experiencing sampling method”. In: *Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services*. MobileHCI ’08. Amsterdam, The Netherlands: Association for Computing Machinery, 2008, 471–476. ISBN: 9781595939524. DOI: 10.1145/1409240.1409316. URL: https://doi.org/10.1145/1409240.1409316.

[109] Liuping Wang et al. “mirrorU: Scaffolding Emotional Reflection via In-Situ Assessment and Interactive Feedback”. In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI EA ’18. Montreal QC, Canada: Association for Computing Machinery, 2018, 1–6. ISBN: 9781450356213. DOI: 10.1145/3170427.3188517. URL: https://doi.org/10.1145/3170427.3188517.

[110] James A Russell, Anna Weiss, and Gerald A Mendelsohn. “Affect grid: a single-item scale of pleasure and arousal.” In: *Journal of personality and social psychology* 57.3 (1989), p. 493.

[111] PMA Desmet. “Measuring emotion”. In: *M. Blythe, A Monk, K. Overbeeke, & P* (2003).

[112] John P. Pollak, Phil Adams, and Geri Gay. “PAM: a photographic affect meter for frequent, in situ measurement of affect”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’11. Vancouver, BC, Canada: Association for Computing Machinery, 2011, 725–734. ISBN: 9781450302289. DOI: 10.1145/1978942.1979047. URL: https://doi.org/10.1145/1978942.1979047.

[113] Swarnali Banik et al. “Towards Reducing Continuous Emotion Annotation Effort During Video Consumption: A Physiological Response Profiling Approach”. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8.3 (Sept. 2024). DOI: 10.1145/3678569. URL: https://doi.org/10.1145/3678569.

[114] Nina Rajcic and Jon McCormack. “Mirror Ritual: An Affective Interface for Emotional Self-Reflection”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, 1–13. ISBN: 9781450367080. DOI: 10.1145/3313831.3376625. URL: https://doi.org/10.1145/3313831.3376625.

[115] Si Chen et al. “Mirror Hearts: Exploring the (Mis-)Alignment between AI-Recognized and Self-Reported Emotions”. In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI EA ’23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394222. DOI: 10.1145/3544549.3585607. URL: https://doi.org/10.1145/3544549.3585607.

[116] Evangelos Karapanos. “Beyond Experience Sampling: Evaluating Personal Informatics with Technology-Assisted Reconstruction”. In: *arXiv preprint arXiv:1207.1821* (2012).

[117] Arthur A Stone et al. “A population approach to the study of emotion: diurnal rhythms of a working day examined with the Day Reconstruction Method.” In: *Emotion* 6.1 (2006), p. 139.

[118] Junze Li et al. *DiaryHelper: Exploring the Use of an Automatic Contextual Information Recording Agent for Elicitation Diary Study*. 2024. arXiv: 2404.19738 [cs.HC]. URL: https://arxiv.org/abs/2404.19738.

[119] Ruolan Wu et al. “MindShift: Leveraging Large Language Models for Mental-States-Based Problematic Smartphone Use Intervention”. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI ’24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300.

DOI: 10.1145/3613904.3642790. URL: https://doi.org/10.1145/3613904.3642790.

[120] Taewan Kim et al. "DiaryMate: Understanding User Perceptions and Experience in Human-AI Collaboration for Personal Journaling". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642693. URL: https://doi.org/10.1145/3613904.3642693.

[121] Subigya Nepal et al. "MindScape Study: Integrating LLM and Behavioral Sensing for Personalized AI-Driven Journaling Experiences". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8.4 (Nov. 2024). DOI: 10.1145/3699761. URL: https://doi.org/10.1145/3699761.

[122] Petr Slovak et al. "Designing for emotion regulation interventions: an agenda for HCI theory and research". In: *ACM Transactions on Computer-Human Interaction* 30.1 (2023), pp. 1–51.

[123] Soowon Kang et al. "Understanding emotion changes in mobile experience sampling". In: *Proceedings of the 2022 CHI conference on human factors in computing systems*. 2022, pp. 1–14.

[124] Varun Mishra et al. "Investigating contextual cues as indicators for EMA delivery". In: *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. UbiComp '17. Maui, Hawaii: Association for Computing Machinery, 2017, 935–940. ISBN: 9781450351904. DOI: 10.1145/3123024.3124571. URL: https://doi.org/10.1145/3123024.3124571.

[125] John Torous et al. "Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements". In: *BMJ Ment Health* 21.3 (2018), pp. 116–119.

[126] Shanley Corvite et al. "Data Subjects' Perspectives on Emotion Artificial Intelligence Use in the Workplace: A Relational Ethics Lens". In: *Proc. ACM Hum.-Comput. Interact.* 7.CSCW1 (2023). DOI: 10.1145/3579600. URL: https://doi.org/10.1145/3579600.

[127] Amit Baumel et al. "Objective user engagement with mental health apps: systematic search and panel-based usage analysis". In: *Journal of medical Internet research* 21.9 (2019), e14567.

[128] Aditya Ponnada et al. "Longitudinal User Engagement with Microinteraction Ecological Momentary Assessment (μEMA)". In: *Proc. ACM Interact. Mob. Wearable*

Ubiquitous Technol. 9.3 (Sept. 2025). DOI: 10.1145/3749541. URL: https://doi.org/10.1145/3749541.

[129] Christie N Scollon, Chu Kim-Prieto, and Ed Diener. "Experience sampling: Promises and pitfalls, strengths and weaknesses". In: *Journal of Happiness studies* 4.1 (2003), pp. 5–34.

[130] Surjya Ghosh et al. "Designing an Experience Sampling Method for Smartphone Based Emotion Detection". In: *IEEE Transactions on Affective Computing* 12.4 (2021), pp. 913–927. DOI: 10.1109/TAFFC.2019.2905561.

[131] Panyu Zhang et al. "A Reproducible Stress Prediction Pipeline with Mobile Sensor Data". In: *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 8.3 (2024), pp. 1–35.

[132] Varun Mishra et al. "Evaluating the reproducibility of physiological stress detection models". In: *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 4.4 (2020), pp. 1–29.

[133] Yunjo Han et al. "Systematic Evaluation of Personalized Deep Learning Models for Affect Recognition". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8.4 (2024), pp. 1–35.

[134] James A Russell. "A circumplex model of affect." In: *Journal of personality and social psychology* 39.6 (1980), p. 1161.

[135] Emily Zhou et al. "AffectEval: A Modular and Customizable Framework for Affective Computing". In: *arXiv preprint arXiv:2504.21184* (2025).

[136] Mouhannad Ali et al. "A globally generalized emotion recognition system involving different physiological signals". In: *Sensors* 18.6 (2018), p. 1905.

[137] MinSeop Lee et al. "Emotion recognition using convolutional neural network with selected statistical photoplethysmogram features". In: *Applied Sciences* 10.10 (2020), p. 3501.

[138] Liang Zhao et al. "Stress detection via multimodal multitemporal-scale fusion: A hybrid of deep learning and handcrafted feature approach". In: *IEEE Sensors Journal* 23.22 (2023), pp. 27817–27827.

[139] Jolly Ehiabhi and Haifeng Wang. "A systematic review of machine learning models in mental health analysis based on multi-channel multi-modal biometric signals". In: *BioMedInformatics* 3.1 (2023), pp. 193–219.

[140] Maciej Dzieżyc et al. “Can we ditch feature engineering? end-to-end deep learning for affect recognition from physiological sensor data”. In: *Sensors* 20.22 (2020), p. 6535.

[141] Kaiyang Zhou et al. “Conditional prompt learning for vision-language models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16816–16825.

[142] Md Kafiul Islam, Amir Rastegarnia, and Saeid Sanei. “Signal artifacts and techniques for artifacts and noise removal”. In: *Signal Processing Techniques for Computational Health Informatics*. Springer, 2020, pp. 23–79.

[143] Shkurta Gashi et al. “Detection of Artifacts in Ambulatory Electrodermal Activity Data”. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4.2 (June 2020). DOI: 10.1145/3397316. URL: https://doi.org/10.1145/3397316.

[144] K.L. Venkatachalam, Joel E. Herbrandson, and Samuel J. Asirvatham. “Signals and Signal Processing for the Electrophysiologist”. In: *Circulation: Arrhythmia and Electrophysiology* 4.6 (2011), pp. 974–981. DOI: 10.1161/CIRCEP.111.964973. eprint: https://www.ahajournals.org/doi/pdf/10.1161/CIRCEP.111.964973. URL: https://www.ahajournals.org/doi/abs/10.1161/CIRCEP.111.964973.

[145] Yali Zheng et al. *Tiny-PPG: A Lightweight Deep Neural Network for Real-Time Detection of Motion Artifacts in Photoplethysmogram Signals on Edge Devices*. 2023. arXiv: 2305.03308 [eess.SP]. URL: https://arxiv.org/abs/2305.03308.

[146] Beatrice Rammstedt et al. “A short scale for assessing the big five dimensions of personality: 10 item big five inventory (BFI-10)”. In: *methods, data, analyses* 7.2 (2013), p. 17.

[147] Oye Gureje and B Obikoya. “The GHQ-12 as a screening tool in a primary care setting”. In: *Social psychiatry and psychiatric epidemiology* 25 (1990), pp. 276–280.

[148] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[149] Benjamin Elizalde et al. *CLAP: Learning Audio Concepts From Natural Language Supervision*. 2022. arXiv: 2206.04769 [cs.SD].

[150] Yizhou Wang et al. *VaQuitA: Enhancing Alignment in LLM-Assisted Video Understanding*. 2023. arXiv: 2312.02310 [cs.CV].

[151] Anja Thieme, Danielle Belgrave, and Gavin Doherty. "Machine Learning in Mental Health: A Systematic Review of the HCI Literature to Support the Development of Effective and Implementable ML Systems". In: *ACM Trans. Comput.-Hum. Interact.* 27.5 (2020). ISSN: 1073-0516. DOI: 10.1145/3398069. URL: https://doi.org/10.1145/3398069.

[152] Hyun K Kim et al. "Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment". In: *Applied ergonomics* 69 (2018), pp. 66–73.

[153] Benjamin J Li et al. "A public database of immersive VR videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures". In: *Frontiers in psychology* 8 (2017), p. 2116.

[154] S. Koelstra et al. "DEAP: A Database for Emotion Analysis ;Using Physiological Signals". en. In: *IEEE Transactions on Affective Computing* 3.1 (Jan. 2012), pp. 18–31. ISSN: 1949-3045. DOI: 10.1109/T-AFFC.2011.15. URL: http://ieeexplore.ieee.org/document/5871728/ (visited on 01/10/2025).

[155] Luma Tabbaa et al. "VREED: Virtual Reality Emotion Recognition Dataset Using Eye Tracking & Physiological Measures". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.4 (Dec. 2022), 178:1–178:20. DOI: 10.1145/3495002. URL: https://doi.org/10.1145/3495002 (visited on 07/26/2023).

[156] Ghufran Shafiq and Kalyana Chakravarthy Veluvolu. "Multimodal chest surface motion data for respiratory and cardiovascular monitoring applications". In: *Scientific data* 4.1 (2017), pp. 1–12.

[157] Unai Zalabarria et al. "A low-cost, portable solution for stress and relaxation estimation based on a real-time fuzzy algorithm". In: *IEEE Access* 8 (2020), pp. 74118–74128.

[158] Noppawit Aeimpreeda et al. "Study of drowsiness from simple physiological signals testing: A signal processing perspective". In: *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE. 2020, pp. 738–741.

[159] K Maheshkumar et al. "Validation of PC-based sound card with biopac for digitalization of ECG recording in short-term HRV analysis". In: *North American journal of medical sciences* 8.7 (2016), p. 307.

[160] Jainendra Shukla et al. "Feature Extraction and Selection for Emotion Recognition from Electrodermal Activity". In: *IEEE Transactions on Affective Computing* 12.4 (2021), pp. 857–869. DOI: 10.1109/TAFFC.2019.2901673.

[161] Van-Tu Ninh et al. *An Improved Subject-Independent Stress Detection Model Applied to Consumer-grade Wearable Devices*. 2022. arXiv: 2203.09663 [cs.LG].

[162] Alberto Greco et al. "cvxEDA: A convex optimization approach to electrodermal activity processing". In: *IEEE transactions on biomedical engineering* 63.4 (2015), pp. 797–804.

[163] Dominique Makowski et al. "NeuroKit2: A Python toolbox for neurophysiological signal processing". In: *Behavior research methods* (2021), pp. 1–8.

[164] Victor Sanh et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *arXiv preprint arXiv:1910.01108* (2019).

[165] Alexis Conneau et al. "Unsupervised cross-lingual representation learning at scale". In: *arXiv preprint arXiv:1911.02116* (2019).

[166] Stanisław Saganowski et al. "Emognition dataset: emotion recognition with self-reports, facial expressions, and physiology using wearables". en. In: *Scientific Data* 9.1 (Apr. 2022). Publisher: Nature Publishing Group, p. 158. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01262-0. URL: https://www.nature.com/articles/s41597-022-01262-0 (visited on 01/10/2025).

[167] Seyedmajid Hosseini et al. "A multimodal sensor dataset for continuous stress detection of nurses in a hospital". In: *Scientific Data* 9.1 (2022), p. 255.

[168] Fanny Larradet et al. "Toward emotion recognition from physiological signals in the wild: approaching the methodological issues in real-life data collection". In: *Frontiers in psychology* 11 (2020), p. 1111.

[169] Yekta Said Can, Bhargavi Mahesh, and Elisabeth André. "Approaches, applications, and challenges in physiological emotion recognition—a tutorial overview". In: *Proceedings of the IEEE* (2023).

[170] Patricia J Bota et al. "A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals". In: *IEEE access* 7 (2019), pp. 140990–141020.

[171] Feng Tian et al. "Emotional arousal in 2D versus 3D virtual reality environments". In: *PloS one* 16.9 (2021), e0256211.

[172] Gina Gilpin, James Gain, and Gosia Lipinska. "The physiological signature of sadness: A comparison between text, film and virtual reality". In: *Brain and Cognition* 152 (2021), p. 105734.

[173] Elizabeth A Kensinger and Daniel L Schacter. "Processing emotional pictures and words: Effects of valence and arousal". In: *Cognitive, Affective, & Behavioral Neuroscience* 6.2 (2006), pp. 110–126.

[174] Lisa Feldman Barrett. "Valence is a basic building block of emotional life". In: *Journal of Research in Personality* 40.1 (2006), pp. 35–55.

[175] Lisa Feldman Barrett and James A Russell. "The structure of current affect: Controversies and emerging consensus". In: *Current directions in psychological science* 8.1 (1999), pp. 10–14.

[176] Min Qin et al. "General Health Questionnaire-12 reliability, factor structure, and external validity among older adults in India". In: *Indian Journal of Psychiatry* 60.1 (2018), p. 56.

[177] Beatrice Rammstedt and Oliver P John. "Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German". In: *Journal of research in Personality* 41.1 (2007), pp. 203–212.

[178] Rebeca Romo-De León et al. "EEG and Physiological Signals Dataset from Participants during Traditional and Partially Immersive Learning Experiences in Humanities". In: *Data* 9.5 (2024), p. 68.

[179] Margaret M Bradley and Peter J Lang. "Measuring emotion: the self-assessment manikin and the semantic differential". In: *Journal of behavior therapy and experimental psychiatry* 25.1 (1994), pp. 49–59.

[180] J. Richard Jennings et al. "Alternate Cardiovascular Baseline Assessment Techniques: Vanilla or Resting Baseline". In: *Psychophysiology* 29.6 (1992), pp. 742–750. DOI: https://doi.org/10.1111/j.1469-8986.1992.tb02052.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.1992.tb02052.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8986.1992.tb02052.x.

[181] Biopac. *Data Acquisition, Loggers, Amplifiers, Transducers, Electrodes: BIOPAC*. https://www.biopac.com/. 2023.

[182] David Papineau. "Realism and epistemology". In: *Mind* 94.375 (1985), pp. 367–388.

[183] Virginia Braun and Victoria Clarke. "Using thematic analysis in psychology". In: *Qualitative research in psychology* 3.2 (2006), pp. 77–101.

[184] Martin Gjoreski et al. "Monitoring stress with a wrist device using context". In: *Journal of biomedical informatics* 73 (2017), pp. 159–170.

[185] Seyedmajid Hosseini et al. “A multimodal sensor dataset for continuous stress detection of nurses in a hospital”. en. In: *Scientific Data* 9.1 (June 2022). Publisher: Nature Publishing Group, p. 255. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01361-y. URL: https://www.nature.com/articles/s41597-022-01361-y (visited on 01/10/2025).

[186] Martin Gjoreski et al. “Continuous stress detection using a wrist device: in laboratory and real life”. In: *proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct*. 2016, pp. 1185–1193.

[187] Paul Ricoeur. “Philosophical hermeneutics and theological hermeneutics”. In: *Studies in religion/Sciences religieuses* 5.1 (1975), pp. 14–33.

[188] Klaus R Scherer. “Appraisal theory.” In: (1999).

[189] Lisa Feldman Barrett. “The theory of constructed emotion: an active inference account of interoception and categorization”. In: *Social cognitive and affective neuroscience* 12.1 (2017), pp. 1–23.

[190] ELLIOT Aronson, TD Wilson, and RM Akert. “Interpersonal Attraction”. In: *E. Aronson et al.(5 th Ed.) Social Psychology* (2005), pp. 316–355.

[191] Mel Slater and Martin Usoh. “Presence in immersive virtual environments”. In: *Proceedings of IEEE virtual reality annual international symposium*. IEEE. 1993, pp. 90–96.

[192] Crescent Jicol et al. “Designing and Assessing a Virtual Reality Simulation to Build Resilience to Street Harassment”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. ¡conf-loc¿, ¡city¿New Orleans¡/city¿, ¡state¿LA¡/state¿, ¡country¿USA¡/country¿, ¡/conf-loc¿: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3502129. URL: https://doi.org/10.1145/3491102.3502129.

[193] Jaakko Tervonen et al. “In Search of Harmful Stress”. In: *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*. UbiComp ’21. Virtual, USA: Association for Computing Machinery, 2021, 215–217. ISBN: 9781450384612. DOI: 10.1145/3460418.3479335. URL: https://doi.org/10.1145/3460418.3479335.

[194] Robin Burchard et al. “WashSpot: Real-Time Spotting and Detection of Enacted Compulsive Hand Washing with Wearable Devices”. In: *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*. UbiComp/ISWC ’22 Adjunct. Cambridge, United Kingdom: Association for Computing

Machinery, 2023, 483–487. ISBN: 9781450394239. DOI: 10.1145/3544793.3563428. URL: https://doi.org/10.1145/3544793.3563428.

[195] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. "Laughter Recognition Using Non-Invasive Wearable Devices". In: *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*. PervasiveHealth'19. Trento, Italy: Association for Computing Machinery, 2019, 262–271. ISBN: 9781450361262. DOI: 10.1145/3329189.3329216. URL: https://doi.org/10.1145/3329189.3329216.

[196] Enrique Garcia-Ceja et al. "Depresjon: A Motor Activity Database of Depression Episodes in Unipolar and Bipolar Patients". In: *Proceedings of the 9th ACM Multimedia Systems Conference*. MMSys '18. Amsterdam, Netherlands: Association for Computing Machinery, 2018, 472–477. ISBN: 9781450351928. DOI: 10.1145/3204949.3208125. URL: https://doi.org/10.1145/3204949.3208125.

[197] Tao Bi et al. "Towards Chatbot-Supported Self-Reporting for Increased Reliability and Richness of Ground Truth for Automatic Pain Recognition: Reflections on Long-Distance Runners and People with Chronic Pain". In: *Companion Publication of the 2021 International Conference on Multimodal Interaction*. ICMI '21 Companion. Montreal, QC, Canada: Association for Computing Machinery, 2021, 43–53. ISBN: 9781450384711. DOI: 10.1145/3461615.3485670. URL: https://doi.org/10.1145/3461615.3485670.

[198] Kim Olivia Snooks et al. "Context-Aware Wearables: The Last Thing We Need is a Pandemic of Stray Cats". In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI EA '21. ¡conf-loc¿, ¡city¿Yokohama¡/city¿, ¡country¿Japan¡/country¿, ¡/conf-loc¿: Association for Computing Machinery, 2021. ISBN: 9781450380959. DOI: 10.1145/3411763.3450367. URL: https://doi.org/10.1145/3411763.3450367.

[199] Divy Thakkar et al. "When is Machine Learning Data Good?: Valuing in Public Health Datafication". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. ¡conf-loc¿, ¡city¿New Orleans¡/city¿, ¡state¿LA¡/state¿, ¡country¿USA¡/country¿, ¡/conf-loc¿: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3501868. URL: https://doi.org/10.1145/3491102.3501868.

[200] Han Yu and Akane Sano. "Semi-Supervised Learning for Wearable-Based Momentary Stress Detection in the Wild". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7.2 (2023). DOI: 10.1145/3596246. URL: https://doi.org/10.1145/3596246.

[201] Luca Romeo et al. “Multiple Instance Learning for Emotion Recognition Using Physiological Signals”. In: *IEEE Transactions on Affective Computing* 13.1 (2022), pp. 389–407. DOI: 10.1109/TAFFC.2019.2954118.

[202] William H Walker et al. “Circadian rhythm disruption and mental health”. In: *Translational psychiatry* 10.1 (2020), p. 28.

[203] Hamidan Z. Wijasena, Ridi Ferdiana, and Sunu Wibirama. “A Survey of Emotion Recognition using Physiological Signal in Wearable Devices”. In: *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*. 2021, pp. 1–6. DOI: 10.1109/AIMS52415.2021.9466092.

[204] Angeliki Metallinou and Shrikanth Narayanan. “Annotation and processing of continuous emotional attributes: Challenges and opportunities”. In: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 2013, pp. 1–8. DOI: 10.1109/FG.2013.6553804.

[205] Nan Gao et al. “Critiquing Self-report Practices for Human Mental and Wellbeing Computing at Ubicomp”. In: *arXiv preprint arXiv:2311.15496* (2023).

[206] Harmanpreet Kaur et al. ““I Didn’t Know I Looked Angry”: Characterizing Observed Emotion and Reported Affect at Work”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517453. URL: https://doi.org/10.1145/3491102.3517453.

[207] Vivian Genaro Motti. “Assistive wearables: opportunities and challenges”. In: *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. UbiComp/ISWC ’19 Adjunct. London, United Kingdom: Association for Computing Machinery, 2019, 1040–1043. ISBN: 9781450368698. DOI: 10.1145/3341162.3349573. URL: https://doi.org/10.1145/3341162.3349573.

[208] Niels van Berkel et al. “Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports”. In: *Int. J. Hum.-Comput. Stud.* 125.C (2019), 118–128. ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2018.12.002. URL: https://doi.org/10.1016/j.ijhcs.2018.12.002.

[209] Stefan Schneider et al. “Comparability of emotion dynamics derived from ecological momentary assessments, daily diaries, and the day reconstruction method: Observational study”. In: *Journal of Medical Internet Research* 22.9 (2020), e19201.

[210] Soowon Kang et al. “Understanding Emotion Changes in Mobile Experience Sampling”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3501944. URL: https://doi.org/10.1145/3491102.3501944.

[211] Renwen Zhang et al. “Designing for Emotional Well-being: Integrating Persuasion and Customization into Mental Health Technologies”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445771. URL: https://doi.org/10.1145/3411764.3445771.

[212] Rosalind Picard. *W.,(1997). Affective Computing*. 1997.

[213] Lisa Feldman Barrett et al. “Knowing what you’re feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation”. In: *Cognition & Emotion* 15.6 (2001), pp. 713–724.

[214] Kathy Baxter, Catherine Courage, and Kelly Caine. *Understanding your users: a practical guide to user research methods*. Morgan Kaufmann, 2015.

[215] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.

[216] Kristen A Lindquist et al. “Emotion perception, but not affect perception, is impaired with semantic memory loss.” In: *Emotion* 14.2 (2014), p. 375.

[217] Samuel J Stratton. “Population research: convenience sampling strategies”. In: *Prehospital and disaster Medicine* 36.4 (2021), pp. 373–374.

[218] Victoria Clarke and Virginia Braun. “Thematic analysis”. In: *The journal of positive psychology* 12.3 (2017), pp. 297–298.

[219] Yonghai Yu and Yun Bi. “A study on “5W1H” user analysis on interaction design of interface”. In: *2010 IEEE 11th International Conference on Computer-Aided Industrial Design & Conceptual Design 1*. Vol. 1. IEEE. 2010, pp. 329–332.

[220] Subigya Nepal et al. “Contextual AI Journaling: Integrating LLM and Time Series Behavioral Sensing Technology to Promote Self-Reflection and Well-being using the MindScape App”. In: *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–8.

[221] Maria Dolores C Tongco. “Purposive sampling as a tool for informant selection”. In: (2007).

[222] Aditya Bhattacharya, Simone Stumpf, and Katrien Verbert. "Representation Debiasing of Generated Data Involving Domain Experts". In: *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. UMAP Adjunct '24. Cagliari, Italy: Association for Computing Machinery, 2024, 516–522. ISBN: 9798400704666. DOI: 10.1145/3631700.3664910. URL: https://doi.org/10.1145/3631700.3664910.

[223] Bhargavi Mahesh et al. "Requirements for a Reference Dataset for Multimodal Human Stress Detection". In: *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. 2019, pp. 492–498. DOI: 10.1109/PERCOMW.2019.8730884.

[224] Christina Kelley, Bongshin Lee, and Lauren Wilcox. "Self-tracking for Mental Wellness: Understanding Expert Perspectives and Student Experiences". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: Association for Computing Machinery, 2017, 629–641. ISBN: 9781450346559. DOI: 10.1145/3025453.3025750. URL: https://doi.org/10.1145/3025453.3025750.

[225] Mark G Haviland. "Structure of the twenty-item Toronto Alexithymia Scale". In: *Journal of personality assessment* 66.1 (1996), pp. 116–125.

[226] David Preece et al. "The psychometric assessment of alexithymia: Development and validation of the Perth Alexithymia Questionnaire". In: *Personality and Individual Differences* 132 (2018), pp. 32–44.

[227] Nan Gao et al. "Investigating the Reliability of Self-report Data in the Wild: The Quest for Ground Truth". In: *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*. UbiComp/ISWC '21 Adjunct. Virtual, USA: Association for Computing Machinery, 2021, 237–242. ISBN: 9781450384612. DOI: 10.1145/3460418.3479338. URL: https://doi.org/10.1145/3460418.3479338.

[228] Mohammad Hossein Jarrahi, Ali Memariani, and Shion Guha. "The Principles of Data-Centric AI". In: *Commun. ACM* 66.8 (July 2023), 84–92. ISSN: 0001-0782. DOI: 10.1145/3571724. URL: https://doi.org/10.1145/3571724.

[229] Veniamin Veselovsky et al. *Localized Cultural Knowledge is Conserved and Controllable in Large Language Models*. 2025. arXiv: 2504.10191 [cs.CL]. URL: https://arxiv.org/abs/2504.10191.

[230] Anupriya Tuli et al. "Harmony: close knitted mhealth assistance for patients, caregivers and doctors for managing SMIs". In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*.

UbiComp '16. Heidelberg, Germany: Association for Computing Machinery, 2016, 1144–1152. ISBN: 9781450344623. DOI: 10.1145/2968219.2968301. URL: https://doi.org/10.1145/2968219.2968301.

[231] Matthew K Nock et al. "The emotion reactivity scale: development, evaluation, and relation to self-injurious thoughts and behaviors". In: *Behavior therapy* 39.2 (2008), pp. 107–116.

[232] Richard D Lane et al. "The Levels of Emotional Awareness Scale: A cognitive-developmental measure of emotion". In: *Journal of personality assessment* 55.1-2 (1990), pp. 124–134.

[233] Bahar Sert and Selami Varol Ülker. "A Review of LIWC and Machine Learning Approaches On Mental Health Diagnosis". In: *Social Review of Technology and Change* 1.2 (2023), pp. 71–92.

[234] Lawrence A Palinkas et al. "Purposeful sampling for qualitative data collection and analysis in mixed method implementation research". In: *Administration and policy in mental health and mental health services research* 42 (2015), pp. 533–544.

[235] Saleema Amershi et al. "Guidelines for human-AI interaction". In: *Proceedings of the 2019 chi conference on human factors in computing systems*. 2019, pp. 1–13.

[236] Susanna Loeb et al. "Descriptive Analysis in Education: A Guide for Researchers. NCEE 2017-4023." In: *National Center for Education Evaluation and Regional Assistance* (2017).

[237] Nikki Rickard et al. "Development of a mobile phone app to support self-monitoring of emotional well-being: a mental health digital innovation". In: *JMIR mental health* 3.4 (2016), e6202.

[238] MetricWire Inc. *MetricWire*. 2025. URL: https://www.metricwire.com.

[239] *Daylio - Mood Tracker and Diary*. https://daylio.net. 2024.

[240] Aditya Vaidyam, John Halamka, John Torous, et al. "Enabling research and clinical use of patient-generated health data (the mindLAMP Platform): digital phenotyping study". In: *JMIR mHealth and uHealth* 10.1 (2022), e30557.

[241] *Headspace - Mindfulness and Meditation App*. https://www.headspace.com. 2024.

[242] Happify Research. *Happify*. Mobile application. 2025. URL: https://www.happify.com/research/.

[243] Annika Howells, Itai Ivtzan, and Francisco Jose Eiroa-Orosa. “Putting the ‘app’in happiness: a randomised controlled trial of a smartphone-based mindfulness intervention to enhance wellbeing”. In: *Journal of happiness studies* 17 (2016), pp. 163–185.

[244] *Calm - Meditation and Sleep App*. https://www.calm.com. 2024.

[245] Apple Inc. *Apple Watch*. 2024. URL: https://www.apple.com/watch/.

[246] Samsung Electronics. *Samsung Galaxy Watch*. 2024. URL: https://www.samsung.com/global/galaxy/galaxy-watch/.

[247] Fitbit, Inc. *Fitbit Wearables*. 2024. URL: https://www.fitbit.com/global/us/products.

[248] Oura Health Oy. *Oura Ring*. 2024. URL: https://ouraring.com.

[249] Whoop, Inc. *WHOOP Wearable*. 2024. URL: https://www.whoop.com.

[250] *Wysa - Mental Health Support*. https://www.wysa.io. 2024.

[251] *Woebot - Your Self-Care Expert*. https://woebothealth.com. 2024.

[252] Jukka-Pekka Onnela et al. “Beiwe: A data collection platform for high-throughput digital phenotyping”. In: *Journal of Open Source Software* 6.68 (2021), p. 3417.

[253] Denzil Ferreira, Vassilis Kostakos, and Anind K Dey. “AWARE: mobile context instrumentation framework”. In: *Frontiers in ICT* 2 (2015), p. 6.

[254] Google Inc. *Paco: Personal Analytics Companion*. 2015. URL: https://code.google.com/archive/p/paco/.

[255] ilumivu. *Ecological Momentary Assessment (mEMA) App*. 2024. URL: https://ilumivu.com/solutions/ecological-momentary-assessment-app/.

[256] Sabrina Thai and Elizabeth Page-Gould. “ExperienceSampler: An open-source scaffold for building smartphone apps for experience sampling.” In: *Psychological Methods* 23.4 (2018), p. 729.

[257] Kristof Meers et al. “mobileQ: A free user-friendly application for collecting experience sampling data”. In: *Behavior Research Methods* 52 (2020), pp. 1510–1515.

[258] Taewan Kim et al. “MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients’ Journaling”. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI ’24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642937. URL: https://doi.org/10.1145/3613904.3642937.

[259] Yubin Kim et al. *Health-LLM: Large Language Models for Health Prediction via Wearable Sensor Data*. 2024. arXiv: 2401.06866 [cs.CL]. URL: https://arxiv.org/abs/2401.06866.

[260] Ruth Malkinson. “Cognitive-behavioral grief therapy: The ABC model of rational-emotion behavior therapy”. In: *Psihologijske teme* 19.2 (2010), pp. 289–305.

[261] Taewan Kim et al. “MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients’ Journaling”. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Vol. 55. CHI ’24. ACM, May 2024, 1–20. DOI: 10.1145/3613904.3642937. URL: http://dx.doi.org/10.1145/3613904.3642937.

[262] Amy Wenzel. “Basic strategies of cognitive behavioral therapy”. In: *Psychiatric Clinics* 40.4 (2017), pp. 597–609.

[263] Jane Kaye et al. “Dynamic consent: a patient interface for twenty-first century research networks”. In: *European journal of human genetics* 23.2 (2015), pp. 141–146.

[264] Mike A. Merrill et al. *Transforming Wearable Data into Health Insights using Large Language Model Agents*. 2024. arXiv: 2406.06464 [cs.AI]. URL: https://arxiv.org/abs/2406.06464.

[265] Marc Schröder, Hannes Pirker, and Myriam Lamolle. “First suggestions for an emotion annotation and representation language”. In: *Proceedings of LREC*. Vol. 6. 2006, pp. 88–92.

[266] Inexika Inc. *iMoodJournal – Mood Tracking Mobile Application*. https://www.imoodjournal.com/. 2025.

[267] Mindsera. *Mindsera: AI-powered journal for mental fitness*. https://www.mindsera.com/. 2025.

[268] Luke Stark and Jesse Hoey. “The Ethics of Emotion in Artificial Intelligence Systems”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, 2021, 782–793. ISBN: 9781450383097. DOI: 10.1145/3442188.3445939. URL: https://doi.org/10.1145/3442188.3445939.

[269] Lakmal Meegahapola et al. “Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries”. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6.4 (Jan. 2023). DOI: 10.1145/3569483. URL: https://doi.org/10.1145/3569483.

[270] Elena Smets et al. “Large-scale wearable data reveal digital phenotypes for daily-life stress detection”. In: *NPJ digital medicine* 1.1 (2018), p. 67.

[271] Ananya Bhattacharjee et al. “Integrating Individual and Social Contexts into Self-Reflection Technologies”. In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI EA ’23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394222. DOI: 10.1145/3544549.3573803. URL: https://doi.org/10.1145/3544549.3573803.

[272] Donghoon Shin et al. “Exploring the Effects of AI-assisted Emotional Support Processes in Online Mental Health Community”. In: *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI EA ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391566. DOI: 10.1145/3491101.3519854. URL: https://doi.org/10.1145/3491101.3519854.

[273] Alaa Abd-Alrazaq et al. “Wearable artificial intelligence for detecting anxiety: systematic review and meta-analysis”. In: *Journal of medical Internet research* 25 (2023), e48754.

[274] Alaa Abd-Alrazaq et al. “Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression”. In: *NPJ Digital Medicine* 6.1 (2023), p. 84.

[275] Niels van Berkel, Denzil Ferreira, and Vassilis Kostakos. “The Experience Sampling Method on Mobile Devices”. In: *ACM Comput. Surv.* 50.6 (Dec. 2017). ISSN: 0360-0300. DOI: 10.1145/3123988. URL: https://doi.org/10.1145/3123988.

[276] Ha Le et al. “Feasibility and Utility of Multimodal Micro Ecological Momentary Assessment on a Smartwatch”. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI ’25. New York, NY, USA: Association for Computing Machinery, 2025. ISBN: 9798400713941. DOI: 10.1145/3706598.3714086. URL: https://doi.org/10.1145/3706598.3714086.

[277] Matteo Busso et al. “DiversityOne: A multi-country smartphone sensor dataset for everyday life behavior modeling”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9.1 (2025), pp. 1–49.

[278] Yuwei Zhang et al. “SensorLM: Learning the Language of Wearable Sensors”. In: *arXiv preprint arXiv:2506.09108* (2025).

[279] Kevin Doherty, Andreas Balaskas, and Gavin Doherty. “The Design of Ecological Momentary Assessment Technologies”. In: *Interacting with Computers* 32.1 (2020), pp. 257–278. DOI: 10.1093/iwcomp/iwaa019.

[280] Romualdo Gondomar and Enric Mor. “Understanding Agency in Human-Computer Interaction Design”. In: *Human-Computer Interaction. Theory, Methods and Tools: Thematic Area, HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2021, 137–149. ISBN: 978-3-030-78461-4. DOI: 10.1007/978-3-030-78462-1_10. URL: https://doi.org/10.1007/978-3-030-78462-1_10.

[281] Richard M Ryan and Edward L Deci. “Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being.” In: *American psychologist* 55.1 (2000), p. 68.

[282] Yung-Ju Chang et al. “An investigation of using mobile and situated crowdsourcing to collect annotated travel activity data in real-word settings”. In: *International Journal of Human-Computer Studies* 102 (2017), pp. 81–102.

[283] Preethi Srinivas et al. “Context-sensitive ecological momentary assessment: application of user-centered design for improving user satisfaction and engagement during self-report”. In: *JMIR mHealth and uHealth* 7.4 (2019), e10894.

[284] Wei Wang et al. “Designing adaptive user interfaces for mHealth applications targeting chronic disease: A User-centric approach”. In: *ACM Trans Softw Eng Methodol* 1.1 (2024).

[285] Yung-Ju Chang, Gaurav Paruthi, and Mark W. Newman. “A field study comparing approaches to collecting annotated activity data in real-world settings”. In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp ’15. Osaka, Japan: Association for Computing Machinery, 2015, 671–682. ISBN: 9781450335744. DOI: 10.1145/2750858.2807524. URL: https://doi.org/10.1145/2750858.2807524.

[286] Gabriela Villalobos-Zúñiga et al. “Informed Choices, Progress Monitoring and Comparison with Peers: Features to Support the Autonomy, Competence and Relatedness Needs, as Suggested by the Self-Determination Theory”. In: *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. MobileHCI ’21. Toulouse & Virtual, France: Association for Computing Machinery, 2021. ISBN: 9781450383288. DOI: 10.1145/3447526.3472039. URL: https://doi.org/10.1145/3447526.3472039.

[287] Arthur A Stone, Stefan Schneider, and Joshua M Smyth. “Evaluation of pressing issues in ecological momentary assessment”. In: *Annual Review of Clinical Psychology* 19.1 (2023), pp. 107–131.

[288] Nicolas Simonazzi et al. “The Geneva Emotion Wheel Mobile Interface: an Instrument to Report Emotions on Android Devices”. In: *ERGO’IA 2021-De l’Interaction Homme-Machine à la Relation Homme-Machine, comment concevoir des systèmes performants et éthiques*. 2021.

[289] R Siqueira Reis, AA Hino, and CR Añez. “Perceived stress scale”. In: *J. health Psychol* 15.1 (2010), pp. 107–114.

[290] Debra Trampe, Jordi Quoidbach, and Maxime Taquet. “Emotions in everyday life”. In: *PloS one* 10.12 (2015), e0145450.

[291] R Michael Bagby, James DA Parker, and Graeme J Taylor. “Twenty-five years with the 20-item Toronto Alexithymia Scale”. In: *Journal of psychosomatic research* 131 (2020), p. 109940.

[292] David A Preece et al. “The Emotion Regulation Questionnaire-Short Form (ERQ-S): A 6-item measure of cognitive reappraisal and expressive suppression”. In: *Journal of Affective Disorders* 340 (2023), pp. 855–861.

[293] Bruce W Smith et al. “The brief resilience scale: assessing the ability to bounce back”. In: *International journal of behavioral medicine* 15 (2008), pp. 194–200.

[294] Leo A Goodman. “Snowball sampling”. In: *The annals of mathematical statistics* (1961), pp. 148–170.

[295] Satu Elo and Helvi Kyngäs. “The qualitative content analysis process”. In: *Journal of advanced nursing* 62.1 (2008), pp. 107–115.

[296] James J Gross et al. “Emotion regulation: Conceptual and empirical foundations”. In: *Handbook of emotion regulation* 2.1 (2014), pp. 3–20.

[297] Pragya Singh et al. “EEVR: A Dataset of Paired Physiological Signals and Textual Descriptions for Joint Emotion Representation Learning”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 15765–15778.

[298] Ira J Roseman and Craig A Smith. “Appraisal theory”. In: *Appraisal processes in emotion: Theory, methods, research* (2001), pp. 3–19.

[299] World Health Organization. *Mental health atlas 2024*. World Health Organization, 2025.

[300] World Health Organization. *World mental health report: Transforming mental health for all*. World Health Organization, 2022.

[301] Jakob E. Bardram and Aleksandar Matic. "A Decade of Ubiquitous Computing Research in Mental Health". In: *IEEE Pervasive Computing* 19.1 (2020), pp. 62–72. DOI: 10.1109/MPRV.2019.2925338.

[302] Anja Thieme, Danielle Belgrave, and Gavin Doherty. "Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 27.5 (2020), pp. 1–53.

[303] Subigya Nepal et al. "Capturing the College Experience: A Four-Year Mobile Sensing Study of Mental Health, Resilience and Behavior of College Students during the Pandemic". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8.1 (Mar. 2024). DOI: 10.1145/3643501. URL: https://doi.org/10.1145/3643501.

[304] Jean Costa et al. "BoostMeUp: Improving Cognitive Performance in the Moment by Unobtrusively Regulating Emotions with a Smartwatch". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3.2 (June 2019). DOI: 10.1145/3328911. URL: https://doi.org/10.1145/3328911.

[305] Xuhai Xu et al. "TypeOut: Leveraging Just-in-Time Self-Affirmation for Smartphone Overuse Reduction". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517476. URL: https://doi.org/10.1145/3491102.3517476.

[306] Stephanie Balters et al. "Calm Commute: Guided Slow Breathing for Daily Stress Management in Drivers". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4.1 (Mar. 2020). DOI: 10.1145/3380998. URL: https://doi.org/10.1145/3380998.

[307] Kwangyoung Lee and Hwajung Hong. "Designing for Self-Tracking of Emotion and Experience with Tangible Modality". In: *Proceedings of the 2017 Conference on Designing Interactive Systems*. DIS '17. Edinburgh, United Kingdom: Association for Computing Machinery, 2017, 465–475. ISBN: 9781450349222. DOI: 10.1145/3064663.3064697. URL: https://doi.org/10.1145/3064663.3064697.

[308] Amit Baumel, Stav Edan, and John M Kane. "Is there a trial bias impacting user engagement with unguided e-mental health interventions? A systematic comparison of published reports and real-world usage of the same programs". In: *Translational behavioral medicine* 9.6 (2019), pp. 1020–1033.

[309] Jessica M Lipschitz et al. “The engagement problem: a review of engagement with digital mental health interventions and recommendations for a path forward”. In: *Current treatment options in psychiatry* 10.3 (2023), pp. 119–135.

[310] Katharine A Smith et al. “Engagement and attrition in digital mental health: current challenges and potential solutions”. In: *npj Digital Medicine* 8.1 (2025), p. 398.

[311] Judith Borghouts et al. “Barriers to and facilitators of user engagement with digital mental health interventions: systematic review”. In: *Journal of medical Internet research* 23.3 (2021), e24387.

[312] Jieun Lim et al. “Exploring Context-Aware Mental Health Self-Tracking Using Multimodal Smart Speakers in Home Environments”. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI ’24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642846. URL: https://doi.org/10.1145/3613904.3642846.

[313] Silja Litvin et al. “The impact of a gamified mobile mental health app (eQuoo) on resilience and mental health in a student population: large-scale randomized controlled trial”. In: *JMIR Mental Health* 10 (2023), e47285.

[314] Stephanie G Six et al. “Examining the effectiveness of gamification in mental health apps for depression: systematic review and meta-analysis”. In: *JMIR mental health* 8.11 (2021), e32199.

[315] Ananya Bhattacharjee et al. “Investigating the Role of Situational Disruptors in Engagement with Digital Mental Health Tools”. In: *Proc. ACM Hum.-Comput. Interact.* 9.7 (Oct. 2025). DOI: 10.1145/3757487. URL: https://doi.org/10.1145/3757487.

[316] Jacinta Jardine et al. “Between Rhetoric and Reality: Real-world Barriers to Uptake and Early Engagement in Digital Mental Health Interventions”. In: *ACM Trans. Comput.-Hum. Interact.* 31.2 (Feb. 2024). ISSN: 1073-0516. DOI: 10.1145/3635472. URL: https://doi.org/10.1145/3635472.

[317] Frensen Salim and Sunjun Kim. ““I Can Feel What I Used”: A Diary Study of Smartwatch Features and Emotional Experiences”. In: *Proceedings of the 25th International Conference on Mobile Human-Computer Interaction*. MobileHCI ’23 Companion. Athens, Greece: Association for Computing Machinery, 2023. ISBN: 9781450399241. DOI: 10.1145/3565066.3608689. URL: https://doi.org/10.1145/3565066.3608689.

[318] Matthew Barker-Canler et al. “Flexible Minimalist Self-Tracking to Support Individual Reflection”. In: *ACM Trans. Comput.-Hum. Interact.* 31.3 (Aug. 2024). ISSN:

1073-0516. DOI: 10.1145/3660339. URL: https://doi.org/10.1145/3660339.

[319] Jiaying "Lizzy" Liu et al. "From Regulation to Support: Centering Humans in Technology-Mediated Emotion Intervention in Care Contexts". In: *Proc. ACM Hum.-Comput. Interact.* 9.7 (Oct. 2025). DOI: 10.1145/3757605. URL: https://doi.org/10.1145/3757605.

[320] W Gerrod Parrott. *Emotions in social psychology: Essential readings*. psychology press, 2001.

[321] Jingoog Kim and Mary Lou Maher. "Conceptual metaphors for designing smart environments: device, robot, and friend". In: *Frontiers in Psychology* 11 (2020), p. 198.

[322] David Watson and Lee Anna Clark. "The PANAS-X: Manual for the positive and negative affect schedule-expanded form". In: *Unpublished manuscript, University of Iowa* (1994).

[323] Gloria Willcox. "The feeling wheel: A tool for expanding awareness of emotions and increasing spontaneity and intimacy". In: *Transactional Analysis Journal* 12.4 (1982), pp. 274–276.

[324] Albert Ellis. "The revised ABC's of rational-emotive therapy (RET)". In: *Journal of rational-emotive and cognitive-behavior therapy* 9.3 (1991), pp. 139–172.

[325] Shan Feng et al. "How self-tracking and the quantified self promote health and well-being: systematic review". In: *Journal of Medical Internet Research* 23.9 (2021), e25171.

[326] Graham Gibbs. "Learning by doing: A guide to teaching and learning methods". In: *Further Education Unit* (1988).

[327] Viswanath Venkatesh et al. "User acceptance of information technology: Toward a unified view". In: *MIS quarterly* (2003), pp. 425–478.

[328] Lucy Yardley et al. "Understanding and promoting effective engagement with digital behavior change interventions". In: *American journal of preventive medicine* 51.5 (2016), pp. 833–842.

[329] Pedro Sanches et al. "HCI and Affective Health: Taking stock of a decade of studies and charting future research directions". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, 1–17. ISBN: 9781450359702.

DOI: 10.1145/3290605.3300475. URL: https://doi.org/10.1145/3290605.3300475.

[330] Kevin Doherty and Gavin Doherty. "Engagement in HCI: Conception, Theory and Measurement". In: *ACM Comput. Surv.* 51.5 (Nov. 2018). ISSN: 0360-0300. DOI: 10.1145/3234149. URL: https://doi.org/10.1145/3234149.

[331] Camille Nadal et al. "Patient Acceptance of Self-Monitoring on a Smartwatch in a Routine Digital Therapy: A Mixed-Methods Study". In: *ACM Trans. Comput.-Hum. Interact.* 31.1 (Nov. 2023). ISSN: 1073-0516. DOI: 10.1145/3617361. URL: https://doi.org/10.1145/3617361.

[332] Janghee Cho et al. "Reflection in Theory and Reflection in Practice: An Exploration of the Gaps in Reflection Support among Personal Informatics Apps". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3501991. URL: https://doi.org/10.1145/3491102.3501991.

[333] Elisabeth T. Kersten-van Dijk et al. "Personal Informatics, Self-Insight, and Behavior Change: A Critical Review of Current Literature". In: *Hum.-Comput. Interact.* 32.5–6 (Nov. 2017), 268–296. ISSN: 0737-0024. DOI: 10.1080/07370024.2016.1276456. URL: https://doi.org/10.1080/07370024.2016.1276456.

[334] KS Jacob et al. "Mental health systems in countries: where are we now?" In: *The Lancet* 370.9592 (2007), pp. 1061–1077.

[335] Sudhir K Khandelwal et al. "India mental health country profile". In: *International review of psychiatry* 16.1-2 (2004), pp. 126–141.

[336] China Mills and Eva Hilberg. "The construction of mental health as a technological problem in India". In: *Critical Public Health* 30.1 (2020), pp. 41–52.

[337] Deborah Lupton. *The quantified self*. John Wiley & Sons, 2016.

[338] Eleanor R. Burgess et al. "What's In Your Kit? Mental Health Technology Kits for Depression Self-Management". In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI '25. New York, NY, USA: Association for Computing Machinery, 2025. ISBN: 9798400713941. DOI: 10.1145/3706598.3713585. URL: https://doi.org/10.1145/3706598.3713585.

[339] Philip Schmidt et al. "Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection". In: *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ICMI '18. New York, NY, USA: Association

for Computing Machinery, Oct. 2018, pp. 400–408. ISBN: 978-1-4503-5692-3. DOI: 10.1145/3242969.3242985. URL: https://dl.acm.org/doi/10.1145/3242969.3242985 (visited on 01/10/2025).

[340] Jun-Zhi Xiang et al. "A multi-modal deep learning approach for stress detection using physiological signals: integrating time and frequency domain features". In: *Frontiers in Physiology* 16 (2025), p. 1584299.

[341] Rita Meziati Sabour et al. "UBFC-Phys: A Multimodal Database For Psychophysiological Studies of Social Stress". In: *IEEE Transactions on Affective Computing* 14.1 (Jan. 2023). Conference Name: IEEE Transactions on Affective Computing, pp. 622–636. ISSN: 1949-3045. DOI: 10.1109/TAFFC.2021.3056960. URL: https://ieeexplore.ieee.org/document/9346017/?arnumber=9346017 (visited on 01/10/2025).

[342] Ehsanul Haque Nirjhar and Theodora Chaspari. "Modeling Gold Standard Moment-to-Moment Ratings of Perception of Stress from Audio Recordings". In: *IEEE Transactions on Affective Computing* (2024). Conference Name: IEEE Transactions on Affective Computing, pp. 1–18. ISSN: 1949-3045. DOI: 10.1109/TAFFC.2024.3435502. URL: https://ieeexplore.ieee.org/document/10614869 (visited on 01/10/2025).

[343] Sudarshan Pant et al. "PhyMER: Physiological Dataset for Multimodal Emotion Recognition With Personality as a Context". In: *IEEE Access* 11 (2023). Conference Name: IEEE Access, pp. 107638–107656. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3320053. URL: https://ieeexplore.ieee.org/document/10265252/?arnumber=10265252 (visited on 01/10/2025).

[344] Win-Ken Beh et al. *MAUS: A Dataset for Mental Workload Assessmenton N-back Task Using Wearable Sensor*. arXiv:2111.02561 [eess]. Nov. 2021. DOI: 10.48550/arXiv.2111.02561. URL: http://arxiv.org/abs/2111.02561 (visited on 01/10/2025).

[345] Valentina Markova, Todor Ganchev, and Kalin Kalinkov. "CLAS: A Database for Cognitive Load, Affect and Stress Recognition". In: *2019 International Conference on Biomedical Innovations and Applications (BIA)*. Nov. 2019, pp. 1–4. DOI: 10.1109/BIA48344.2019.8967457. URL: https://ieeexplore.ieee.org/document/8967457/?arnumber=8967457 (visited on 01/10/2025).

[346] Christoph Anders et al. "Unobtrusive measurement of cognitive load and physiological signals in uncontrolled environments". en. In: *Scientific Data* 11.1 (Sept. 2024). Publisher: Nature Publishing Group, p. 1000. ISSN: 2052-4463. DOI: 10.1038/s41597-024-03738-7. URL: https://www.nature.com/articles/s41597-024-03738-7 (visited on 01/10/2025).

[347] Tong Xue et al. “CEAP-360VR: A Continuous Physiological and Behavioral Emotion Annotation Dataset for 360ˆ\circ VR Videos”. In: *IEEE Transactions on Multimedia* 25 (2023). Conference Name: IEEE Transactions on Multimedia, pp. 243–255. ISSN: 1941-0077. DOI: 10.1109/TMM.2021.3124080.

[348] João Areias Saraiva et al. “Scientisst move: Annotated wearable multimodal biosignals recorded during everyday life activities in naturalistic environments”. In: *Circulation* 101 (2023), e215–e220.

[349] Ary L Goldberger et al. “PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals”. In: *circulation* 101.23 (2000), e215–e220.

[350] Matias Laporte, Martin Gjoreski, and Marc Langheinrich. “LAUREATE: A Dataset for Supporting Research in Affective Computing and Human Memory Augmentation”. en. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7.3 (Sept. 2023), pp. 1–41. ISSN: 2474-9567. DOI: 10.1145/3610892. URL: https://dl.acm.org/doi/10.1145/3610892 (visited on 01/10/2025).

[351] Alexander Heimerl et al. *ForDigitStress: A multi-modal stress dataset employing a digital job interview scenario*. arXiv:2303.07742 [cs]. Mar. 2023. DOI: 10.48550/arXiv.2303.07742. URL: http://arxiv.org/abs/2303.07742 (visited on 01/10/2025).

[352] Linda Becker, Alexander Heimerl, and Elisabeth André. “ForDigitStress: presentation and evaluation of a new laboratory stressor using a digital job interview-scenario”. In: *Frontiers in Psychology* 14 (2023), p. 1182959.

[353] Ramesh Kumar Sah et al. “Adarp: A multi modal dataset for stress and alcohol relapse quantification in real life setting”. In: *2022 IEEE-EMBS International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE. 2022, pp. 1–4.

[354] Parastoo Alinia et al. “Associations between physiological signals captured using wearable sensors and self-reported outcomes among adults in alcohol use disorder recovery: development and usability study”. In: *JMIR Formative Research* 5.7 (2021), e27891.

[355] Wonse Jo et al. *MOCAS: A Multimodal Dataset for Objective Cognitive Workload Assessment on Simultaneous Tasks*. arXiv:2210.03065 [cs]. June 2024. DOI: 10.48550/arXiv.2210.03065. URL: http://arxiv.org/abs/2210.03065 (visited on 01/10/2025).

[356] Andrea Hongn et al. “Wearable Physiological Signals under Acute Stress and Exercise Conditions”. In: *Scientific Data* 12.1 (2025), p. 520.

[357] Timnit Gebru et al. “Datasheets for Datasets”. In: *CoRR* abs/1803.09010 (2018). arXiv: 1803.09010. URL: http://arxiv.org/abs/1803.09010.

[358] Ganeshan Malhotra et al. “Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations”. In: *Proceedings of the fifteenth ACM international conference on web search and data mining*. 2022, pp. 735–745.

[359] Aaron Hurst et al. “Gpt-4o system card”. In: *arXiv preprint arXiv:2410.21276* (2024).