Investigating the Use of Physiological and Behavioral Signals to Facilitate Empathic Human-AI Interaction for Daily Stress Management

Poorvesh Dongre

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Computer Science and Applications

Denis Gračanin, Chair

Benjamin Knapp

Sang Won Lee

John Richey

Mark Billinghurst

01 December 2025

Blacksburg, Virginia

Keywords: Physiological Computing, Biocybernetic Adaptation, Affect Recognition, Stress Detection, Machine Learning (ML), Deep Learning (DL), Large Language Models (LLMs)

Investigating the Use of Physiological and Behavioral Signals to Facilitate Empathic Human-AI Interaction for Daily Stress Management

Poorvesh Dongre

(ABSTRACT)

This dissertation explores the design and evaluation of Empathic Large Language Models (EmLLMs) for general mental health support. EmLLMs use physiological and behavioral signals to infer users' mental states (affective and cognitive) and accordingly generate empathic messages as adaptive interventions. Three core research goals guided this work: (1) systematically reviewing state-of-the-art methods for stress and affect recognition with physiological signals and for designing physiologically adaptive systems, (2) developing and evaluating physiology-driven EmLLM prototypes that integrate stress detection with LLM-based dialogue for stress intervention, and (3) evaluating the performance and stability of multimodal LLMs using behavioral signals for emotion recognition and supportive message generation. Findings from the systematic review highlight that physiological signals provide valuable insights into stress and affect, and that systems with physiology-driven adaptation are effective at improving both user experiences and mental health interventions. Autoethnographic and pilot studies with graduate students on different prototypes of physiology-driven EmLLMs demonstrate promise for daily stress management, and expert evaluations provide further insights into refining the design of physiology-driven EmLLMs for real-world and clinical use. Performance and stability evaluations of multimodal LLMs show that multimodal behavioral inputs, including voice and facial features, enhance emotion recognition and reasoning. However, model behavior varies across modalities, underscoring the need for

robust evaluation, customization strategies, and protective safeguards for mental health applications. Overall, this dissertation offers a systematic review, empirical insights, and design guidelines for developing empathic, engaging, and effective digital mental health systems.

Investigating the Use of Physiological and Behavioral Signals to Facilitate Empathic Human-AI Interaction for Daily Stress Management

Poorvesh Dongre

(GENERAL AUDIENCE ABSTRACT)

Managing stress and emotional well-being is a growing challenge, especially for students and working adults. This dissertation explores how new forms of Artificial Intelligence (AI) can better understand people's emotions and support their mental health. These systems go beyond traditional digital mental health tools by using physiological signals (e.g., heart rate or skin conductance) and behavioral cues (e.g., voice or facial expressions) to estimate when someone may be stressed or overwhelmed and to respond with supportive, personalized messages. This work has three main parts. First, it reviews current scientific methods for detecting stress and emotion using physiological data and examines how technology can adapt to users' emotional states. Second, it introduces and tests several prototypes that combine physiological sensing with LLM chatbots to help graduate students reflect on and manage daily stress. Third, it evaluates how well the latest multimodal AI models can process behavioral cues to detect emotions and generate empathic responses for mental health support. Across studies, this research shows that physiological and behavioral signals can meaningfully reveal emotional patterns and that AI systems that incorporate these signals can improve user experience and emotional support. However, it also finds that AI behavior can vary across input types, underscoring the importance of careful testing, customization, and safety protections when these systems are used for mental health applications. Overall, this research provides new insights, tools, and design guidelines for creating AI systems that

are not only intelligent but also sensitive, supportive, and safe to use in everyday mental health contexts.

Dedication

Dedicated to Dr. B. R. Ambedkar

Acknowledgments

This work would not have been possible without the guidance and support of my advisor, Dr. Denis Gracanin, and my committee members. Their guidance have been instrumental in achieving my research goals. I want to extend my loving gratitude to my family and friends for their unwavering support throughout my academic journey. I would like to express my sincere thanks to Dr. Mark Billinghurst, Dr. Kunal Gupta, Dr. Santosh Kumar, Dr. Sameer Neupane, Dr. Shiwali Mohan, and Dr. Christian Webb for research collaborations. Finally, I thank all my lab members at the Distributed Virtual Environments (DVE) Lab, several faculty and staff at Virginia Tech for their support.

Contents

List of Figures

List of Tables

List of Abbreviations

BCA Biocybernetic Adaptation

DL Deep Learning

ECG Electrocardiogram

EDA Electrodermal Activity

EMG Electromyography

HBI Human-Building Interaction

HCI Human-Computer Interaction

LLMs Large Language Models

ML Machine Learning

MLLM Multimodal LLM

PhyComp Physiological Computing

PPG Photoplethysmography

RESP Respiration Rate

ST Skin Temperature

Chapter 1

Introduction

Human psychological states significantly impact their mental and physical health. Psychological states refer to affective and cognitive states that influence how people perceive, think, and feel. Affective states involve emotions and moods, while cognitive states encompass mental processes such as attention, memory, and workload. Negative affective states such as stress and anxiety can lead to adverse health outcomes, including cardiovascular diseases, immune system dysfunction, and mental health disorders [12]. Negative cognitive states, characterized by high mental workload, can impair cognitive functions such as attention, memory, and decision-making, potentially leading to mental fatigue [146].

Human psychological states also impact human interaction with computers and other automated systems. Users experiencing negative affective states, such as anger and frustration, when interacting with a system may stop using it [25]. Systems that exacerbate mental workload can overwhelm users, decreasing efficiency and increasing the likelihood of mistakes or accidents [153]. Therefore, monitoring human psychological states can also improve human interaction with computers and other automated systems.

A range of methods have been employed to assess psychological states. Traditional techniques, such as standardized questionnaires, Ecological Momentary Assessments (EMA), and structured interviews, rely on individuals to consciously reflect on and report their affective and cognitive states. While these traditional techniques are valuable for capturing subjective experiences, they are often limited by recall bias, social desirability effects, and low temporal resolution.

Physiological and behavioral approaches offer complementary perspectives for inferring affective or cognitive states by analyzing body signals such as Photoplethysmography (PPG), Electrodermal Activity (EDA), and Skin Temperature (ST), and behavioral cues, such as facial expressions, vocal prosody, speech patterns, eye gaze, posture, and gestures. Physiological sensing involves measuring and analyzing human body signals that serve as objective indicators of the body's internal states. They may function as outcomes, markers, or even correlates of psychological states, depending on the context and the strength of their relationship [28]. Physiological signals are increasingly used to support both digital health and effective interaction with computers and automated systems [23]. In digital health, continuous monitoring of users' physiological signals enables the detection of clinical conditions and the delivery of timely, personalized interventions [199]. In Human-Computer Interaction (HCI), physiological signals are primarily used to evaluate user experience, but they are also used to develop physiologically adaptive systems [70, 127].

This dissertation explores the use of physiological sensing not only to monitor and intervene in users' daily stress levels but also to enhance the user experience of digital mental health tools to support daily stress management. Stress is a widespread and persistent challenge in modern life. Digital mental health tools, including wearables that collect physiological signals to monitor stress and mental health, as well as chatbots that provide stress interventions, have shown promise in managing daily stress. However, these tools face several challenges related to user personalization and engagement, leading to high user drop-off rates and, ultimately, low clinical impact. To address these gaps, this research integrates physiological sensing and Large Language Model (LLM) chatbots to create a physiology-driven system for stress management that (i) monitors and infers user stress using physiological signals and (ii) provides empathic interactions with an LLM chatbot. We also investigate the use of behavioral cues, including textual, vocal, and facial features, to facilitate empathic interactions.

1.1 Background

This section provides essential background on related topics, including the fundamentals of human physiology and physiological sensing; stress and affect recognition using physiological signals; the impact of stress and affective states on mental health; an overview of digital mental health tools; and a brief introduction to physiological computing systems.

1.1.1 Human Physiology and Physiological Sensing

Human physiology is the scientific study of the human body's functions, including the mechanical, physical, biochemical, and bioelectric processes. It focuses on how the body's systems, including the nervous, cardiovascular, respiratory, gastrointestinal, and musculoskeletal systems, interact with one another to respond to various stimuli and maintain homeostasis. The functioning of these systems can be objectively measured using physiological signals.

The nervous system plays a central role in coordinating the body's systems. It is categorized into the Central Nervous System (CNS), which includes the brain and spinal cord, and the Peripheral Nervous System (PNS), which connects the CNS to the rest of the body. A key component of the PNS is the Autonomic Nervous System (ANS), which regulates involuntary body functions, including Heart Rate (HR), Respiration Rate (RR), and digestion. The ANS is further categorized into two key components: the sympathetic and parasympathetic nervous systems. The sympathetic nervous system prepares the body for stressful situations by constricting blood vessels and increasing HR and RR. The parasympathetic nervous system promotes relaxation and energy conservation by relaxing the muscles and reducing HR and RR.

Physiological signals that capture involuntary PNS responses and provide insight into ANS function are typically referred to as peripheral physiological signals. These signals comprise measures of the heart, including PPG and Electrocardiogram (ECG); measures of the skin, including EDA and ST; and measures of the muscle, including Electromyogram (EMG). In contrast, physiological signals of the CNS focus on capturing neural activity and brain function. These include the Electroencephalogram (EEG), which measures brain electrical activity; Functional Magnetic Resonance Imaging (fMRI), which tracks changes in cerebral blood oxygenation; and Functional Near-Infrared Spectroscopy (fNIRS), which monitors cerebral blood flow. The measurement of physiological signals, both peripheral and neural, for inferring an individual's physical and psychological state is referred to as physiological sensing [23, 42, 70, 71, 73].

1.1.2 Stress and Affect Recognition using Physiological Signals

Stress is the body's natural response to any internal or external event that disrupts homeostasis and can be categorized in several ways. Based on frequency, stress can be categorized as acute (a short-term response to immediate pressures), episodic acute (repeated acute stress episodes), and chronic (long-term, persistent stress that can significantly harm mental and physical health). Based on sentiment, stress can be categorized as eustress and distress. Eustress is positive stress associated with motivation and excitement, whereas distress is negative stress associated with unpleasant experiences. Stress can also be categorized by origin: psychological stress arises from emotional or cognitive strain; physical stress stems from bodily demands or environmental factors; and psycho-physiological stress refers to the body's integrated psychological and physiological response to perceived challenges or threats. In this research, we focus on distress and psycho-physiological stress; therefore, any subsequent references to "stress" should be understood as referring to these two forms.

Affect is a broad term encompassing emotional states, moods, and other subjective feelings. It refers to the external expression of internal experiences and is often measured along dimensions such as valence and arousal [208]. Valence refers to the pleasantness or unpleasantness of an experience, ranging from negative to positive. Arousal refers to the intensity or activation level of an experience, ranging from calm to excited. Stress and affect are closely related but distinct constructs. Both involve experiences that can trigger psycho-physiological responses. However, stress results from perceived threats and challenges and is associated with negative valence and high arousal. Whereas affective states can arise spontaneously and encompass a broader range of experiences, with varying levels of valence and arousal [208]. In essence, all stress is a form of affect, but not all affect constitutes stress.

Recent research has made significant strides in stress and affect recognition by leveraging various modalities, including facial expressions, speech, text, and their multimodal combinations using Machine Learning (ML) and Deep Learning (DL) techniques. For instance, Mollahosseini et al. [170] used Convolutional Neural Networks (CNNs) to extract complex features from facial expressions to classify discrete emotions and valence and arousal levels. Eyben et al. [69] extracted prosodic features from speech, including pitch, jitter, and shimmer, and applied ML and DL methods to classify discrete affective states. Tzirakis et al. [233] combined audio-visual inputs, using CNNs to process visual data and Recurrent Neural Networks (RNNs) for audio signals, to estimate valence and arousal. Poria et al. [196] used CNNs and RNNs to model both spatial and sequential patterns in conversational language for text-based emotion recognition. A limitation of using these modalities for stress and affect recognition is that users can intentionally mask or exaggerate their expressions, thereby reducing the reliability of the inferences.

A key advantage of physiological signals is their close connection to the ANS, making them difficult to control or manipulate consciously. This characteristic enhances their reliability in stress and affect recognition, particularly in situations where individuals are unwilling or unable to express themselves overtly. However, the relationship and causality between physiological signals and psychological states, including stress and affect, have long been debated in the existing literature. Cacioppo et al. [28] and Bradley and Lang [24] emphasized that physiological responses can be meaningfully associated with psychological states, but only under carefully controlled experimental conditions. Kreibig et al. [134] cautioned that even under such conditions, the correlations between physiological changes and psychological states are not always consistent.

Early works on stress and affect recognition using physiological signals were foundational but had significant limitations. For instance, studies by Picard et al. [193] and Haag et al. [99] used physiological signals from only one subject, limiting the generalizability of their findings. However, with advances in physiological sensing technologies, several publicly available datasets now include data from multiple participants exposed to diverse experimental or real-world conditions. For example, the Dataset for Emotion Analysis using Physiological signals (DEAP) [131] integrated EEG signals with other physiological signals of 32 participants to create a multimodal database for affect recognition. The Wearable Stress and Affect Detection (WESAD) dataset [212] was collected from 8 participants exposed to diverse stimuli and captures multimodal physiological signals. The Mobile Open Observation of Daily Stressors (MOODS) dataset [108] collected physiological signals from 122 participants in naturalistic, everyday environments. These datasets have enabled researchers to develop models for recognizing stress and affect, achieving increasingly robust and accurate performance.

1.1.3 Impact of Stress and Affect on Mental Health

Stress and affect play a critical role in shaping human psychology and behavior, thereby profoundly influencing mental health. As discussed, eustress is a form of positive stress that can enhance motivation and excitement. Similarly, positive affective states, such as joy, gratitude, and contentment, are strongly associated with psychological resilience and serve as protective factors against the development of mental health disorders [82, 83]. Positive stress and affective states not only help in maintaining good mental health but also promote physical health [198].

In contrast, chronic stress and sustained experiences of negative affective states, such as sadness, fear, and anxiety, are closely linked to adverse mental health outcomes. For instance, chronic stress has been shown to elevate cortisol levels and impair emotional regulation mechanisms [155]. It can also narrow cognitive focus, hinder problem-solving abilities, and increase vulnerability to negative thought patterns [144]. Chronic sadness may evolve into depressive symptoms [125], while persistent fear and anxiety can lead to hyper-vigilance and an overactive stress response [195]. Therefore, continuous monitoring and effective management of stress and affective states are essential for supporting and maintaining mental health.

1.1.4 Digital Mental Health Tools

Digital mental health tools encompass a wide range of solutions, including wearables (mobile apps paired with wearables and standalone wearable apps) and mental health chatbots designed to support mental health and well-being. Wearables have gained prominence because they passively and continuously collect user data, primarily physiological signals, to infer users' stress and affect and provide just-in-time interventions. During the past two decades, the scientific community has made significant advances in using wearables for stress and affect detection [91, 108, 183, 242]. Wearable data has been leveraged to create innovative visualizations for self-reflection [129, 130, 209, 223, 247] and delivered real-time interventions [19, 109]. There are also wearable apps that provide biofeedback, helping users engage in relaxation or mindfulness practices and improve emotional regulation [76].

Mental health chatbots provide support through human-like, natural-language conversations. The origins of mental health chatbots can be traced back to ELIZA, one of the earliest programs developed to simulate human conversation through pattern matching and scripted responses [240]. Since then, a range of rule-based and semi-automated chatbots have emerged, offering psychoeducation, coping strategies, and therapeutic dialogue. For instance, Woebot is a fully automated chatbot based on cognitive behavioral therapy (CBT) principles [78]. Wysa supports mental health through empathy-focused conversations [111].

Recent studies have observed a notable shift from traditional rule-based systems to natural language processing (NLP)-powered ML mental health chatbots. They enable more dynamic, responsive interactions by adapting their responses based on user input, emotional tone, and conversational context [1]. This transition has been further accelerated by the development of LLMs, which demonstrate remarkable capabilities in understanding and generating human-like language. LLM-powered mental health chatbots engage in more nuanced, empathic, and contextually aware conversations [104].

1.1.5 Physiological Computing

Technological systems that incorporate users' physiological signals into their functionality to improve user experience are called Physiological Computing (PhyComp) systems [73]. Such systems leverage users' physiological signals to create an alternate communication channel and facilitate bi-directional interactions between the user and the system, making them more intuitive and responsive [71]. PhyComp systems can be broadly categorized into interface control-type systems and systems with biocybernetic adaptation [74].

In interface control-type systems, users' physiological signals are directly translated into meaningful commands, enabling them to control an external device. An example of such systems is Brain-Computer Interfaces (BCIs), in which the brain's electrocortical activity is translated into a sequence of commands that control the interface. The operationalization of interface control-type systems depends on user intentionality, meaning the user intends to achieve a specific selection or outcome. These systems offer hands-free interaction, which is particularly beneficial for individuals with mobility impairments or for enhancing accessibility in complex environments.

In systems with biocybernetic adaptation, users' physiological signals are continuously monitored to infer their psychological states, and the system dynamically adjusts its behavior to enhance user engagement, performance, or well-being. Examples of such systems include emotion-adaptive games that adjust the game difficulty to increase engagement [251] and stress-adaptive training platforms that adjust the task difficulty to improve performance [221]. Biocybernetic adaptations are a derivative of biofeedback, which provides real-time feedback on physiological signals to users, enabling them to regulate their physiology consciously. The unique feature that distinguishes these systems from biofeedback and interface-control systems is that physiological signals drive system adaptations without requiring the user's intentional input.

1.2 Motivation

The motivation for this research stems from existing challenges in receiving professional mental health services and limitations of wearables and mental health chatbots in providing mental health support. This section elaborates on these challenges, laying the groundwork for our motivation to develop the proposed physiology-driven system for daily stress management among graduate students.

1.2.1 Stress among Graduate Students

As discussed, stress is a state of emotional or mental burden caused by adverse or demanding circumstances. While eustress can enhance focus, motivation, and performance, chronic distress poses serious risks to both mental and physical health. High levels of distress have become a pervasive and persistent issue, particularly among graduate students who often face high academic expectations, financial insecurity, uncertain career prospects, and a lack of work-life balance. Several studies have highlighted the growing concern about stress and mental health among graduate students. For instance, a global survey conducted in 2019 found that 36% of Ph.D. students sought help for anxiety or depression related to their academic work [175]. Moreover, a systematic review and meta-analysis of studies published through 2019 revealed that 24% of Ph.D. students reported clinically significant symptoms of depression, while 17% exhibited symptoms of anxiety [210].

Stress among Graduate Students at Virginia Tech

Further evidence of stress among graduate students at Virginia Tech is provided by a study conducted as part of this research, which was initially aimed at validating a conceptual framework for Human-Building Interaction (HBI) [205]. The framework uses sensor technologies to enhance occupant experience and well-being. It involves (1) identifying the building type and its primary occupants, (2) evaluating occupants' needs, (3) using sensor systems to monitor aspects of occupants' needs, (4) analyzing the sensor data to make meaningful inferences, and (5) making necessary adjustments in the building environment to address occupant needs. The main components of the framework include (1) sensor systems, (2) processing systems, and (3) adaptive systems. The sensor systems monitor occupants and their environment, the processing system uses various inference models to infer occupant states and environmental conditions, and the adaptive systems adjust the environment based on these inferred states and conditions.

To validate the conceptual framework, focus groups and semi-structured interviews were conducted with 19 participants, 12 undergraduate and 7 graduate students, who regularly spent more than 10 hours per week in a smart academic building at Virginia Tech. All participants were asked semi-structured questions to assess their needs, concerns, and experiences within the building. All sessions were audio-recorded and transcribed verbatim for analysis. The qualitative data were analyzed using a combination of inductive and deductive coding approaches. The deductive codes were derived from constructs in existing HBI theories and frameworks [8]. The thematic analysis revealed five key themes: navigation, learning, comfort, stress management, and safety.

Among these, "stress management" emerged as a prominent concern for graduate students. They highlighted spending a significant amount of time in the academic building and sought some form of stress intervention to manage their everyday stress. Despite these concerns, students often struggle to seek mental health support due to the limited availability and accessibility of mental health services.

1.2.2 Availability and Accessibility of Mental Health Services in Universities

In the U.S., most universities maintain on-campus counseling centers staffed by licensed mental health professionals. These centers provide individual counseling, crisis intervention, group therapy, and preventive wellness programming. However, demand for mental health services in universities is increasing, with several prominent universities falling short of recommended benchmarks [85]. This shortage translates into longer wait times, exacerbating the situation for many students already experiencing mental health concerns [245]. It also raises concerns about the continuity and effectiveness of treatment.

Although the overall percent usage of mental health services available for students has increased, disparities persist across race, gender, and socioeconomic groups. Underrepresented minority and international students remain less likely to use campus services despite comparable or higher levels of distress [245]. This is due to fear of harming their careers, financial insecurity, and limited time [232]. The situation in developing countries such as India and South Africa is far more difficult, with many universities lacking dedicated counseling centers [114]. Moreover, the stigma associated with mental health issues in such countries further suppresses students from seeking help.

1.2.3 Limitations of Digital Mental Health Tools

Digital mental health tools offer a promising solution by addressing affordability and accessibility challenges for graduate students. As discussed, these tools encompass a wide range of technological solutions, including wearables and mental health chatbots. Wearables can monitor users' stress and provide just-in-time interventions. Mental health chatbots can provide support through human-like, natural-language conversations. Despite their promise, digital mental health tools face several limitations.

Limitations of Wearables

Wearables have limitations in detecting stress in individual users. Approaches to inferring stress from wearable data, primarily physiological signals, involve preprocessing the signals, extracting relevant features, and applying machine learning models to the extracted features to map them to stress labels [70]. Physiological signals are sparse and noisy, requiring preprocessing to impute missing values and remove artifacts. Feature extraction is another challenge, as the physiological signals require carefully designed features to capture stress-related patterns [212]. Signal variability across users due to individual differences in psychophysiological reactivity and lifestyle further complicates stress detection [112, 228]. Moreover, the use of static labels oversimplifies stress as a categorical construct.

Wearables for stress monitoring and intervention also have limitations related to user experience. Many wearables provide limited user interaction, offering redundant or generic feedback that does not adapt to the user's context or changing needs. Such feedback can reduce perceived usefulness and lead to disengagement over time [52]. In addition, repeated notifications may contribute to intervention fatigue, where users become desensitized to alerts or abandon the app altogether [174]. The lack of personalization in both monitoring and intervention delivery limits the user experience and long-term engagement [163]. Lastly, without transparent explanations of stress inferences or meaningful interaction mechanisms, users may develop mistrust in the system's accuracy and stop using it [228].

Limitation of Mental Health Chatbots

Mental health support chatbots are promising tools for stress intervention as they can provide accessible, nonjudgmental, and cost-effective assistance. However, the drop-off rates among mental health chatbots are high. A key limitation is the lack of therapeutic alliance, as many users find conversations with chatbots shallow, impersonal, and lacking genuine empathy [4]. Limited personalization further undermines effectiveness, with many systems delivering generic or repetitive responses that do not adapt to individual stress profiles, contextual factors, or user preferences [2, 109, 231]. Narrow interaction modalities also constrain user engagement, as most chatbots rely solely on text-based input and output. Moreover, the absence of dynamic engagement features, such as interactive prompts, longitudinal tracking, and context-aware reminders, reduces motivation for continued use, leaving the responsibility of maintaining the interaction entirely on the user [22].

1.3 Proposed Approach

Our approach integrates wearables and mental health chatbots to design and develop a system that addresses the key limitations of each. By leveraging chatbots, the system overcomes the limitations of wearables by transforming psychophysiological inferences into personalized and empathic feedback, reducing redundant notifications, mitigating intervention fatigue, and fostering sustained user engagement through adaptive dialogue. By leveraging wearables, the system overcomes the limitations of mental health chatbots by providing an additional modality to objectively and continuously infer users' stressful states, enabling the chatbot to move beyond shallow, text-only interactions toward more empathic, adaptive, and personalized conversations. To summarize, we propose a system that continuously infers user stress from wearables and integrates these inferences into chatbot interactions to improve user experience and promote effective stress management.

This research also examines the application of predictive and generative AI, specifically Deep Learning (DL) and Large Language Models (LLMs), in developing the integrated system. DL techniques can overcome the limitation of stress inference from physiological signals by learning robust representations directly from raw or minimally processed data, reducing the reliance on hand-crafted features. They enable the development of personalized models that account for inter-individual variability in physiology, lifestyle, and stress reactivity. In addition, self-supervised and multitask learning approaches enable DL models to leverage large-scale unlabeled data and to jointly learn related affective states, thereby improving generalization across users and contexts.

LLMs, on the other hand, show greater promise for enhancing user engagement than traditional rule-based chatbots [104]. Unlike earlier systems that often deliver scripted or generic responses, LLM chatbots use advanced NLP techniques to infer user intent and emotional state, enabling more contextually appropriate interactions. They can sustain human-like conversations, with the ability to "listen" and respond in a manner that feels empathic and supportive. Moreover, LLMs adapt dynamically by remembering past conversations, adjusting the tone, style, or complexity of responses, and tailoring support over time to the individual's evolving needs. These capabilities make LLM chatbots more engaging, more responsive to user context, and better equipped to build trust and rapport.

1.4 Research Question, Goals, and Tasks

Based on the discussion so far, the overarching research question can be framed as:

How can physiological and behavioral signals be integrated with LLMs to design systems that improve user experience and intervention effectiveness in daily stress management?

This broad research question can be further categorized into the following research goals and tasks, which address both the user-centered and technical limitations discussed.

Research Goal 1 (RG1): Conduct a systematic literature review to understand the latest trends in using physiological signals for stress and affect recognition, and designing systems with physiology-driven/biocybernetic adaptation.

This research goal can be further categorized into the following tasks:

Task 1: Psycho-physiological Stress and Affect Recognition- Systematically review the latest research on using physiological signals for stress and affect recognition to identify commonly used signals, pre-processing steps, feature extraction methods, modeling approaches, and validation strategies.

Task 2: Design of Biocybernetic Adaptations- Systematically review the existing systems with physiology-driven/biocybernetic adaptation to identify their applications, design features, including adaptation medium and mechanism, and evaluation procedures.

Task 3: Deriving Design Guidelines- Synthesize findings from Tasks 1 and 2 to derive design guidelines that inform the design and development (RG2) of a physiology-driven system for daily stress management.

Research Goal 2 (RG2): Design, implement, and evaluate digital mental health systems that integrate physiological signals with LLMs to monitor user stress and intervene by adapting their behavior based on the stress level.

This research goal can be further categorized into the following tasks:

Task 4: Integrate and Evaluate Existing Tools- Manually integrate an off-the-shelf wearable for stress monitoring and LLMs chatbots for stress intervention to establish a baseline integrated system. Evaluate its user experience and effectiveness to inform the design of a purpose-built system.

Task 5: Design and Develop an Integrated System- Design and develop an integrated system that integrates physiological signals from wearables with LLM chatbots for daily stress management. This task can be further categorized into the following subtasks.

Subtask 1: System Architecture Design- Design the system architecture, including a physiological signal acquisition pipeline, stress recognition models, LLM-based dialogue management, adaptation mechanisms, and user interface.

Subtask 2: Model Development and Personalization- Develop stress detection models from physiological data and implement personalization strategies to account for inter-individual variability in stress responses.

Task 6: Technical and Usability Evaluation- Conduct user studies to evaluate the technical and user-centered aspects of the integrated system. This task can be further categorized into the following subtasks.

Subtask 3: Pilot Study with Graduate Students- Conduct a preliminary evaluation with graduate students to evaluate the system's stress recognition accuracy, mental health support, and user experience in a real-world setting.

Subtask 4: Design Evaluation with Mental Health Experts- Conduct user studies with mental health experts to assess the design functionality and limitations of the integrated system.

Research Goal 3 (RG3): Investigate the complementary role of multimodal data

(audio, visual, text) in enhancing affect recognition and adaptive interventions within integrated AI systems for mental health support.

This research goal can be further categorized into the following tasks:

Task 7: Multimodal Affect Recognition- Evaluate the performance and stability of multimodal LLMs for affect recognition, reasoning, and empathic message generation from combinations of audio, visual, and text data.

Chapter 2

Review of Literature

In the last chapter, the concepts of physiological computing and biocybernetic adaptations were briefly introduced. We also discussed in detail the motivation behind designing a digital mental health system with physiology-driven adaptations for stress management.

This chapter reviews the two main components of systems with biocybernetic adaptation: psycho-physiological inference and system adaptation. It systematically reviews the latest literature on the use of physiological signals for stress and affect recognition by identifying commonly used signals, preprocessing procedures, feature extraction methods, inference algorithms, and validation techniques. This chapter also reviews the latest literature on the design of systems with biocybernetic adaptation, identifying their design features, including the adaptation medium and mechanisms. Lastly, the evaluation procedures of such systems are examined, providing insights into the accuracy of psycho-physiological inference and the effectiveness of system adaptation.

2.1 Systematic Review Methodology

The broad research question guiding this review is: "What are the trends and best practices for designing, developing, and evaluating systems with biocybernetic adaptation, and how can they be used for everyday stress management?" This broad research question can be broken down into the following sub-questions.

RQ1 What are the common practices for stress and affect recognition from physiological signals?

(a) What are the commonly used datasets, emotion induction techniques, data collection devices, and labeling approaches?

(b) What are the commonly used physiological signals, pre-processing procedures, physiological features, and modeling methods?

RQ2 What are the common applications, design features, and evaluation procedures of systems with biocybernetic adaptation?

(a) What are the common applications and design features, including adaptation medium and mechanism?

(b) What are the commonly used evaluation techniques and the impact of system adaptation on users?

This review adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure transparency and rigor in answering the proposed research questions [186]. Covidence, an online tool for systematic review management, was used to facilitate screening, data extraction, and organization [41].

To find relevant papers, different keyword search queries were experimented with and the following query was finalized: *("physiological data" OR "physiological signals" OR "biosignals") AND ("affective computing" OR "affect recognition" OR "emotion recognition" OR "stress recognition") AND ("adaptive user interface" OR "dynamic user interface" OR "user interface adaptation").* This query was used in three databases: ACM Digital Library, IEEE Xplore Digital Library, and Scopus, and it resulted in 984 conference proceedings and journal papers in the last five and a half years (2019 to mid-2024). These papers were further

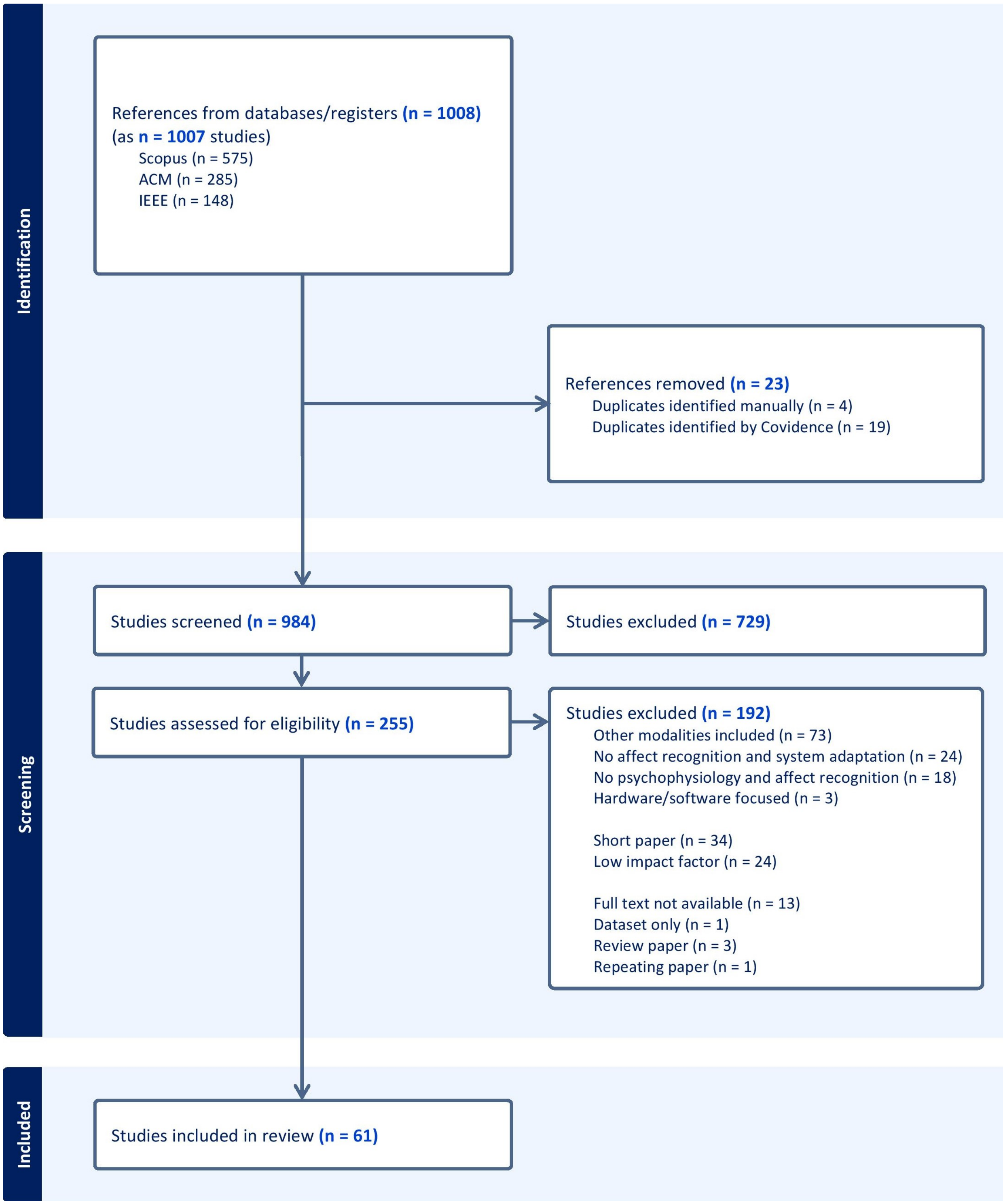


Figure 2.1: Systematic Review Flowchart

evaluated using the following inclusion and exclusion criteria to identify the most relevant papers.

Inclusion Criteria

1. The paper focuses on stress and affect recognition using peripheral physiological signals.
2. The paper focuses on system adaptations driven by peripheral physiological signals used for stress and affect.

Exclusion Criteria

1. The paper focuses on recognizing cognitive states, such as attention and cognitive workload.
2. The paper focuses on the use of neurophysiological signals, such as EEG.
3. The paper uses other modalities, such as video and audio, along with physiological signals for stress and affect recognition.
4. Short papers, extended abstracts, and posters and papers published in a conference or journal with an impact factor of less than 1.

The papers were filtered for peripheral physiological signals because they are particularly suitable for stress and affect recognition due to their strong relationship to autonomic responses [23]. Moreover, with the availability of several commercial devices, acquiring peripheral physiological signals is less obtrusive than acquiring neurophysiological signals, making them suitable for everyday use. Similarly, excluding papers that rely on other modalities allows for assessing the efficacy of physiological signals alone, eliminating confounding variables.

By excluding short papers, extended abstracts, and posters, the review focuses on comprehensive papers that provide in-depth experimental results, validated methods, and thorough analysis that can be used to draw robust and generalizable conclusions. Moreover, to maintain a high evidence standard and ensure the review's reliability, only papers published in conferences or journals with an impact factor of at least one are included.

The inclusion and exclusion criteria were first used to screen the titles and abstracts of all 984 papers, resulting in 255 papers for full review. The title and abstract screening process involved two independent reviewers, achieving a Cohen's Kappa coefficient of 0.5318, indicating moderate agreement. Later, the same criteria were followed for a full review, resulting in 61 papers for final data extraction. Two independent reviewers also participated in the full-paper screening, and Cohen's Kappa coefficient was 0.56659, reflecting moderate agreement. Lastly, the two reviewers collaborated to extract data from the final 61 shortlisted papers.

The results of data extraction revealed that the 61 shortlisted papers could be broadly categorized into three main research focus areas: (1) stress and affect recognition using physiological signals (46 papers), (2) system adaptation (6 papers), and (3) recognition and adaptation (9 papers). Papers focusing on affect recognition used proprietary and/or publicly available datasets with physiological signals to develop ML and/or DL models for affect recognition. These papers detailed preprocessing, feature extraction, modeling, and validation methods. Papers focusing on system adaptation involved designing systems with biocybernetic adaptation in domains such as training, mental health, and entertainment, and evaluating the impact of these adaptations on users. Papers focusing on recognition and adaptation presented a holistic approach that combined model development with biocybernetic adaptation and user evaluations. Figure 1 shows this categorization for the reviewed papers published from 2019 to mid-2024.

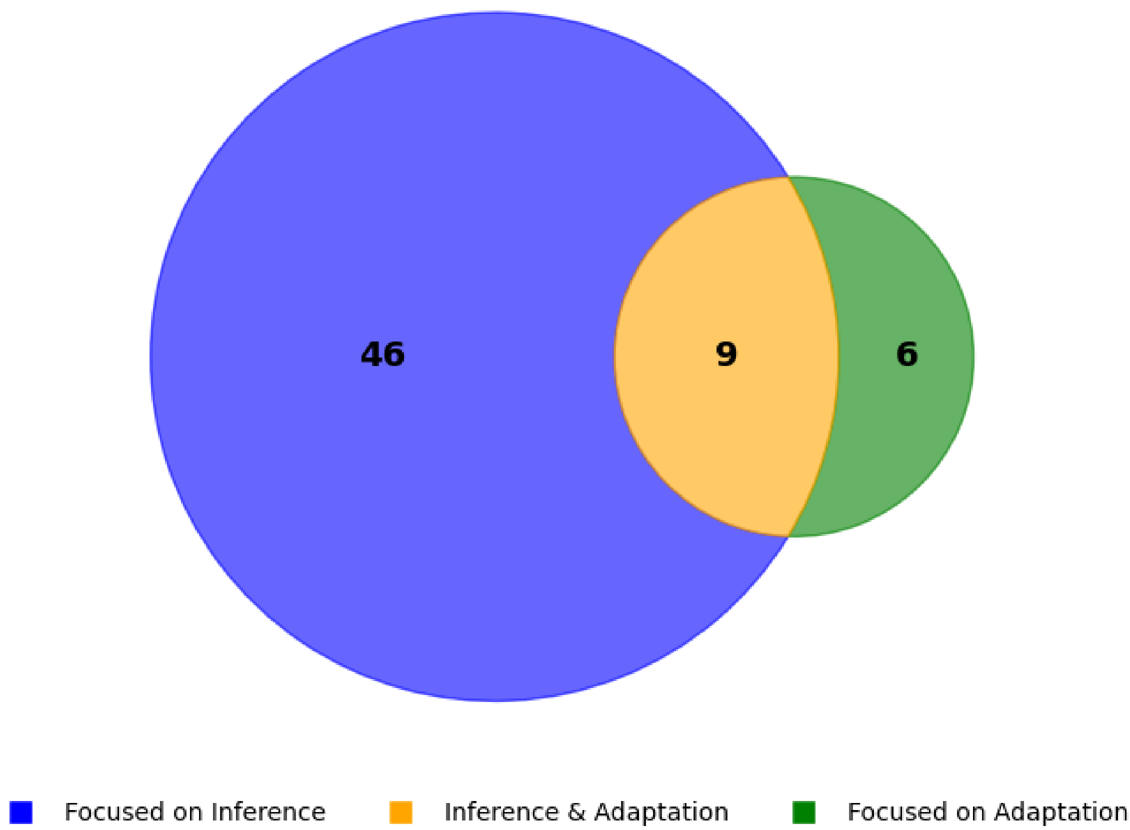


Figure 2.2: Research Focus Areas

2.2 Psycho-physiological Inference of Stress and Affect

As discussed, psycho-physiological inference refers to the process of inferring internal states from the measurement and analysis of bodily responses. It is a key component of systems with biocybernetic adaptation. In this review, the psycho-physiological inference of stress and affect across 55 papers (46 on stress and affect recognition and 9 on recognition and adaptation) was performed using ML and DL models trained on peripheral physiological signals from various private and publicly available datasets. The typical ML pipeline for psycho-physiological inference of stress and affect involves collecting physiological signals, preprocessing the signals, extracting relevant features, training ML models, and evaluating the trained models. DL can incorporate signal preprocessing and feature engineering in model development, simplifying the modeling process [113]. This section examines how the reviewed papers approached each of these components.

2.2.1 Dataset Description

High-quality datasets are essential for developing models for recognizing stress and affect from physiological signals. The quality of datasets depends on various factors, such as the stimuli used, data collection devices, and data labeling methods. Therefore, this subsection examines the quality of the psycho-physiological inference datasets used for stress and affect recognition in the reviewed papers.

Dataset Types

Of the 55 papers that used physiological signals for stress and affect recognition, 27 focused on utilizing publicly available datasets to develop the inference models. These datasets provide standardized benchmarks to facilitate cross-study comparisons in affective computing research. Commonly used publicly available datasets include WESAD (11 papers), AMIGOS (5 papers), DEAP (4 papers), MAHNOB-HCI (3 papers), K-EmoCon (3 papers), SWELL (3 papers), CASE (3 papers), driveDB (3 papers), DREAMER (2 papers), and CLAS (2 papers). Notably, several papers combined multiple publicly available datasets to improve inference performance. For instance, Zhu et al. [259] combined DREAMER, MAHNOB-HCI, and WESAD; Harper et al. [102] used both DREAMER and AMIGOS; while Karavidas et al., Vos et al., and Bhatti et al. integrated SWELL and WESAD [21, 118, 237]. Other less commonly used publicly available datasets include ASCERTAIN [225], RECOLA [204], AffectiveROAD [100], MERTI-Apps [156], and PPG-Dalia [202].

28 papers used private or proprietary datasets to develop stress and affect recognition models. Individual researchers collected these datasets to meet specific study requirements. Although these datasets enable tailored applications, they often lack the standardization required for cross-study comparison. Notably, publicly available datasets were used in papers on psycho-physiological inference, whereas private datasets were used in papers on system adaptation. It is also important to note that some reviewed papers used a combination of publicly available and private datasets for model training and testing. For example, Lee et al. [139] used the MERTI-Apps dataset to train an affect recognition model and a private dataset to customize their model. We review the quality parameters of both publicly available and private datasets.

Stimuli

Stimuli are fundamental for eliciting physiological responses. As shown in Table 2.2, video clips are the most commonly used stimuli in papers using both publicly available and private datasets. Videos combine visual and auditory elements with narrative structures and can

Table 2.1: Datasets Used for Stress and Affect Recognition

Private Datasets	**Public Datasets**									
28 papers	27 papers									
	WESAD [213]	AMIGOS [166]	DEAP [131]	MAHNOB-HCI [219]	K-EmoCon [189]	SWELL-KW [132]	CASE [216]	driveDB [103]	DREAMER [120]	CLAS [161]
[5, 11, 18, 29, 48, 72, 87, 90, 94, 98, 116, 119, 121, 122, 133, 142, 148, 149, 158, 159, 165, 184, 185, 192, 201, 220, 235, 236]	[16, 21, 53, 80, 112, 118, 135, 147, 218, 237, 259]	[35, 66, 102, 206, 215]	[66, 140, 141, 188]	[80, 206, 259]	[53, 249, 260]	[21, 118, 237]	[21, 53, 164]	[123, 162, 259]	[102, 259]	[53, 206]

depict emotionally charged situations. In publicly available datasets, including WESAD, AMIGOS, DEAP, MAHNOB-HCI, CASE, DREAMER, ASCERTAIN, and CLAS, videos were used to elicit emotions like happiness, sadness, fear, and relaxation. Moreover, videos were used in 6 papers as stimuli for private datasets.

Cognitive tasks were commonly used as stimuli to evoke stress and anxiety. WESAD uses social stress tests and cognitive tasks, such as the Trier Social Stress Test (TSST) and mental arithmetic challenges, to evoke stress and anxiety [213]. SWELL-KW utilizes work-related tasks to simulate real-world cognitive challenges that elicit varying cognitive load and stress levels [132]. Popular video games such as Tetris and PUBG were primarily used in studies that used proprietary datasets [11, 158] to elicit various affective states among users.

To elicit stress and emotion in more naturalistic environments, K-EmoCon [189] used debates on social issues, SWELL-KW [132] used simulated office environments, and PPG-Dalia [202] used a wide range of real-life activities. driveDB [103] and AffectiveROAD [100] used driving scenarios. 6 papers that used private datasets also employed natural environments, including activities of daily living, as stimuli to capture users' affective states. Other commonly used stimuli in papers with private datasets include Virtual Reality (VR) (4 papers) and audio clips (3 papers). Interestingly, we found no publicly available datasets in our review that collect physiological signals using games, VR, and audio as stimuli.

Data Collection Devices

As discussed, peripheral physiological signals reflect PNS activity. Devices that collect peripheral physiological signals are less obstrusive. They are usually easier to set up, lighter, more portable, and less costly than devices that collect CNS physiological activity. However, the devices used to collect peripheral physiological signals in the reviewed studies, across both

Table 2.2: Stimuli Used to Evoke Stress and Emotions

Stimuli	**Private Datasets**	**Public Datasets**
Video	[116, 133, 139, 142, 159, 235]	WESAD, AMIGOS, DEAP, MAHNOB-HCI, CASE, DREAMER, MERTI-Apps, ASCERTAIN, CLAS
Cognitive Tasks	[5, 48, 87, 121, 147, 201]	WESAD, SWELL-KW, CLAS, RECOLA
In-the-Wild	[29, 98, 122, 185, 220, 236]	K-EmoCon, SWELL-KW, PPG-DaLiA
Driving	[72]	driveDB, AffectiveROAD
Images	[90]	MAHNOB-HCI, CLAS
Games	[11, 94, 149, 158, 159, 184]	
VR	[94, 165, 192, 201]	
Audio	[18, 87, 133]	

public and private datasets, exhibited varying levels of obtrusiveness. Device obtrusiveness refers to the extent to which a device interferes with a user's comfort and usability due to its size, placement, cost, or required setup [127].

Obtrusive devices include the BioSemi ActiveTwo and Biopac MP150. The BioSemi ActiveTwo, commonly used in datasets such as AMIGOS, DEAP, and MAHNOB-HCI, records high-resolution EEG and peripheral signals, including ECG and EDA. Its obtrusiveness stems from the complex setup and placement of multiple electrodes, which can affect user comfort. The Biopac MP150, used in several private datasets, is versatile, capturing signals such as ECG, EDA, EMG, EEG, and RR. However, its bulky design and extensive setup time make it cumbersome for practical use. Although compact, the Shimmer3 remains obtrusive due to its configuration and sensor-placement constraints. It records ECG, EMG, EDA, and PPG and contributes to datasets like AMIGOS and DREAMER.

The Polar H10, RespiBAN, and Zephyr BioHarness are less obtrusive due to their lightweight, portable designs. These chest-worn devices are primarily used for accurate ECG monitoring and contribute to publicly available datasets, including K-EmoCon (Polar H10), WESAD (RespiBAN), and SWELL (Zephyr BioHarness). However, their requirement for specific chest placement may cause mild discomfort or limit user mobility during extended use. Despite this, they balance usability and signal accuracy, making them more practical for real-world applications than larger, more complex systems.

The Empatica E4 wristband provides unobtrusive peripheral physiological data collection and is commonly used in publicly available datasets (e.g., WESAD and SWELL) and in papers reporting on private datasets (11 papers). The Empatica E4 collects EDA, PPG, ST, and ACC data, making it suitable for controlled and naturalistic data collection. Other wristbands, such as the Microsoft Band, Biovotion Armband, and Garmin Forerunner, provide peripheral physiological data collection with minimal interference with daily activities. Table 2.3 shows all the devices used to collect physiological signals in the reviewed papers using public and private datasets.

Data Labeling

Data labeling is the process of associating raw physiological signals with the ground truth and is crucial in stress and affect recognition with supervised ML and DL approaches. In the reviewed papers, self-reports are the most commonly used labeling method, and the Self-Assessment Manikin (SAM) is the most common survey instrument for labeling physiological signals. SAM is a non-verbal, pictorial survey that measures users' subjective reactions to various stimuli. It is used in publicly available datasets, including WESAD, AMIGOS, DEAP, MAHNOB-HCI, and CASE, and in 5 of 30 papers that use private datasets. Other commonly used self-reporting surveys include emotion intensity ratings, the Subjective Unit of Distress (SUD), the State-Trait Anger Expression Inventory-2 (STAXI-2), perceived stress,

Table 2.3: Physiological Data Collection Devices

Data Collection Devices	Private Dataset	Public Dataset
Highly Obtrusive e.g., Biopac MP150, BioSemi ActiveTwo, Shimmer3	[87, 94, 116, 139, 142, 148, 149, 184, 192, 201, 235]	AMIGOS, DEAP, MAHNOB-HCI, DREAMER, CLAS, MERTI-Apps, ASCERTAIN, CASE
Less Obtrusive e.g., Polar H10, RespiBAN, Zephyr BioHarness	[11]	K-EmoCon, WESAD, SWELL, driveDB, CLAS, RECOLA, AffectiveROAD
Unobtrusive e.g., Empatica E4, Microsoft Band 2, Biovotion Armband,	[5, 18, 29, 48, 90, 98, 119, 122, 123, 143, 158, 159, 162, 165, 185, 220, 236]	WESAD, SWELL, PPG-DaLiA

and the Game Experience Questionnaire (GEQ).

Ecological Momentary Assessment (EMA) is also a self-reporting method that collects real-time data on an individual's experiences in their natural environment. Unlike traditional surveys such as SAM, EMA captures data as events occur in daily life, thereby minimizing recall bias and providing insights into temporal fluctuations. The eDiary app [137], Mobile App [53], and RankTrace [154] are examples of EMA techniques used in papers that use private datasets.

Task-based labeling involves assigning labels to physiological signals that correspond to specific tasks or activities during data collection. These tasks include watching videos (WESAD, CASE, AMIGOS, DREAMER, and ASCERTAIN), report writing (SWELL-KW), paired debates (K-EmoCon), and survival tasks requiring collaboration and decision-making (RECOLA). Task-based labels involved watching videos and listening to music, and were also commonly used in papers reporting private datasets. Several public and private datasets have employed hybrid data labeling approaches to enhance the validity of the assigned labels. For example, WESAD used self-reports and task-based labels, MAHNOB-HCI used self-reports and external annotators, and AMIGOS used self-reports, task-based labels, and external annotators to label the data. Table 2.4 shows all the physiological signal labeling approaches.

Table 2.4: Data Labeling Approaches

Data Labeling	Private Datasets	Public Datasets
Self-report e.g., SAM, EMA, SUD, STAXI-2, GEQ	[18, 48, 72, 87, 90, 94, 98, 119, 121, 122, 133, 139, 142, 149, 158, 159, 185, 192, 201, 220, 235, 236]	WESAD, AMIGOS, MAHNOB-HCI, DEAP, DREAMER, ASCERTAIN, MERTI-Apps, K-EmoCon, CASE, CLAS, driveDB, SWELL-KW
Task-Based	[87, 90, 116, 133, 147]	WESAD, AMIGOS, MAHNOB-HCI, CASE, CLAS, driveDB, SWELL-KW, AffectiveROAD, PPG-Dalia
External Raters	[90, 165]	MAHNOB-HCI, AMIGOS, RECOLA, K-EmoCon

2.2.2 Modeling Description

Developing accurate ML models for stress and affect recognition involves multiple stages, including selecting the most relevant physiological signals, data preprocessing, feature engineering, and model training and testing. DL can incorporate data preprocessing and feature engineering into model training, thereby simplifying the modeling process. To understand the best modeling practices, this subsection details the stages of the modeling process as reported in the reviewed papers.

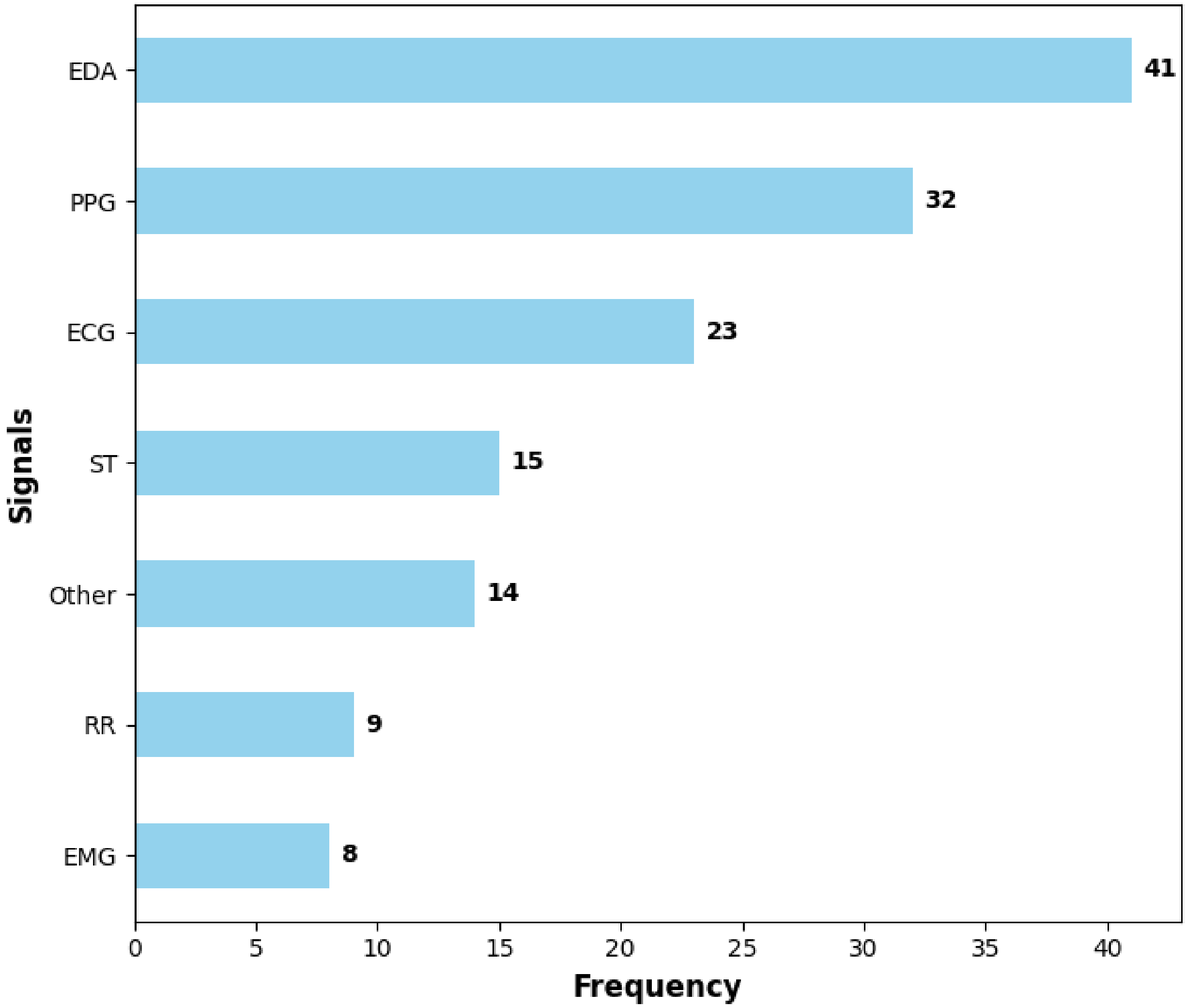


Figure 2.3: Physiological Signal Frequency Barchart

Physiological Signals

Across reviewed papers, various physiological signals were used to develop ML models for stress and affective recognition. The most commonly used peripheral physiological signals in papers using publicly available datasets are EDA (20 papers), PPG (14 papers), ECG (13 papers), ST (8 papers), and Respiratory Rate (RR) (6 papers). 3 of 26 papers that used public datasets also employed EMG to recognize affective states. Other non-physiological

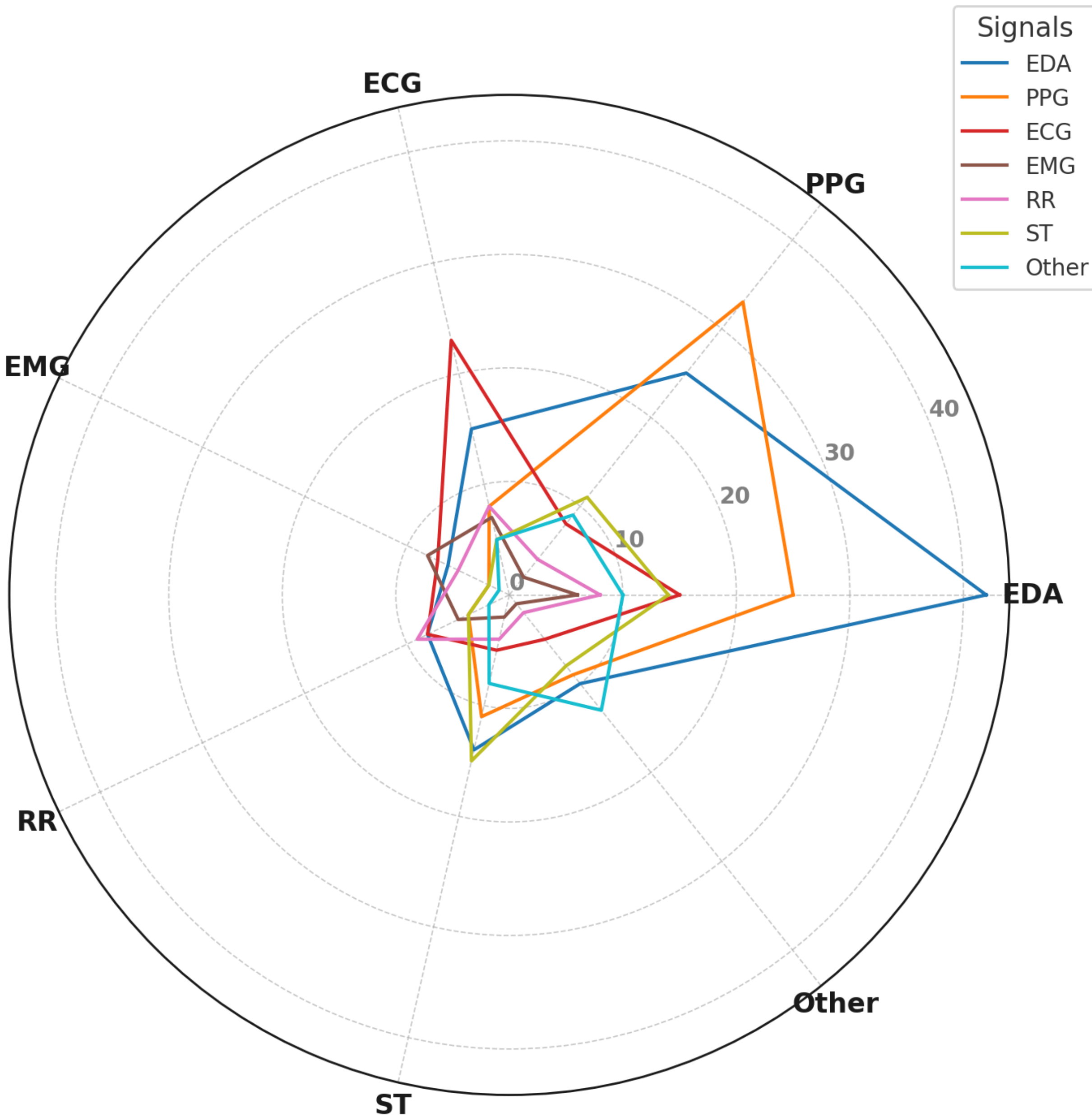


Figure 2.4: Radar Chart of Physiological Signal Combination

signals include the user's Accelerometer (ACC), which is not a measure of human physiology but is an indicator of human activity. The peripheral physiological signals used for stress and affect recognition in papers with private datasets include EDA (21 papers), PPG (18 papers), ECG (10 papers), ST (6 papers), EMG (5 papers), and RR (3 papers). Notably, these signals were used in various combinations to enhance stress and affect recognition accuracy.

The peripheral physiological signals most commonly used together overall are EDA and PPG (25 papers). This is consistent with the physiological responses of the sympathetic and parasympathetic nervous systems. EDA, driven by sweat gland activity, is a hallmark of sympathetic activation. PPG provides insights into heart rate and Heart Rate Variability (HRV), modulated by sympathetic and parasympathetic activation. Combining these signals leverages the complementary nature of sympathetic and parasympathetic physiological responses to enhance affect recognition accuracy. Figure 2.3 provides a quick overview of the frequency of the signals used in papers using both public and private datasets for stress and affect recognition. Figure 2.4 shows the distribution and combination of various signals used for stress and affect recognition. Each axis corresponds to one signal modality, and each colored polygon represents the relative frequency with which that signal appears as a primary feature in the surveyed literature.

Table 2.5: Signal Preprocessing Techniques

	Smoothing / Artifact Removal	**Signal Segmentation**	**Other Techniques**
EDA	Low-pass filters [11, 18, 21, 35, 66, 87, 94, 121, 139, 165, 192, 206, 236] Band-pass filters [35, 94, 118, 139, 164, 236] Artifact detection [11, 18, 29, 165, 206] Moving average filters [139, 206] Polynomial filters [139] Wavelet transform [118]	Fixed-length Window [11, 92] Overlapping Window [92, 164] Short-length [139] Event-based [147]	Decomposition [18, 21, 29, 35, 87, 90, 94, 118, 119, 133, 148] Normalization [66, 90, 92, 118, 123, 133, 147] Convolution [121, 147] Deconvolution [18, 118]
PPG	Low-pass filters [11, 66, 87] Band-pass filters [21, 90, 139] Artifact detection [29]	Fixed-length Window [11, 72, 92, 141] Short-length [139, 140, 142] Pulse-based [140, 142]	Normalization [66, 90, 92, 123, 140, 141, 142] Signal Splitting and Trend Removal [140] Peak-to-Peak Interval (PPI) [72]

Continued on next page

Table 2.5: Signal Preprocessing Techniques (Continued)

	Smoothing / Artifact Removal	**Signal Segmentation**	**Other Techniques**
ECG	Band-pass filters [21, 94, 121, 206, 215, 235] Low-pass filters [11, 66] Pan-Tompkins algorithm [26, 206] Data cleaning [259]	Fixed-length Window [11, 26, 72, 215, 235] Sliding Window [26, 35, 235, 259] Overlapping Window [26, 35, 235, 259] Frame-based [26, 259]	Normalization [26, 72, 102, 235, 259] QRS Complex Detection [72, 235] Interbeat Interval (IBI) [72, 102]
Skin Temp (ST)	Band-pass filters [21]	N/A	N/A
Resp Rate (RR)	Band-pass filters [21, 121] Low-pass filters [11, 121] Moving average filters [94]	Fixed-length Window [11]	N/A
EMG	Low-pass filters [11] Denoising with discrete wavelet transformation [201]	Fixed-length Window [201]	Minimal-redundancy-maximal-relevance (mRmR) criterion [201] Normalization [201]

Continued on next page

Table 2.5: Signal Preprocessing Techniques (Continued)

	Smoothing / Artifact Removal	**Signal Segmentation**	**Other Techniques**
Multi	Band-pass filters [94, 121, 184, 188] Low-pass filters [121, 149, 192] Downsampling [185, 206] Fast Fourier Transform [122]	Sliding Window [21, 53, 80, 98, 121, 135, 158, 159, 192, 220] Fixed-length Window [48, 98, 185, 188, 236] Event-based [121, 149, 158, 192]	Normalization [11, 21, 80, 135, 149, 158, 162, 184, 192, 260] Feature Imputation [162, 165, 220]

Signal Preprocessing Techniques

Signal preprocessing techniques help eliminate noise and artifacts from physiological signals that can interfere with stress and affect recognition. The signal preprocessing techniques used in the reviewed papers can be divided into three categories: smoothing and artifact removal, signal segmentation, and other miscellaneous techniques. Table 2.5 indicates the usage of each technique and the papers employing them.

Smoothing aims to reduce noise, whereas artifact-removal techniques focus on removing artifacts caused by movement and electrode disconnections in physiological signals. Low-pass filters (22 papers) and band-pass filters (21 papers) are among the most commonly applied filters to remove artifacts from peripheral physiological signals. The moving average filter (3 papers) was used to remove noise from EDA and RR.

Signal segmentation involves dividing the continuous signal into more manageable or meaningful segments for stress and affect recognition. Fixed-length window segmentation (18 papers) is widely applied across all physiological signals in the reviewed papers. It divides the signal into equal-length segments based on time- or sample-based fixed windows. Sliding or overlapping window segmentation (33 papers) was also commonly used, employing overlapping windows to capture more temporal context within each segment.

Other miscellaneous techniques for signal preprocessing include normalization and standardization (30 papers), signal decomposition (11 papers), peak detection (5 papers), and imputation (3 papers). Normalization and standardization ensure that signals from different users, recording sessions, or devices are comparable. Signal decomposition techniques are primarily applied to EDA signals to separate tonic and phasic components. Peak detection techniques (3 papers) specifically analyze the heart rate and HRV. These include peak-to-peak detection in PPG signals, QRS detection, and inter-beat interval (IBI) calculation. Lastly, mentions of imputation techniques (3 papers) fill in the missing values, a common phenomenon in physiological signals.

Table 2.6: Physiological Signal Features

	Time Domain	**Frequency Domain**	**Time-Frequency**	**Nonlinear and Other Features**
EDA	Mean [18, 29, 35, 87, 90, 94, 118, 119, 123, 148, 164, 206, 260] Standard Deviation [18, 29, 35, 87, 90, 118, 119, 123, 133, 148, 164, 206, 260] Min [18, 35, 87, 90, 94, 118, 123, 148, 164, 206] Max [18, 35, 87, 90, 94, 118, 123, 148, 164, 206] Skewness [18, 35, 118, 119, 123, 133, 206] Kurtosis [26, 35, 206, 215] SCR-related [29, 123, 164, 206, 236] Peak-related [11, 29, 87, 206, 236] Mobility [133]	Low Frequency (LF) [35, 123] High Frequency (HF) [35, 123] Very Low Frequency (VLF) [35, 123]	N/A	Fractal Dimension Entropy [133] Refined Composite Multiscale Dispersion Entropy (RCMDE) [133] Multiscale Entropy [35] Morphological Features [11, 18, 206, 236]

Continued on next page

Table 2.6: Physiological Signal Features (Continued)

	Time Domain	**Frequency Domain**	**Time-Frequency**	**Nonlinear and Other Features**
PPG	Mean [29, 87, 90, 116, 123, 141, 236, 260] Standard Deviation [29, 90, 116, 123, 141, 260] Root Mean Square of Successive Differences (RMSSD) [29, 90, 116, 123, 141] Kurtosis [35, 119, 123, 236] Min [87, 90, 119] Max [87, 90, 119] Interbeat Interval (IBI) [122, 236] Pulse Transit Time (PTT) [72]	Low Frequency (LF) [29, 116, 123, 141, 236] High Frequency (HF) [29, 116, 123, 141, 236] Low Frequency/High Frequency Ratio (LF/HF ratio) [29, 116, 123, 141, 236] High Frequency Power [29, 116, 123, 141, 236] Very Low Frequency (VLF) [29, 116, 123, 141]	N/A	Poincaré Plot Indices (SD1 and SD2) [123] Multiscale Entropy (MSE) [260] BVP Specific Features [11, 92] Local Maxima and Minima Features [123] HRVAS Toolbox Features [122] Deep Learning Features [66]

Continued on next page

Table 2.6: Physiological Signal Features (Continued)

	Time Domain	**Frequency Domain**	**Time-Frequency**	**Nonlinear and Other Features**
ECG	Root Mean Square of Successive Differences (RMSSD) [16, 29, 72, 121, 206, 215, 235] Mean [26, 35, 121, 206, 215, 235, 260] Standard Deviation [26, 35, 121, 206, 215, 235, 260] Heart Rate Variability (HRV) [35, 90, 206, 215, 235] pNN50 [16, 206, 215] Max [35, 90, 206, 235] Min [35, 206, 235] Pulse Transit Time [72, 122]	Low Frequency (LF) [16, 26, 29, 35, 72, 215, 235] High Frequency (HF) [16, 26, 29, 35, 72, 215, 235] LF/HF ratio [16, 26, 35, 215, 235] Very Low Frequency (VLF) [16, 35, 215, 235] Power Spectral Density (PSD) [35] Number LF (nLF)[35]	Wavelet Scattering [215]	Poincaré Plot Indices (SD1 and SD2) [16, 123, 215, 235] Shannon Entropy [35, 215] Sample Entropy [215] Local Maxima and Minima Features [123] HRVAS Toolbox Features [122] Deep Learning Features [66]

Continued on next page

Table 2.6: Physiological Signal Features (Continued)

	Time Domain	**Frequency Domain**	**Time-Frequency**	**Nonlinear and Other Features**
Skin Temp (ST)	Mean [260] Standard Deviation [260] Kurtosis [260] Skewness [260]	N/A	N/A	Temperature Curvature and Derivative Measures [122] Thermal Event Markers [235] Toolbox Features and Derivative Analysis [26]
Resp Rate (RR)	Breathing Rate [139]	N/A	N/A	N/A

Continued on next page

Table 2.6: Physiological Signal Features (Continued)

	Time Domain	**Frequency Domain**	**Time-Frequency**	**Nonlinear and Other Features**
EMG	Mean Absolute Value (MAC) [201] Zero Crossing (ZC) [201] Slope Sign Changes (SSC) [201] Root Mean Square (RMS) [201]	Power Spectral Density (PSD) [201]	Wavelet Coefficients [201]	N/A

Continued on next page

Table 2.6: Physiological Signal Features (Continued)

	Time Domain	**Frequency Domain**	**Time-Frequency**	**Nonlinear and Other Features**
Multi	Mean [121, 149, 158, 159, 162, 165, 184, 185, 188, 220, 237, 249, 260] Standard Deviation [11, 149, 159, 162, 165, 184, 185, 188, 220, 237, 249] Root Mean Square (RMS) [121, 184, 185, 188, 192, 220] Peak values/amplitude [184, 185, 188, 237] Variance [184, 185, 220, 237] Range [11, 149, 220]	Low Frequency (LF) [149, 158, 162, 185, 188, 192, 206] High Frequency (HF) [149, 158, 162, 185, 188, 192, 206] LF/HF ratio [149, 158, 185, 192, 206] PSD [185, 188, 206, 249] Very Low Frequency (VLF) [149, 162, 206] Ultra-Low Frequency (ULF) [162]	Wavelet Transform Coefficients [185] Wavelet Packet Entropy [188] Wavelet Coefficients [192]	Approximate Entropy [188] Sample Entropy [188] Fuzzy Entropy [188] Poincaré Plot Indices (SD1 and SD2) [185] Machine Learning Features [185, 249]

Feature Extraction

Raw physiological signals often lack meaningful or interpretable information, making feature extraction a critical step in stress and affect recognition. The features extracted from the physiological signals in the reviewed papers can be categorized as time-domain, frequency-domain, time-frequency, and non-linear. Table 2.6 shows the usage of these features in the reviewed papers.

The focus on time-domain features was often on statistical measures across physiological signals. These include mean (42 papers), standard deviation (38 papers), minimum (16 papers), maximum (16 papers), skewness (8 papers), and kurtosis (5 papers). Mean provides the average signal level, showing the baseline physiological response, while the Standard deviation indicates the extent of fluctuations around this baseline. Minimum and maximum values capture the range of the response, identifying the lowest and peak levels. Skewness highlights the asymmetry in response distribution, showing whether responses are biased toward higher or lower values. Kurtosis reveals the presence of extreme values, indicating intense or unusual responses. Other commonly used time-domain features include Root Mean Square (RMS) (7 papers) and Root Mean Square of Successive Differences (RMSSD) (12 papers). RMS measures the overall magnitude of a signal, while RMSSD captures rapid changes by focusing on the differences between consecutive signal values.

Frequency-domain features involve dividing the signal into frequency bands: Very Low Frequency (VLF) (13 papers), Low Frequency (LF) (21 papers), and High Frequency (HF) (21 papers). These bands represent distinct physiological processes linked to various affective states, with VLF often associated with long-term regulatory processes, LF reflecting a mix of sympathetic and parasympathetic activity, and HF primarily related to parasympathetic activity. Also commonly used were the ratios of the LF and HF bands (15 papers). A higher LF/HF ratio indicates increased sympathetic dominance (often associated with stress or arousal), while a lower LF/HF ratio indicates stronger parasympathetic influence (linked to relaxation or calmness). Another commonly used frequency-domain feature was Power Spectral Density (PSD), which quantifies the intensity of various frequency bands over time.

Time-frequency features (5 papers) were also used, simultaneously providing information about physiological signals' time and frequency components.

Nonlinear features typically include entropy features (10 papers), including Shannon Entropy, Approximate Entropy, Sample Entropy, Multiscale Entropy, and Fuzzy Entropy. Entropy features in physiological signals measure their complexity or regularity, providing insight into the underlying physiological processes. Shannon Entropy quantifies the overall uncertainty or information content in a signal. Approximate Entropy evaluates the likelihood that similar patterns of observations remain similar in the future. Sample Entropy improves on Approximate Entropy by reducing bias and dependence on dataset length. Multiscale Entropy extends Sample Entropy across multiple temporal scales to capture both short- and long-term patterns in physiological dynamics. Fuzzy Entropy introduces fuzzy membership functions to enhance robustness against noise and small datasets. Morphological features (4 papers) were also used to capture complex signal characteristics, such as sudden changes, irregular patterns, or specific structural traits, in EDA signals. However, researchers are increasingly using ML and DL features that are more generalizable and have been shown to outperform hand-crafted features, highlighting the potential of data-driven approaches for feature extraction [98].

Feature Selection

Selecting the most relevant features is valuable for increasing stress and affect recognition accuracy and reducing computational requirements. Statistical methods (9 papers) are the most commonly used feature selection methods. These methods employ various statistical tests to assess the relationship between features and affective states, thereby eliminating irrelevant or redundant features. Correlation-based techniques such as Pearson Correlation used in [94, 141, 192] and Spearman Correlation used in [133] identify features strongly associated with affective states. Hypothesis testing methods like Chi-Square used in [92], Analysis of Variance (ANOVA) used in [164], and Scalable Hypothesis Testing used in [98] involve keeping features with significant relationships with the affective states based on the p-value. Other techniques, such as Mutual Information used in [121], also focus on retaining features that reduce uncertainty about the affective states, and SelectKBest used in [162] ranks features using scoring functions and selects the top K features.

Dimensionality reduction techniques simplify the extracted features by transforming features to a lower dimension while preserving critical information. Principal Component Analysis (PCA) used in [66, 123, 149, 215], transforms features into orthogonal components that maximize variance, reducing dimensionality without significant loss of information. Similarly, Minimum Redundancy Maximum Relevance (MRMR), used in [92, 201], selects features that are highly relevant to the target variable while minimizing redundancy among selected features, ensuring both efficiency and diversity in the feature set.

Specialized algorithms further enhance feature selection for complex or noisy datasets. The RELIEFF algorithm, used in [26, 72], estimates feature importance based on how well features distinguish nearby instances. While statistical methods dominate traditional feature selection approaches, DL techniques can automatically extract and select the most relevant features during model training.

Modeling Methods

Overall, 55 papers used ML and/or DL models for stress and affect recognition. Among these, 26 used ML algorithms, 19 employed DL architectures, and 10 used a combination of ML and DL. Figure 2.5 shows the distribution of the commonly used ML and DL algorithms from 2019 to mid-2024.

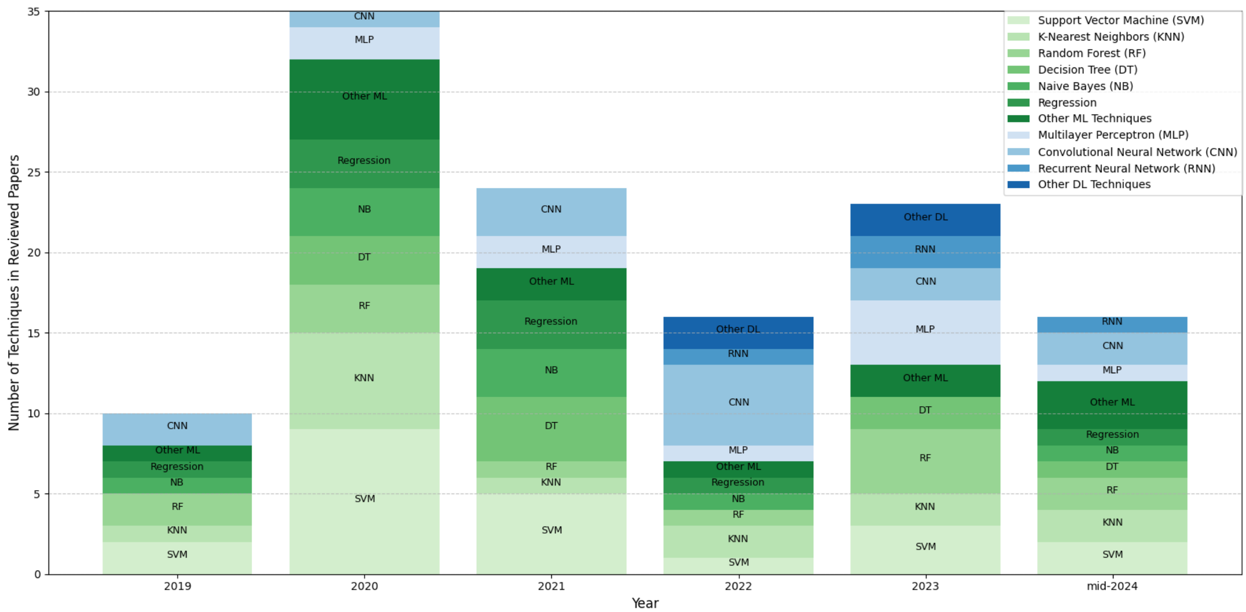


Figure 2.5: Machine Learning and Deep Learning Models Developed for Stress and Affect Recognition using Physiological Signals in the Reviewed Papers

Machine Learning: Support Vector Machines (SVMs), used in 22 papers, are the most commonly used ML algorithm for stress and affect recognition using physiological signals. SVMs are supervised ML models that identify the optimal hyperplane to separate data into different classes. They effectively handle high-dimensional data and are robust to over-fitting, making them suitable for stress and affect recognition using physiological signals. 14 papers used K-Nearest Neighbor (KNN), an instance-based learning algorithm that classifies new data points based on the majority class among the 'k' closest training examples and captures subtle variations in physiological signals. Random Forest (RF) (13 papers) and Decision Trees (DT) (10 papers) capture non-linear relationships and provide insights into feature importance, explaining which physiological signals/features are most indicative of stress or other affective states. 9 papers used Regression models, offering a nuanced understanding beyond the categorical classification of affective states from physiological signals. Naive Bayes classifiers, used in 8 papers, are computationally efficient and beneficial when physiological data is limited.

Other ML techniques include Extreme Gradient Boosting (XGBoost), Extreme Learning Machines (ELM), Self-Organizing Maps (SOM), Zero R, Locally Weighted Learning (LWL), explainable ML, and unsupervised ML algorithms. Extreme Gradient Boosting (XGBoost) used in [11, 35, 165] captures intricate patterns through boosting techniques. ELM used in [188] provides a fast training approach with randomly initialized input weights. SOM, as used in [26], creates low-dimensional representations of high-dimensional data, aiding in visualizing emotional states. Methods such as ZeroR [102] and LWL [48] serve as baseline classifiers or localized models that adapt to nearby data points. Explainable ML techniques, including SHAP (SHapley Additive exPlanations), were also used in [16] to improve interpretability by identifying feature contributions. Unsupervised ML algorithms such as Expectation-Maximization (EM) cite de2019study and k-means clustering [48, 122] can identify inherent patterns in physiological signals and facilitate stress and affect recognition by informing supervised ML models.

Deep Learning: DL models for stress and affect recognition using physiological signals involved using Convolutional Neural Networks (CNNs) (14 papers), Long Short-Term Memory Networks (LSTMs) (4 papers), and Multilayer Perceptrons (MLPs) (10 papers). 1D CNNs (13 papers) apply convolutional filters along a single dimension, making them well-suited for processing time-series data such as physiological signals. They process time-series data effectively by learning spatial and hierarchical features in raw physiological signals. 2D CNNs, typically used for image data, are applied to spectrograms of physiological signals in [66]. They capture spatio-temporal features in these two-dimensional representations of physiological signals, which is beneficial in analyzing frequency-domain information.

LSTMs are Recurrent Neural Networks (RNNs) that can learn long-term dependencies in sequential data. They were combined with CNNs in CNN-LSTM architectures to capture spatial features and temporal dependencies over time in physiological signals for stress and affect [29, 102, 218, 260]. MLPs are fully connected neural networks that can approximate complex and nonlinear relationships between physiological features and affective labels. They were applied to hand-crafted features extracted using scientific methods and DL features extracted from CNN, LSTM, and CNN-LSTM architectures for affect recognition.

Advanced DL models used for stress and affect recognition in reviewed papers include Transformers used in [249], Autoencoders used in [142, 206], Variational Autoencoders (VAEs) used in [259], and contrastive learning used in [53]. Transformers use self-attention mechanisms to capture long-range dependencies. Autoencoders and VAEs are generative models that learn latent representations of data. These models help reduce dimensionality and capture complex patterns in physiological signals. Contrastive learning involves a smaller "student model" that learns representations by comparing its outputs with those of a larger, more complex "teacher model." There is a growing research interest in using these advanced DL models for time-series modeling for stress and affect recognition.

Cross-Validation

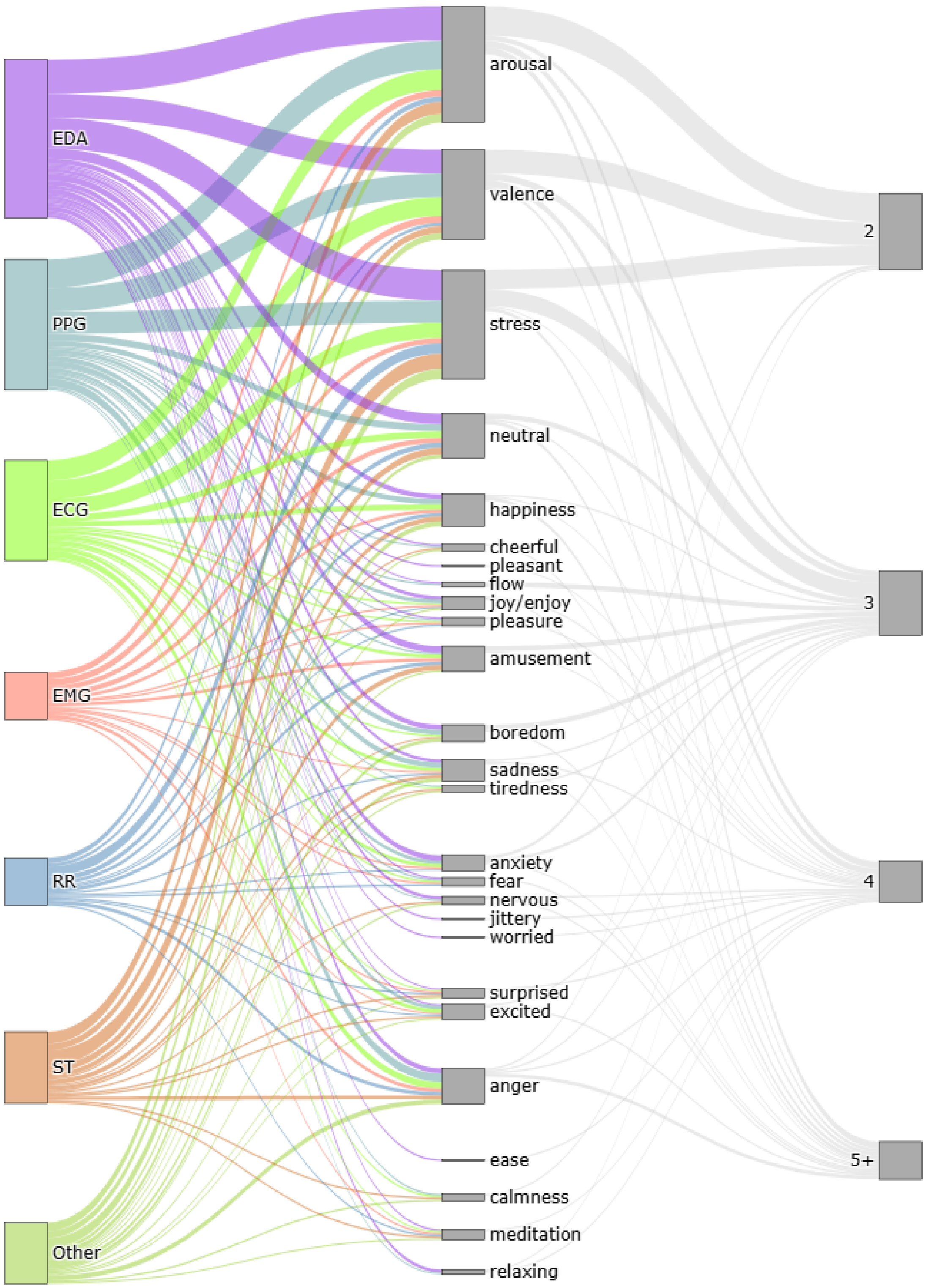
Cross-validation is a resampling technique to assess how well a model generalizes to new data. It provides a more reliable performance estimate than a single train-test split by rotating which data subsets are used for training and validation, reducing overfitting and helping select the best model or hyperparameters. The most frequently used cross-validation techniques are 10-fold cross-validation (17 papers) and leave-one-out cross-validation (LOOCV) (11 papers). 10-fold cross-validation involves dividing the dataset into 10 equal subsets, using nine subsets for training and one subset for testing. In LOOCV, the model is trained on all samples except one, which is held out for testing, and this process is repeated for every sample in the dataset. The reviewed papers also used Leave-One-Subject-Out (LOSO), Leavek-Subjects-Out (LkSO), Leave-One-Day-Out, and Leave-One-Event-Out cross-validation, which are modifications of LOOCV to develop more robust and generalizable stress and affect recognition models. Some papers also opted for straightforward data splitting as an alternative to, or in tandem with, cross-validation (5 papers).

Classes and Labels

In the developed stress and affect recognition models, the states were most commonly recognized as valence and arousal labels. These labels commonly included binary classes (15 papers) such as high and low valence or high and low arousal, and four classes (5 papers) that combined valence and arousal into four quadrants, including High Valence-High Arousal (HVHA), High Valence-Low Arousal (HVLA), Low Valence-High Arousal (LVHA), and Low Valence-Low Arousal (LVLA). Additionally, some models utilized more granular scales, such as 5-point or 10-point scales, to capture the intensity of valence and arousal.

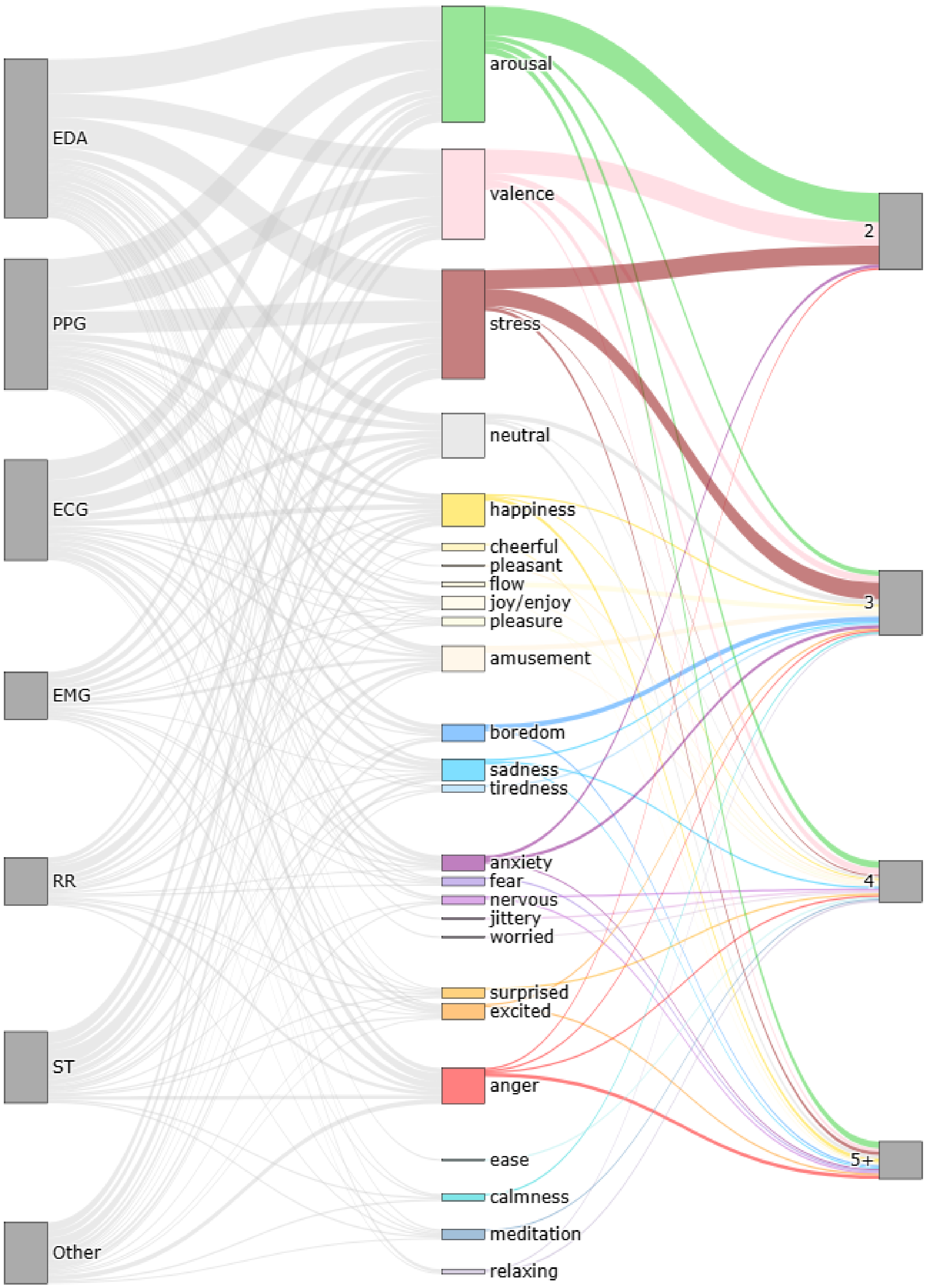
23 papers used categorical labels to infer stress, with the majority using binary (10 papers) and ternary (9 papers) classes. In binary systems, stress is labeled as either "stress" or "no stress" (or "relaxed"). In ternary systems, stress is labeled as "Low, Medium, High" or "Baseline, Stress, Amusement" or "Neutral, Stressful, Relaxing" or "Boredom, Flow, Stress." Stress was also classified into four distinct stress states (neutral, stress, amusement, meditation) [80] and as an ordinal progression (0.25, 0.5, 0.75, 1.0) [112]. Moreover, finer distinctions are made using scales (e.g., 1-10) [236] or multi-class systems that integrate stress with other states like focus, amusement, or phobia-related distress [218].

Other affective states are recognized as subjective labels with different classes, including joy, anger, sadness, and pleasure (5 papers). Fairclough et al. [72] developed a model for binary classification of anger as high and low levels. Additionally, some models used scaled



(a) Physiological signals → affective states.

Figure 2.6: Sankey diagrams of stress and affect recognition: (a) input physiological signals to affective states; (b) affective states to classification schemes.



(b) Affective states → classification schemes.

Figure 2.6: Sankey diagrams (continued).

categories, assigning emotions to high, medium, and low intensity [118]. Less commonly used affective states recognized using physiological data include jittery and flow. Figure 2.6 is a Sankey Diagram that shows the proportional flow of classification categories (right) to affective states (center) and to physiological signals (left), highlighting trends in how different signals are used to recognize various affective states with varying complexity.

2.3 System Adaptation

As discussed, system adaptation involves the dynamic adjustment of system behavior or user interface based on the affective states inferred from physiological signals. It is the second key component of systems with biocybernetic adaptation after psycho-physiological inference. Among the 61 shortlisted papers, 15 focused on the system's ability to adapt based on the affective states recognized from physiological signals. Among these, 6 papers focused specifically on designing system adaptations, while 9 focused on both affect recognition and system adaptations (Figure 2.2).

Papers focusing solely on system adaptation investigate the design of adaptations and evaluate their impact on users. Papers focusing on inference and adaptation describe the processes of stress and affect recognition using ML and/or DL, the design of system adaptation, and the evaluation of the impact of adaptation on users. This section details the system design of the reviewed systems with biocybernetic adaptation based on the MAPA framework proposed by Munoz et al. [172]. This framework provides a structured approach for categorizing systems with biocybernetic adaptations by emphasizing four key components: medium, application area, psycho-physiological state, and adaptation technique. Additionally, this review discusses the technical and user evaluations of such systems, providing insights into the accuracy of real-time stress and affect recognition and the effectiveness of system adaptation for users.

2.3.1 Medium

Medium refers to the means of communication through which the system adaptation occurs. This medium can include physical objects, such as robots, virtual interfaces, such as Augmented Reality (AR) and Virtual Reality (VR), or hybrid physical-virtual components. Each medium has a unique set of advantages and disadvantages for biocybernetic adaptations, as given below.

Physical Objects

Physical objects enable direct interaction with the environment, allowing for tangible feedback. They can seamlessly integrate into users' daily activities and provide greater personalization, thereby enhancing user engagement and the effectiveness of system adaptation. However, physical objects have limited customizability and adaptability. In this review, Leonidis et al. [143] used physical objects in a smart home to design a pervasive system with biocybernetic adaptations. Chen et al. [34] designed biocybernetic adaptations for an umbrella, whereby the umbrella changes color in response to the user's real-time physiological signals.

Virtual Interfaces

Virtual interfaces enable more dynamic and customizable adaptations than physical objects. AR and VR, used as a medium in 6 papers, offer high levels of immersion and control. Desktop applications (4 papers) and mobile applications (2 papers) were also commonly used as virtual media, enabling users to experience adaptive responses across various settings. Lee et al. [139] used CARLA, an open-source simulator for autonomous driving.

2.3.2 Application

Application refers to the context or domain where the BCA systems are used. The reviewed papers had three main application areas of systems with biocybernetic adaptation: training (6 papers), mental health (5 papers), and entertainment (4 papers).

Training

As discussed, users' cognitive and affective states influence their task performance. By continuously adapting to the user's affective state, systems with biocybernetic adaptations can create engaging and personalized training experiences to enhance performance. Téllez et al. [229] used an immersive VR football game to train users to improve their performance by adjusting elements of the VR environment based on stress levels measured from physiological signals. Reidy et al. [201] also used an immersive VR game to train users to improve their performance on working memory and episodic memory tasks, guided by affective state recognition from physiological signals and prior performance.

Adiani et al. [5] designed a Career Interview Readiness in Virtual Reality (CIRVR) platform to train autistic users for job interviews. The CIRVR platform adapts by interrupting when high stress is detected in users' physiological signals, providing a brief pause intended to alleviate stress before resuming the interview. Fairclough et al. [72] designed a biofeedback-like system that displays to users a geolocated visualization of vehicle parameters, photographs, and psycho-physiological inferences about their emotions, enabling them to regulate their emotions and improve driving performance. As discussed, biocybernetic adaptations derive from biofeedback but differ in that they use physiological signals to automatically drive system adaptations, without requiring conscious user input. Lee et al. [139] designed a system for autonomous cars that recognizes the driver's affective states with their physiological signals to switch between manual and auto driving to improve driving performance.

Mental Health

Systems with biocybernetic adaptations were commonly used to increase the effectiveness of mental health interventions and practices, such as exposure therapy, meditation, and emotion regulation. Systems that adapt in response to fear detected from the user's physiological signals are used in exposure therapy. For instance, Arquissandas et al. [14] developed an adaptive AR game that exposed users to a virtual snake, and Petrescu et al. [192] proposed an adaptive VR acrophobia game that exposed users to heights.

The CaLmi system [143] is a pervasive system with biocybernetic adaptation that helps occupants regulate their everyday stress by exposing them to soothing lights, sounds, and fragrances, thereby supporting meditation. The TROI system by Dissanayake et al. [53] employed a smartphone application to display users' everyday affective states, and Chen et al. [34] designed an umbrella with biocybernetic adaptation that changes color based on users' psycho-physiological states. The TROI system and the biocybernetic umbrella are more biofeedback-based systems because they establish conscious interaction with users, enabling them to regulate their emotions.

Entertainment

Systems with biocybernetic adaptation were also employed to design more engaging and entertaining gaming experiences. Orozco-Mora et al. [184], and Frachi et al. [81] designed immersive VR games that adjust difficulty based on users' stress levels, measured from physiological signals, to facilitate user engagement. Karavidas et al. [118] designed an adaptive kitchen game that adjusts its difficulty based on the user's emotions recognized from their physiological signals, ensuring it is challenging enough to match the user's skills and keep them engaged. Katada et al. [119] proposed a physiologically adaptive dialogue system that adjusts its conversational responses to make the interaction more engaging and emotionally attuned. The CIRVR platform [5] also proposed developing a virtual interviewer that adapts its conversational style based on user stress detected from physiological signals. Chiossi et al. [37] designed an immersive social VR game that adjusts the number of Non-Player Characters (NPCs) in the virtual environment based on inferences from users' physiological signals to maintain proxemic comfort.

2.3.3 Psycho-Physiological States

Several physiological signals were used to infer psychological states, triggering system adaptation. The commonly used psychological states include stress (4 papers), anxiety (2 papers), valence (2 papers), and arousal (2 papers), which are recognized from a combination of physiological signals, including EDA (11 papers), PPG (10 papers), ECG (4 papers), and EMG (2 papers).

Stress and Anxiety

Stress and anxiety, as indicated by EDA, PPG, and ECG, were the most commonly used psycho-physiological inferences in the reviewed systems with biocybernetic adaptation. The CaLmi smart home system [143] monitors EDA and PPG signals from the Empatica E4 wristband to detect binary stress. It also incorporates contextual data to enhance the detection accuracy, including the occupant's profile, activities, habits, and responsibilities. Similarly, the CIRVR platform [5] uses EDA and PPG signals from the Empatica E4 wristband for binary stress detection.

The immersive VR football game in [229] leverages PPG signals collected from the Polar OH1 sensor to classify heart rate into a 5-level stress scale. The immersive zombie-survival game in [184] uses EDA, ECG, and EMG signals to classify stress into three levels and dynamically adjust the game difficulty.

The AR exposure therapy system [14] collects EDA and ECG signals from the Bitalino device to infer user anxiety levels. Similarly, the adaptive kitchen game [118] uses EDA signals collected with the Bitalino sensor to recognize real-time boredom, flow, and anxiety levels across low, medium, and high categories.

Valence and Arousal

Valence and arousal inferred from EDA and PPG signals were also commonly used to trigger biocybernetic adaptation. In the market and museum immersive VR game [201], facial EMG signals are captured using the Faceteq sensing HMD foam to classify user affect into four valence-arousal combinations: energetic-positive (high valence, high arousal), calm-positive (high valence, low arousal), energetic-negative (low valence, high arousal), and calm-negative (low valence, low arousal).

In the biocybernetic adaptations for autonomous cars [139], PPG and EDA signals from Biopack's MP150 system are used to infer valence and arousal levels. These levels correspond to driving scenarios, including reduced speed perception, delayed incident response, diminished situational judgment, and maintained driving capability, and are translated into control signals for the autonomous vehicle. The TROI system [53] collects EDA, PPG, and ST signals from the Empatica E4 wristband to separately recognize arousal and valence into three classes (high, neutral, and low). These signals also classify six categorical emotions: happy, sad, angry, nervous, cheerful, and neutral.

Other States

Fairclough et al. [72] developed a biofeedback-like system that combines ECG and PPG signals from the Shimmer3 sensor with driving features, such as car speed, to classify high and low anger episodes. In the physiologically adaptive dialogue system [119], EDA and PPG signals collected from the Empatica E4 wristband are used to classify high- and low-enjoyment classes.

No State

Some studies did not infer specific psychological states but instead used rule-based deductions from physiological signals to trigger adaptation. For example, the Labyrinth Game [81], a VR game for proxemic comfort [37], and the umbrella with biocybernetic adaptation [34] used changes in EDA signals to trigger system adaptation.

2.3.4 Adaptation Technique

Adaptation techniques are computational methods used to achieve system adaptation. Among 15 papers focusing on systems with biocybernetic adaptation, 11 used ML and DL techniques, whereas 4 used rule-based approaches to infer stressful and affective states from physiological signals to trigger system adaptation. ML algorithms include SVMs used by [119, 184, 201], kNN used by [72, 184, 201], RF used by [5, 184], LDA used by [72, 201], and Logistic Regression used by [143] for stress and affect recognition. A few authors developed multiple models and used one as the final classifier for inference and adaptation. DL learning models primarily employ 1D CNNs to infer affective states from physiological data and have been used by [53, 139, 192].

Rule-based approaches rely on predefined rules, thresholds, or heuristics derived from expert knowledge or empirical data to guide system behavior. The rule-based approach in [37, 81, 229] involved measuring baseline physiological signals, monitoring current signals, and calculating the difference between baseline and current signals as indicators of change in psychological states to trigger system adaptation. The rule-based adaptation in affective umbrella [34] uses predefined thresholds, and the system adapts when signals exceed or fall below them.

Table 2.7: Summary of Papers with System Adaptation based on the Medium, Application, Psycho-physiological states, and Adaptation technique (MAPA) Framework by Munoz et al. [172]

Sr	**Paper**	**Medium**	**Application**	**Psycho-physiological States**	**Adaptation Technique**
1	Fairclough et al. [72]	Mobile application	Driving performance via emotional awareness	Anger using ECG, PPG, and other physiological signals	ML
2	Petrescu et al. [192]	Immersive VR	Exposure therapy (exposure to heights)	Fear using EDA and PPG	ML
3	Lee et al. [139]	Driving simulator	Driving performance in autonomous vehicles	Valence and arousal using EDA and PPG	DL
4	Leonidis et al. [143]	Smart home	Everyday stress management	Stress using EDA and PPG	ML

SN	Paper	Medium	Application	Psycho-physiological States	Adaptation Technique
5	Chiossi et al. [37]	Immersive VR	Proxemic comfort and engagement in VR	No state; EDA and ECG	Rule-based
6	Karavidas et al. [118]	Desktop application	Engagement and performance in time management game	Boredom, amusement, and anxiety using EDA	ML
7	Téllez et al. [229]	Immersive VR	Sports training and performance	Boredom, amusement, and anxiety using EDA	Rule-based
8	Arquissandas et al. [14]	AR	Exposure therapy (exposure to snakes)	Fear using EDA and ECG	Unknown
9	Orozco-Mora et al. [184]	Immersive VR	Engagement in a First Person Shooting (FPS) zombie game	Stress using ECG and EMG	ML
10	Adiani et al. [5]	Desktop VR	Training for job interviews	Stress using ECG and EDA	ML
11	Reidy et al. [201]	Immersive VR	Engagement and training for memory improvement	Valence and arousal using EMG	ML
12	Chen et al. [34]	Umbrella	Emotion regulation	No state; EDA and PPG	Rule-based

SN	Paper	Medium	Application	Psycho-physiological States	Adaptation Technique
13	Katada et al. [119]	Desktop application	Engagement with conversational agent	Enjoyment using PPG and EDA	DL
14	Frachi et al. [81]	Desktop application	Engagement and performance in labyrinth game	Arousal using EDA, PPG, and ST	Rule-based
15	Dissanayake et al. [53]	Mobile application	Everyday emotion regulation	Valence, Arousal, and discrete emotions using EDA, PPG, and ST	DL

2.3.5 Evaluation

The evaluation of systems with biocybernetic adaptation must assess the accuracy of stress and affect recognition and the effectiveness of system adaptation for the user. However, the reviewed papers focused on evaluating the impact of system adaptation on users, using its expected impact as a proxy for assessing stress and affect recognition accuracy. This approach, while practical, introduces limitations because the accuracy of inference models is assessed indirectly. User-centered evaluation systems with biocybernetic adaptation can be categorized as adaptive or non-adaptive, standalone, or mixed-methods.

Adaptive vs. Non-Adaptive

This evaluation approach involved conducting user studies to compare the adaptive and non-adaptive systems. For example, participants in the market and museum immersive VR game [201] alternated between adaptive and non-adaptive versions of the game, reflecting on their experiences using the Game Experience Questionnaire (GEQ) and interviews. This study targeted older adults as users, indicating enhanced engagement and motivation in the adaptive version, as participants reported feeling more competent and less challenged. Similarly, the labyrinth game [81] involved three levels of game adaptation and post-task evaluations with regular users. Results revealed that increased character movement speed reduced level completion time, whereas slower speed conditions resulted in higher success rates, and adaptive scenarios improved user control and feedback. The immersive zombie-survival game [184] evaluated workload using the NASA-TLX questionnaire among regular users, showing no statistically significant differences between adaptive and non-adaptive scenarios but suggesting a slight increase in perceived demands in the non-adaptive scenario.

Stand-Alone Evaluations

Other approaches involved stand-alone evaluations of systems with biocybernetic adaptation based on their primary functions among various users. For instance, Calmi's evaluation compared adaptive programs and user feedback obtained through interviews and questionnaires [143]. CIRVR [5], designed for individuals with autism, used the System Usability Scale (SUS) to compare usability perceptions between autistic and neurotypical participants. Neurotypical participants rated the platform significantly higher (SUS score: 77.5) compared to autistic participants (57.5), highlighting accessibility challenges for neurodiverse users.

Mixed-Methods Approach

Some papers integrated quantitative and qualitative methods for a more comprehensive analysis. For example, the autonomous driving study [139] measured vehicle response times across situational awareness scenarios with regular users while collecting detailed data on user performance and physiological responses. Results showed that control response times remained well within commercial safety standards, even with the added processing time required for emotion recognition. The affective umbrella [34] study targeted regular users and utilized the SAM scale to capture emotional states before and after each session, supplemented by body map drawings and interviews. Results showed significant differences in arousal and dominance scores across conditions.

Evaluating Inferences

Among all reviewed papers, only the TROI system [53] combined affect recognition ratings with self-reported confidence levels, showcasing the system's accuracy in interpreting users' affective states. This study, involving regular users, demonstrated improved emotional awareness and underscored the importance of integrating intervention methods to enhance broader applicability. Participants expressed that such systems could help them "react better in different situations," emphasizing real-world usability and psychological impact. However, the TROI system's design and functionalities align more closely with biofeedback-type systems.

2.4 Discussion

So far in this chapter, various approaches to developing stress and affect recognition models and to designing and evaluating systems with physiology-driven adaptations have been reviewed. Based on this review, guidelines for designing systems with biocybernetic adaptations were derived. These guidelines are discussed in this section alongside the critical components of stress and affect recognition using physiological signals and system adaptation. Lastly, this section discussed the application of systems with biocybernetic adaptation in emerging technologies.

2.4.1 Guidelines for Designing Systems with Biocybernetic Adaptation

As discussed, systems with biocybernetic adaptation consist of two main components: psychophysiological inference and system adaptation. Therefore, designing such systems requires not only accurate and robust models for inferring states such as stress and affect from physiological signals but also effective adaptation mechanisms that translate these inferences into meaningful changes for the user, including enhancing engagement and reducing stress.

Guidelines for Psychophysiological Inference of Stress and Affect

As discussed, the typical pipeline for developing stress and affect recognition models includes data collection, data processing, model development, and model validation. Data collection involves exposing the participant to a stimulus, recording physiological responses using various devices, and labeling the recorded data. To develop accurate inference models for systems with biocybernetic adaptation, the stimuli to which the participant is exposed in the training dataset should be similar to the conditions the participant will encounter when interacting with the system. For example, the stress recognition model trained on a dataset using arithmetic tests as stressors would work better in a system for arithmetic training. Labeling physiological signals is crucial; however, the reviewed papers mostly relied on self-reports, whereas hybrid labeling can enhance the validity and reliability of the assigned labels. Based on these insights, we propose the first guideline (G1): psychophysiological inference models should be trained with application-relevant stimuli, employ reliable labeling strategies, and ensure that labels are ecologically valid for the intended use case.

Traditionally, data processing involves eliminating noise and artifacts and extracting relevant features from physiological signals. These physiological features and the assigned labels are then used with supervised ML algorithms to infer psychological states, including stress and affect. DL can implement this workflow within a single neural network, thereby developing end-to-end models for stress and affect recognition [113]. DL approaches such as CNNs, LSTMs, and the attention mechanism can extract relevant features from multiple raw signals [257] and allow them to be fused [62, 252, 258]. More importantly, DL enables multi-tasking, transfer learning, and meta-learning, allowing the model to recognize multiple related psychological states from the same data. For example, a multi-task DL model can recognize stress from objective measures (task-based labels) and subjective experiences (self-reports), a transfer-learning DL model can learn to recognize emotions in a new domain (e.g., workplace stress) by leveraging knowledge gained from a related domain (e.g., academic stress), and a meta-learning DL model can quickly adapt to new users by learning how to learn from a few examples, enabling personalized affect recognition systems. Based on these insights, we propose the second guideline (G2): psychophysiological inference models should incorporate advanced DL approaches, including multi-task, transfer, and meta-learning, to ensure robustness, generalizability, and personalization across contexts and users.

Guidelines for System Adaptation

The choice of medium, the means of communication through which the system adaptation occurs, plays a crucial role in the effectiveness of systems with biocybernetic adaptation. Physical objects integrate naturally into users' daily lives, fostering engagement through direct interaction and personalized feedback, but have limited customizability. Despite this, they offer significant advantages in task-oriented scenarios in which tactile feedback enhances the experience [32]. On the other hand, virtual interfaces offer unparalleled adaptability. These media are particularly well-suited to creating dynamic and customizable user experiences across mobile, desktop, and fully immersive VR platforms. Hybrid components that merge physical and virtual elements can be used for applications that require a balance of realism and interaction. The choice of medium should balance the complexity of integrating psychophysiological inferences with ease of use to ensure user engagement and effectiveness. Based on these insights, we propose the third guideline (G3): the choice of medium should align with the application domain, balancing ecological validity, adaptability, and ease of integrating psychophysiological inferences to maximize user engagement and system effectiveness.

In the reviewed papers, the commonly used physiological signals include EDA, PPG, ECG, and EMG, and the commonly inferred psychological states include stress, anxiety, valence, and arousal. The selection of the psychological state depends mainly on the application, and the selection of physiological signals primarily depends on the psychophysiological association. For example, EDA and PPG are frequently used for inferring stress and arousal due to their strong associations with autonomic responses. At the same time, ECG and EMG are preferred in scenarios requiring precise cardiac or facial activity monitoring linked to specific psychological states. Other factors influencing signal selection include ease of acquisition and obtrusiveness. For example, lab-based sensors are suitable for controlled environments, whereas wearable sensors are more suitable for pervasive applications. Based on these insights, we propose the fourth guideline (G4): select psychological states according to the application goals, choose physiological signals with strong psychophysiological associations to those states, and ensure the chosen sensors are appropriate in terms of accuracy, obtrusiveness, and ecological validity for the target medium.

ML and DL techniques were used to develop inference models for stress and affect recognition, and the outputs of these models provide objective triggers for system adaptation. Some systems used physiological triggers, such as deviations from baseline signals, rather than psychophysiological inference models to trigger system adaptations. This approach simplifies implementation and is particularly useful for applications requiring low computational overhead. However, systems that rely solely on physiological triggers lack precision, as physiological changes may not directly correspond to specific psychological states. Moreover, system adaptations may be generic or suboptimal without clear state inference, leading to higher rates of false positives and negatives. Additionally, such systems are less generalizable and scalable across users and applications. Based on these insights, we propose the fifth guideline (G5): system adaptations should be driven by validated psychophysiological inferences rather than raw physiological triggers to ensure precision, reduce false positives and negatives, and enhance generalizability across users and contexts.

The effectiveness of systems with biocybernetic adaptations can be predominantly characterized by their ability to fulfill their intended purpose. For example, the effectiveness of a system designed to train users in arithmetic tests must improve users' test performance. However, systems should be evaluated for both affect recognition accuracy and system adaptation effectiveness to ensure that the outcome is... A multi-phase evaluation strategy, involving separate test phases for psychophysiological inference and system adaptation, followed by holistic testing, can be effective. The inference recognition accuracy should be directly assessed using objective measures, such as comparison with ground-truth data (e.g., self-reports). The effectiveness of system adaptation should be evaluated through its impact on user experience, engagement, task performance, and cognitive and emotional outcomes. Holistic testing can cover both these components to reveal the overall effectiveness of the system. Based on these insights, we propose the sixth guideline (G6): system evaluations should adopt a multi-phase strategy that separately validates inference accuracy and adaptation effectiveness before holistic testing, ensuring both components contribute meaningfully to overall system outcomes.

User studies should involve non-adaptive, randomly adaptive (inaccurate affect recognition), and adaptive (accurate affect recognition) versions of the system to evaluate the added value of adaptation. Combining quantitative and qualitative methods ensures a more comprehensive analysis. Quantitative metrics similar to those used by Gupta et al. [97] can validate the system's technical performance and cognitive and emotional aspects of user experience. In contrast, qualitative insights from user interviews capture subjective experiences. Feedback loops between inference models and adaptation mechanisms should be evaluated for response latency, robustness, and user satisfaction. Longitudinal studies are another important consideration, as they assess the long-term impact of the systems. Finally, evaluations should prioritize real-world applicability by extending testing beyond controlled environments. Based on these insights, we propose the seventh guideline (G7): evaluations should compare adaptive, non-adaptive, and randomly adaptive systems using mixed-method approaches, incorporate longitudinal and real-world testing, and assess feedback loops for latency, robustness, and user satisfaction to establish ecological validity and long-term effectiveness.

2.4.2 Critical Components

Apart from the lesson learned and suggested design guidelines, some critical components within psychophysiological inference and system adaptation need careful consideration. They include accounting for individual variability of physiological responses and deciding the extent of user involvement in system adaptation.

Individual Variability and Personalized Models

Variability in individuals' psychophysiological responses is a significant challenge in designing systems with biocybernetic adaptation. It is influenced by personal factors such as age, gender, mood, experience, and cultural background [133]. For example, variability was found in users' reactions to stimuli, suggesting that unique personal attributes lead to inconsistent physiological responses [260]. This variability often results in the inference models performing well in subject-dependent scenarios but struggling to generalize across different users. This issue can be addressed by having refined normalization and standardization techniques or personalized models for individual users.

Personalized, user-dependent models, which account for each individual's unique physiological and psychological characteristics, generally outperform user-independent models that combine data from diverse individuals [185]. For instance, deep neural networks have shown higher accuracy in detecting emotional states when tailored to individual users, with ECG features enhancing arousal and valence recognition [11]. Even techniques, such as clustering individuals based on similar mood perception [122] or preferences [55, 57], have proven effective, allowing for high accuracy with reduced computational demands. However, subject-specific datasets are smaller, and personalized models often underperform generalized models, highlighting the necessity for a diverse subject pool to improve reliability and generalizability [237].

The Physiological Feedback Spectrum

As highlighted throughout this chapter, systems with biocybernetic adaptation are derived from biofeedback technology but differ fundamentally in that they use physiological signals to automatically drive system-level adaptations, often without requiring active user intervention. This distinction can be used to categorize physiologically adaptive systems based on the degree of user involvement in interpreting and responding to physiological data. This categorization can be visualized in the form of a spectrum or continuum. On one end is explicit feedback, where users actively engage with the data, while implicit feedback lies on the other, with systems adapting autonomously to the user's state. Between these extremes are guided feedback, offering system-generated recommendations, and interactive feedback, where the user's actions and the system's responses dynamically influence each other. These categorizations are described in detail below.

Explicit Feedback: At the left end of the spectrum lies traditional biofeedback, where physiological signals are measured and presented to users to self-regulate through audio [230], visual [191], or haptic cues [167]. For instance, in exposure therapy, a user's real-time stress levels could be visualized as a progress bar, helping them understand their physiological responses to specific stressors. By monitoring these visualizations, users can consciously practice relaxation techniques or adjust their exposure strategy to control their stress levels. This approach requires active engagement, as users must interpret and act on the feedback to achieve the desired outcome.

Guided Feedback builds upon explicit feedback by pairing physiological signals with system-generated recommendations or guidance [72]. Users are provided actionable insights based on the signals, blending self-regulation with external support. For instance, a fitness app monitoring heart rate might suggest taking a break during a workout if it detects elevated stress levels. By providing clear instructions, the system reduced the cognitive load on users, allowing them to focus on following the guidance rather than interpreting raw physiological data. This approach bridges the gap between user agency and system intervention.

Interactive Feedback introduces dynamic engagement, where physiological signals influence real-time changes in the environments or interface, creating a bidirectional interaction between the user and the system [143]. For instance, consider a collaborative brainstorming tool used in a corporate setting. During the virtual brainstorming session, the system monitors participants' physiological signals and detects signs of stress or disengagement in one or more participants. In that case, it might alter the session dynamics by suggesting a quick relaxation break, reshuffling teams, or providing prompts to reframe the discussion topic. Participants can also actively influence the tool by providing verbal or gestural input.

Implicit feedback minimizes user involvement by enabling the system to adapt its behavior based on physiological signals, without requiring explicit interpretation or action from the user. This mechanism is central to systems with biocybernetic adaptation, where system responses are seamlessly aligned with the user's psychophysiological state. For example, in a noise-canceling office environment designed to enhance productivity, wearable sensors may continuously monitor employees' stress levels and focus states. Upon detecting elevated stress, the system can autonomously adjust environmental parameters, such as increasing noise suppression, reducing intrusive sounds, or subtly modifying ambient conditions like lighting. Additional environmental cues, such as gentle nature-inspired sounds, may be introduced to lower cognitive load and promote relaxation. These adaptations occur without requiring user awareness or direct engagement, thereby supporting unobtrusive and continuous regulation of the user's experience.

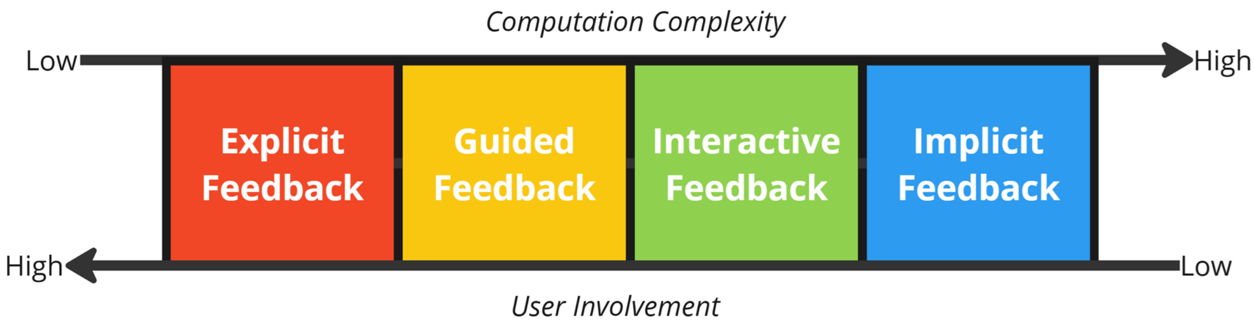


Figure 2.7: Physiological feedback spectrum based on user involvement and computational complexity.

2.4.3 Emerging Areas

Systems with biocybernetic adaptation are rapidly evolving, with emerging areas pushing the boundaries of real-time physiological signal processing and adaptive system design. Among these, Pervasive Biocybernetic Adaptation explores systems designed to operate unobtrusively in dynamic and uncontrolled environments. Simultaneously, biocybernetic adaptation can enable the development of Empathy-Enabled Technologies that facilitate emotionally intelligent human-to-human and human-to-machine interaction. These emerging areas are described in more detail below.

Pervasive Biocybernetic Adaptation

Pervasive systems with biocybernetic adaptation can seamlessly integrate into everyday applications to enhance user experience, performance, and well-being. Examples of such systems in the reviewed papers included the Calmi system, where PPG, EDA, and ST were used to infer stress levels and dynamically adapt to environmental factors in a smart home, including lighting, soundscapes, and scent for stress management [143]. Such systems can also be used in workplaces to mitigate fatigue and improve productivity by incorporating similar ambient adaptations and adjusting workstation ergonomics [15].

In education, such systems can monitor student engagement and dynamically adjust the difficulty of learning materials to maintain focus and optimize outcomes. Wearable fitness devices can use biocybernetic adaptation principles to increase user engagement by providing feedback on physical activity, such as suggesting adjustments to workout intensity. Systems like the Athlete Companion AI monitor fatigue and hydration levels to optimize training regimens and prevent overexertion. In assistive technologies, such systems can be used on wheelchairs to adjust seating angles based on muscle tension inferred from the user's physiological signals.

Despite their promise, pervasive systems with biocybernetic adaptation face significant challenges in real-world deployment. Accurate psychophysiological inference can be compromised by environmental noise, spontaneous fluctuations in physiological signals, and the variability inherent to dynamic, uncontrolled settings [260]. Furthermore, the effectiveness of such systems depends on carefully calibrated adaptations; overly frequent, intrusive, or poorly timed adjustments may lead to user fatigue, reduced engagement, or diminished trust in the technology. Addressing these challenges is therefore critical to ensuring the reliability, usability, and long-term acceptance of biocybernetic adaptation in everyday contexts.

Empathy-Enabled Technologies

Empathy is broadly defined as the capacity to understand and share the feelings of others, and is commonly conceptualized in three forms: emotional, cognitive, and compassionate empathy. Emotional empathy refers to directly sharing or resonating with another person's emotions, cognitive empathy emphasizes perspective-taking and understanding another's viewpoint, and compassionate empathy integrates both to motivate supportive action. Within Human-Computer Interaction (HCI), much of the research on empathy has traditionally focused on systems, particularly VR and AR, as "empathy machines" designed to elicit empathy in users. However, systems can also be imagined as an "empathic entity" that empathizes with users by customizing the user experience, an approach reminiscent of systems with biocybernetic adaptation.

Empathy-enabled technologies integrate these two perspectives by designing systems that not only foster human-to-human empathy but also extend empathy to human-to-computer interactions [60]. Such technologies can be regarded as a natural evolution of biocybernetic adaptation, building upon the ability to recognize and respond to users' psychological states. When integrated with Artificial Intelligence (AI), such technologies advance beyond mere recognition and adaptation, positioning AI as an empathic agent capable of more nuanced inferences and responses [59]. Together, these advancements position empathy-enabled technologies as a critical next step in designing human-centered systems that are adaptive, intelligent, and capable of fostering meaningful emotional connections.

To achieve this, empathy-enabled technologies can leverage both verbal and non-verbal cues, including tone of voice, prosody, facial expressions, body language, visual elements, color, lighting, and music, to more responsive interactions with users [6, 45, 61, 84, 97, 128, 157, 169, 187, 226, 227]. Their applicability spans diverse domains, including customer service [106], healthcare [145], personal assistants [128, 145], and privacy-sensitive contexts [56, 93], where they enhance engagement, trust, and satisfaction through empathically informed responses. Another representative example can be found in mental health, where a virtual therapist may integrate physiological signals such as heart rate with linguistic features to infer stress levels and provide personalized coping strategies.

2.5 Conclusion

This chapter systematically reviewed the state-of-the-art research on psychophysiological inference of stress and affect and physiology-driven system adaptation, the two foundational components of systems with biocybernetic adaptation. The review highlighted the peripheral physiological signals that are widely used for stress and affect recognition, supported by preprocessing, feature extraction, and modeling methods ranging from traditional ML to advanced DL approaches. It examined the importance of dataset quality, ecologically valid stimuli, and labeling strategies in ensuring the reliability and generalizability of inference models.

In parallel, the review analyzed physiology-driven system adaptations across diverse contexts, including training, mental health, and entertainment, using the Medium, Application, Psychophysiological State, and Adaptation mechanism (MAPA) framework [172]. Adaptation mechanisms were shown to vary across physical, virtual, and hybrid mediums, with psychophysiological states such as stress, anxiety, valence, and arousal serving as key triggers. Evaluations underscored the need for rigorous, multi-phase assessment of both inference accuracy and adaptation effectiveness, with real-world and longitudinal studies being particularly important for ecological validity.

From this review, seven design guidelines (G1-G7) were synthesized to inform the development of systems with biocybernetic adaptation. These guidelines emphasize: (1) training inference models with application-relevant stimuli and reliable labeling; (2) leveraging advanced modeling techniques to ensure accuracy and robustness; (3) selecting mediums that balance ecological validity, adaptability, and usability; (4) aligning psychophysiological states and signals with application goals; (5) prioritizing validated psychophysiological inferences over raw physiological triggers; (6) adopting multi-phase evaluations of inference and adaptation; and (7) employing mixed-method, longitudinal, and real-world evaluations for effectiveness.

Finally, this chapter explored emerging directions, particularly pervasive biocybernetic adaptation and empathy-enabled technologies. These areas point toward the integration of biocybernetic principles with pervasive computing and AI-driven empathy, broadening the scope of applications from stress management to emotionally intelligent, human-centered technologies. While challenges remain, the reviewed literature provides a roadmap for advancing robust, adaptive, and empathic systems. Building on these insights, the following chapters will focus on the design, development, and evaluation of systems that incorporate biocybernetic adaptation for daily stress management.

Chapter 3

Wearables meets LLMs for Stress Management

3.1 Introduction

As discussed, traditional approaches to stress management, such as therapy and structured programs, are valuable but often constrained by accessibility, cost, and the lack of immediacy. Wearable apps have emerged as promising tools for stress tracking, offering real-time monitoring of physiological stress indicators to provide timely insights into stress levels [86, 136, 222]. Complementing this, generative AI has been used to develop on-demand mental health support chatbots [138, 150]. While LLM-driven solutions have shown promise in supporting mental health diagnosis [180] and providing mental health first aid [115], they often exhibit behaviors typical of low-quality support [38]. Recent LLM-powered journaling approaches integrating passively collected behavioral patterns have enhanced self-awareness and well-being [177]. However, passive smartphone data may miss momentary events (e.g., anxiety, frustration) that wearables may catch from physiological arousal.

This chapter takes the first steps towards investigating the opportunities and challenges in integrating psychophysiological inferences with LLM chatbots to provide personalized interventions. A duoethnographic methodology, a collaborative autoethnography involving two researchers, was employed to capture detailed, reflexive accounts of integrating these technologies. In a 22-day study, two PhD students used CuesHub [44], an app for smartwatches and smartphones that provides real-time, wearable-triggered stressor journaling. Simultaneously, both researchers used a personalized LLM-based stress intervention chatbot built with OpenAI's custom GPT. Each researcher developed prompt templates tailored to their individual stress triggers, coping styles, and daily routines, using a lightweight framework informed by contemporary stress-management strategies. This personalized approach allowed a more nuanced exploration of how the integration of wearables and LLMs can be adapted to user needs and preferences.

The motivation is to harness the complementary strengths of wearables and LLMs to develop an integrated stress management solution. This research focuses on two questions: (1) How often do users seek interventions for momentary events triggered by wearables? and (2) How does integrating very brief descriptions of momentary events into LLMs enhance stress management interventions? By documenting lived experiences, the aim is to inform the design of an integrated system that can track stress in real time, provide interventions, adapt to the complexities of individual users' needs, and enhance the user experience. The final goal is to articulate design considerations and insights to guide the design of an integrated physiology-driven stress-management system.

Recent advances highlight the potential of integrating wearable apps with LLMs to support both continuous health monitoring and personalized therapeutic interventions. On the monitoring side, Kim et al. [124] proposed Health-LLM, showing that context-enhanced prompting enables LLMs to perform multi-task health prediction across domains such as mental health, cardiac monitoring, and sleep. Fang et al. [75] presented PhysioLLM, an interactive system that integrates statistical analyses of wearable data with LLM reasoning to enhance personalization and user engagement. On the therapeutic applications side, Yang et al. [250] developed ChatDiet, which combines wearable and user data with population nutrition models to provide personalized dietary recommendations with high user alignment, and Ahmed et al. [7] demonstrated the feasibility of using LLMs to analyze Fitbit and survey data to generate personalized well-being recommendations for students.

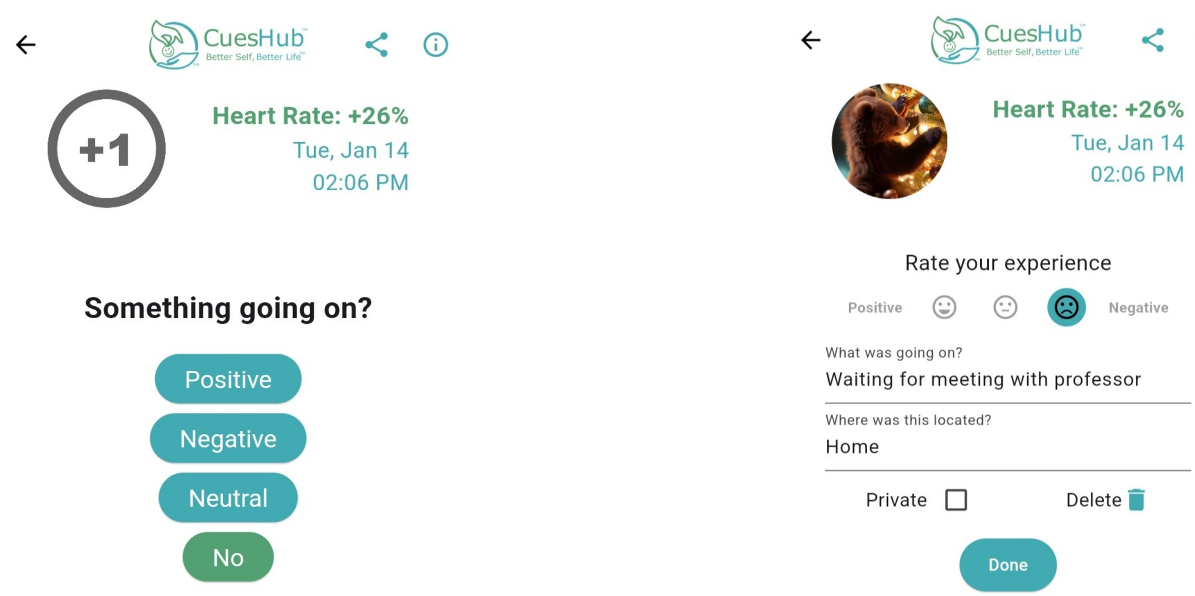
3.2 Methods

3.2.1 Duoethnography

This study employs a *duoethnographic approach*, a qualitative method that explores how two researchers interpret and make sense of shared experiences [40, 211]. It builds on autoethnographic methods [43, 67, 181, 234], which have been applied in diverse contexts, including the use of LLM chatbots for thesis writing, creative expression, and to understand the humanistic elements of self-identity [88, 182, 214]. For this study, two researchers independently interacted with the wearable app and personalized LLM chatbots for stress management, documenting their experiences and reflections in autoethnographic diaries. This approach provided introspective data on how integrating the two systems influenced perceived effectiveness in managing everyday stress. Weekly meetings between the researchers enabled shared reflection, discussion, and comparative analysis of individual perspectives. The study's insights into integrating wearable apps and LLM chatbots for stress management were enriched by combining individual depth with collaborative breadth.

3.2.2 Wearable App for Stressor Monitoring

Both researchers wore Samsung Galaxy Watch 6 devices equipped with the CuesHub smartwatch app [44], serving as the wearable AI platform for detecting physiological events associ-



(a) Valence rating screen of the prompted event

(b) Event description screen of the prompted event

Figure 3.1: CuesHub app screenshots for recording valence and descriptions for events detected by the smartwatch app

ated with stress. Upon detecting a physiological stress event, the smartwatch app prompted researchers via their smartphones, requesting them to provide brief descriptions of the event. Researchers rated the events using one of four options: *Positive*, *Negative*, *Neutral*, or *No* (see Figure 3.1a). For responses indicating *Positive*, *Negative*, or *Neutral*, researchers were further prompted to answer two questions: *What was going on?* and *Where was this located?* (see Figure 3.1b).

3.2.3 Custom LLM Chatbots for Stress Intervention

The LLM chatbots employed in this study used OpenAI's GPT-4 [1] as their underlying language model. To investigate the effectiveness of different interaction paradigms, we explored two contrasting approaches to chatbot design, emphasizing distinct user experiences and

[1]https://openai.com/index/gpt-4/

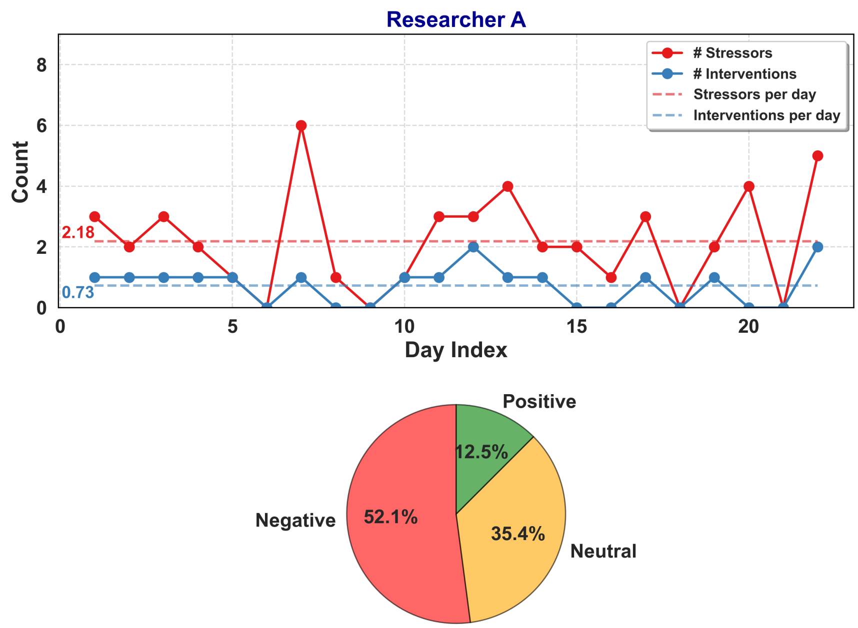
intervention delivery styles. *Researcher A (R_A)* designed their chatbot, named *DeStressify*, to emulate a zero-shot prompting approach akin to intervention systems that automatically deliver stress interventions upon detecting a user's stress [19, 109]. When experiencing the need for intervention, *Researcher A* prompted the chatbot with their stressor and location. Most interactions were transactional, consisting of a single prompt-response exchange where the chatbot provided an intervention in the traditional instructive manner often used in text messaging.

Researcher B (R_B), on the other hand, adopted an end-of-day conversational approach for their chatbot, named *StressGPT*. This chatbot was designed to facilitate a more therapy-like interaction, enabling users to engage in dynamic, multi-turn dialogues. This conversational style aimed to simulate collaborative interactions, with *StressGPT* mimicking personalized coaching sessions by refining suggestions or providing layered support through iterative exchanges, enabling deeper reflection and tailored advice.

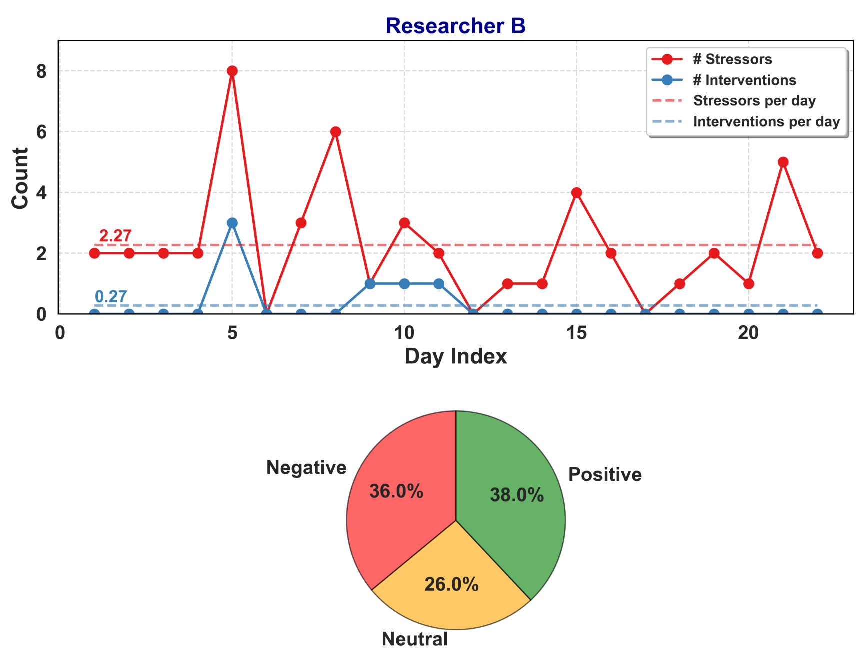
3.2.4 Data Collection and Analysis

Both researchers used the CuesHub app for 22 days. For stress events where researchers felt the need for interventions, they voluntarily engaged with the LLM chatbots. R_A, using *DeStressify*, engaged with the chatbot immediately (or soon) after prompts in the CuesHub app, i.e., following a *just-in-time zero-shot* approach. R_A incorporated the most recently logged stressor and location from the CuesHub app into the chatbot prompt for each interaction.

R_B, in contrast, engaged with *StressGPT* predominantly at the end of the day, consolidating the day's events and reflectively interacting with *StressGPT*. R_B used the logged stressor entries from the CuesHub app as a reference to interact with the chatbot. Please note that even in the end-of-day conversational approach, prompts helped capture events that might



(a) Number of Stressors and Interventions per day and Valence Proportion of Events Detected by CuesHub for R_A



(b) Valence Proportion of Events Detected by CuesHub for each researcher

Figure 3.2: Frequency of stress events and interventions per day and valence distribution of events detected by the CuesHub app

have been overlooked.

After each engagement, researchers rated the intervention on a scale from *Very poor* to *Very good*. Ratings were based on the immediate relevance and usefulness of the interventions in addressing specific stress events. Both researchers employed distinct prompt engineering strategies, guided by catalog patterns described in [241], ensuring tailored and meaningful interactions with their respective LLM chatbots. Both researchers maintained a running journal after each chatbot interaction, reflecting and documenting their perceptions of its relevance, clarity, and impact.

Both researchers conducted a thematic analysis of data from weekly meeting transcripts, journals, and stress events logged through the CuesHub app to examine chatbot usage, intervention effectiveness, user experience, and the connection between wearable-detected stress events and chatbot interactions. The analysis included design documentation detailing chatbot customization and expectations, daily journals reflecting on the relevance and effectiveness of interactions, and engagement logs tracking the frequency and content of interactions.

3.3 Results

The results first describe the number of stress events detected by the wearable app and assess their usefulness by comparing the intervention experiences of the just-in-time zero-shot and end-of-day interactive approaches. Later, the qualitative findings from individual researchers' experiences and reflections in autoethnographic diaries are discussed.

3.3.1 Stress Detected and the Need for Interventions

Diversity in Stress Events Detected

The stress events detected by the app and recorded by the researchers ranged from work-related challenges, such as preparing complex drafts or debugging code, to social stressors, like navigating difficult conversations or attending gatherings. Daily life stressors, including routine tasks such as cooking or managing unexpected delays, were also captured, as were positive stressors, such as celebrating achievements or enjoying successful interactions. Work-related and social stressors often evoked a sense of urgency and anxiety, whereas daily-life stressors tended to accumulate gradually, creating cumulative stress. Positive stressors, on the other hand, brought excitement and a sense of fulfillment, often tied to personal growth or future rewards. The researchers noted that the detected stress event varied in intensity and impact.

Proportion of Stress Events requiring Interventions

The researchers logged 98 events (48 by R_A and 50 by R_B). Of these, 43, 30, and 25 events were negative, neutral, and positive, respectively. Figure 3.2 shows the number of stress events and how many were used for interventions on each day during the study. Of the 98 events, the researchers felt the need to engage with the chatbot in 22 events ($R_A = 16$, $R_B = 6$). Of those, 20 were negative, and 2 were neutral. Researchers required an intervention on only half of those days when at least one stress event was reported. The rating distributions of interventions were: for R_A, *Very poor* (6.25%), *Poor* (18.75%), *Acceptable* (25.00%), *Good* (43.75%), and *Very good* (6.25%); for R_B, *Acceptable* (16.67%), *Good* (66.67%), and *Very good* (16.67%).

Interestingly, certain negative events, such as *"Working on a paper," "Rushing to catch a flight," "Completing pending work,"* and *"Feeling hungry,"* did not require intervention. This was attributed to various factors, including the perception that the event was manageable without assistance, the stressor's temporary or fleeting nature, or a personal preference for self-reliance in handling the situation. This observation highlights that merely detecting a stress state is insufficient; recognizing the specific stressor is critical for delivering tailored interventions. Additionally, stress events such as *"coding issues"* sometimes warranted intervention, but not consistently across all occurrences. This highlights that simply identifying a stress event may be insufficient; participant feedback or additional contextual cues are necessary to determine when an intervention is genuinely warranted.

Number of Interventions Required

R_A primarily used *DeStressify* in real-time immediately after logging stressors that required interventions, whereas R_B engaged in conversation with *StressGPT* towards the end of the day. Figure 3.2 illustrates that while the average number of stressors logged per day was comparable for R_A (2.18) and R_B (2.27), R_A engaged with interventions almost three times more frequently per day (0.73) compared to R_B (0.27). This suggests that real-time interventions are more likely to be utilized when stressors are momentary. Factors such as the transient nature of specific stressors may make immediate intervention more relevant and actionable, providing support precisely when needed. In contrast, engaging with *StressGPT* at the end of the day may reduce the perceived need for interventions, as some stressors may have already dissipated or been resolved by then. Although stressors such as *"meeting with professor"* or *"long wait time"* were common for both researchers, only R_A asked for interventions.

3.3.2 User Experience with the Integrated Systems

This section highlights common themes that emerged from both researchers' experiences with the wearable app and their personal LLM chatbots (*DeStressify* and *StressGPT*), offering insights into the similarities in how these systems functioned and were perceived during interactions.

Need for Continuity and Stressor Awareness in Chatbots

The LLM chatbot's effectiveness was significantly hindered by its inconsistent integration of continuity and contextual awareness, despite having access to the stressors. While it occasionally referenced past interactions, it failed to do so regularly, making conversations feel less personalized and lacking a sense of ongoing support. Rather than tailoring responses to the history of stressors shared with the LLM chatbot, it often provided generic solutions that did not address the evolving nature of the user's challenges. R_B noted, *It felt like the chatbot did not really know who I am and would have given similar responses to anyone asking similar questions.* This absence of context integration underscores the need for LLM chatbots to track and link past stressors to current interactions, thereby enabling a more dynamic, responsive system that delivers consistent, relevant support.

Striking Balance Between Emotional Support and Practicality

A key insight was the need for the LLM chatbots to balance emotional and practical support. While practical strategies like time management and breathing exercises were helpful, they were incomplete without emotional support. Purely emotional responses, without actionable advice, were ineffective in addressing stressors. R_A valued interventions that reframed emotional states, such as linking frustration to the PhD journey, while R_B appreciated the chatbot's emojis, which felt like a human-like touch to the interaction. This underscores the need for a holistic approach in which chatbots provide both emotional support and practical guidance, ensuring that users feel fully supported in managing their stress.

Clarity and Precision in Chatbot's Responses

Both researchers preferred concise responses from the LLM chatbot that directly addressed their issues, as these reduced cognitive load and maintained efficiency in the interaction. R_B noted that the chatbot often provided long, structured responses to simple questions, which made the conversation feel more mechanical and less human-like. On the other hand, R_A particularly appreciated interventions that offered an explicit, singular action or suggestion rather than multiple options or vague guidance. This preference for specificity was especially apparent in emotionally charged or complex situations, where a precise, direct solution was more helpful than a broad range of solutions [38].

Targeted Interventions vs Human-Like Conversations

R_A found that the wearable app, paired with the LLM chatbot, provided targeted interventions tailored to specific stressors with practical, actionable strategies. For work challenges, breaking tasks into smaller steps improved clarity, while physical activities and environmental changes helped reset mental states. Social stressors were addressed through confidence-building interventions, like affirmations and reframing negative thoughts. Positive perspectives on frustrating situations also enhanced engagement and effectiveness. For example, targeted interventions, such as structured debugging for stressors like replicating a paper, were more effective than generic strategies, such as Progressive Muscle Relaxation (see 9.1.2). In contrast, generic responses, such as mindfulness exercises or repeated advice (e.g., "take a deep breath"), often caused frustration, thereby reducing the chatbot's perceived usefulness.

Although R_B found *StressGPT* to be more human-like than other previously used rule-based mental health chatbots. It often behaved like a general-purpose LLM chatbot, responding to queries outside the stress management domain, thereby reducing its relevance. Additionally, it was quick to offer suggestions without fully understanding the user's stressor or context, generating lengthy responses without asking clarifying questions. This lack of active listening made the interaction feel less personal and empathic. The structured, verbose responses felt robotic and detached, resembling those of a general-purpose LLM chatbot rather than a specialized stress-management assistant. Furthermore, rapid text generation disrupted the conversational flow, making it difficult to remain engaged as responses were produced before they could be fully read.

Privacy Concerns

For *DeStressify*, R_A only shared stressors that did not carry sensitive or risky identifiable information. *StressGPT* raised privacy concerns for R_B, especially when discussing sensitive or deeply personal topics. R_B expressed hesitation in fully opening up during interactions, fearing that private information might not be adequately protected or might be misused. This reluctance underscores the need for LLM chatbots to establish trust and security, ensuring that users feel confident sharing personal information without fear of data breaches or misuse. LLM chatbots for mental health must ensure privacy, clear data policies, and transparency, thereby fostering trust and engagement and enabling more effective, personalized support.

3.4 Discussions

This section discusses the opportunities and challenges of using wearable apps with LLM chatbots for stress management, laying the groundwork for developing an integrated system.

3.4.1 Opportunities in Integrating Wearables and LLMs for Stress Management

Prior research underscores the importance of interventions tailored to users' specific stressors [109, 231]. The findings of this study highlight the potential of wearable apps such as CuesHub, which combine physiological data with contextual cues, to personalize interventions. By integrating insights from wearable apps into LLMs, stress management systems can provide more precise and responsive support that aligns with users' needs. Such systems must transition from automatic, event-triggered responses to more selective approaches that prioritize user needs and preferences. By analyzing stressor descriptions, systems can more effectively distinguish events that require intervention, thereby improving relevance and impact.

This study explored two approaches to engaging with the LLM chatbot: real-time and reflective interactions. In the real-time approach, delays in delivering interventions can significantly reduce their effectiveness, as stressors often require immediate attention. Although the reflective approach can tolerate delays, it may still benefit from faster interactions, as timely responses can help users process their experiences more effectively. By incorporating proactive and predictive mechanisms, a system integrating wearables and LLMs can anticipate potential stressors and intervene early, shifting from reactive to preemptive support and enhancing overall stress management.

We also observed that the LLM chatbots sometimes successfully generated interventions based on previously shared events but were limited to events for which users explicitly requested support. Expanding its access to the whole history of wearable-detected events could provide more context, enabling more personalized and informed interventions. Moreover, integrating memory and user history into LLM chatbot systems could significantly enhance interaction continuity, enabling more personalized and consistent support across multiple sessions. This continuity can foster a deeper sense of trust and connection between the user and the chatbot, which is essential for building long-term engagement and effectively addressing recurring stressors.

Lastly, our findings highlighted the potential of a physiology-driven, emotion-aware chatbot to detect heightened stress cues and respond with a combination of practical solutions and emotional comfort, while features like adaptive dialogue could further enrich the user experience. Stress management chatbots could be improved by emulating human-like interactions, incorporating empathic communication and emotional intelligence, thereby better addressing users' emotional needs and fostering more compassionate, relevant exchanges. To conclude, this study highlights both the critical need and the opportunity to integrate wearable apps with LLM chatbots to predict stress and to align interventions with the stressor, ensuring meaningful support without adding unnecessary cognitive load.

3.4.2 Challenges in Integrating Wearables and LLMs for Stress Management

Advances in wearables and LLMs capable of generating adaptive interventions hold significant promise for stress management. A natural assumption might be that seamlessly connecting wearables to LLMs could create the ultimate stress management system. However, this study's findings revealed several challenges in effectively integrating these technologies.

While LLMs excel at generating one-time interventions and engaging users in extended conversational support, their integration with wearables raises significant questions about whether their responses align with users' descriptions of stress events. Firstly, our findings reveal a discrepancy between stress detected by wearable AI and the user's actual need for intervention. Of 98 confirmed stress events, only 22 (approximately 20%) required an intervention, indicating that delivering an intervention for every detected event may be counterproductive.

The wearable app detects both positive and negative stress events. Interestingly, not all negative events require intervention. In this study, fewer than half of the negative events prompted a desire for intervention. Systems that push out interventions based solely on stress detection (even if the AI were powerful enough to distinguish negative events from positive or neutral ones automatically) can create unnecessary burdens, leading to disengagement and reduced utility.

3.5 Study Limitation

This study has several limitations that impact its generalizability. The small sample size of two researchers, both with expertise in wearables and LLMs, introduces potential bias and restricts broader applicability. Additionally, the researcher's perspectives, shaped by both optimism about AI's potential and caution regarding ethical concerns, may have influenced our interpretation of results. The absence of control groups and standardized conditions limits our ability to isolate the effects of wearable-triggered interventions. Furthermore, the study relies primarily on subjective experiences and qualitative data, making it difficult to assess the effectiveness of the intervention objectively. While we have expertise in AI and wearable technologies, our lack of formal training in mental health presents another limitation.

3.6 Conclusion and Future Works

This study demonstrates the potential of integrating wearable apps with LLM chatbots to offer valuable support for stress management, while also highlighting key areas for improvement for such an integration. The findings indicate that the perceived effectiveness and impact of interventions are heavily influenced by factors such as timing, personalization, and contextual relevance. Although the LLM chatbots provided valuable assistance, limitations such as repetitive suggestions, lengthy responses, and limited retention of past interactions often compromised their ability to provide truly impactful support. Looking ahead, systems that integrate wearables with LLMs have the potential to become more adaptive, effective, and user-centered, offering real-time, comprehensive, and personalized support for stress management.

Chapter 4

Physiology-Driven Empathic LLMs for Stress Management

4.1 Introduction

The previous chapter demonstrated the opportunities and challenges of manually integrating wearable apps for stress monitoring with LLM chatbots for stress interventions as a holistic system for stress management. This chapter introduces the physiology-driven Empathic Large Language Models (EmLLMs) framework, which automatically integrates wearables with LLM chatbots. Physiology-driven EmLLMs aim to tightly couple users' internal states with the generative capabilities of LLMs, enabling real-time empathic interaction grounded in the user's state. The physiology-driven EmLLM chatbot prototype for stress management integrates users' stress to inform the LLM's conversational strategies.

Physiology-driven adaptation offers several advantages for stress management. First, physiological signals offer reliable monitoring and insight into stress in everyday contexts and are less susceptible to conscious masking than behavioral cues. Second, integrating these signals with LLMs allows chatbots to recognize when interventions are genuinely needed, reducing the risk of over-notification and intervention fatigue. Third, embedding stress inferences facilitates empathic reasoning in LLM interactions, and EmLLMs can balance emotional validation with actionable guidance, thereby strengthening the therapeutic alliance that is

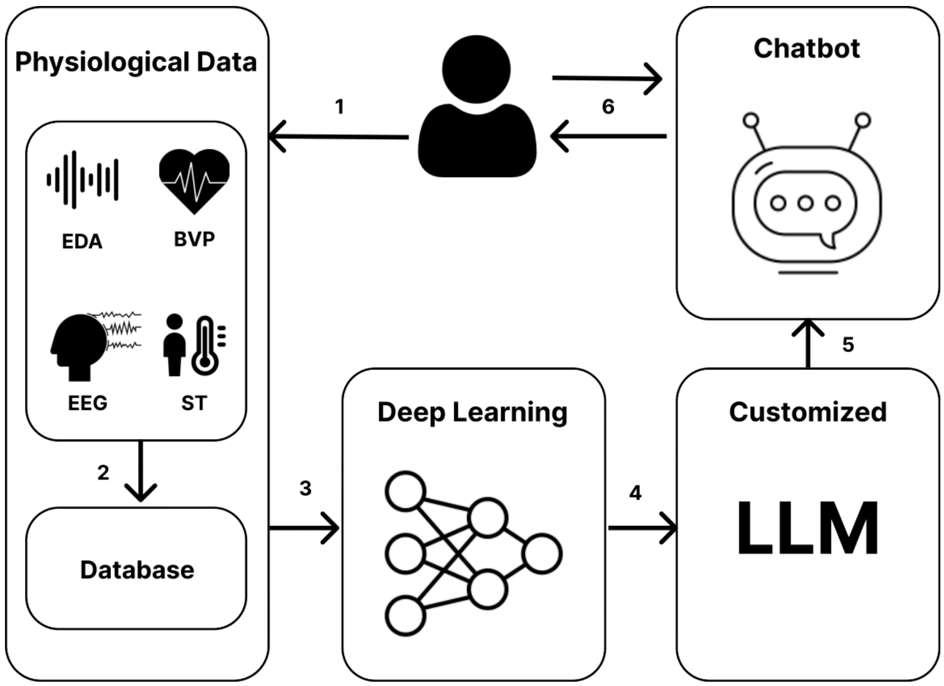


Figure 4.1: Proposed Empathic Large Language Model (EmLLM) Approach

often absent in traditional chatbots.

This chapter builds on insights from the duoethnographic study in the last chapter to outline the design, implementation, and preliminary evaluation of the physiology-driven EmLLM chatbot for daily stress management. We explore how physiology-driven dialogue adaptation can (i) enhance personalization in stress interventions, (ii) improve user engagement and trust, and (iii) support empathic and contextually relevant conversations. By grounding LLM behavior in psychophysiological evidence, the integrated system represents a step toward more human-centered and adaptive digital mental health systems that can address the complex, evolving needs of users in daily life.

4.2 The Physiology-driven EmLLM Framework

EmLLMs, similar to the empathy-enabled technologies 2.4.3, focus on using multimodal user data to facilitate empathy in human-human and human-AI interactions. EmLLMs are responsible not only for inferring users' state from multimodal user data, but also for generating relevant content that aligns with or supports that state. LLMs can infer users' states from the text, audio, visual, and wearable data collected from smartphones, wristbands, smartwatches, and immersive reality headsets. Based on the user's state, LLM content generation can include generating text, audio, video, and 3D objects to enhance the user experience or support the user. Physiology-driven EmLLMs specifically use physiological signals to infer users' states and generate relevant content. To this end, the design of physiology-driven EmLLMs is similar to that of systems with biocybernetic adaptation.

4.3 Prototype Development

Based on the EmLLM framework, a prototype physiology-driven EmLLM chatbot for stress management was developed. This chatbot passively monitors user physiological signals from the Empatica EmbracePlus smartwatch and provides end-of-day interventions [59]. The EmbracePlus smartwatch monitors users' Photoplethysmography (PPG), Electrodermal Activity (EDA), and Skin Temperature (ST) using optical sensors and stores the raw data in an Amazon S3 bucket.

Since LLMs specialize in language comprehension and generation and are not trained to infer psychological states from physiological signals, an end-to-end, multi-channel 1D deep CNN model was used to infer stress from psychophysiological signals for the LLM. The model's inferences are provided to the LLM chatbot as textual prompts to interpret user states from physiological signals. The inference model was trained and tested on the Wearable Stress and Affect Detection (WESAD) dataset [212] using a Leave-One-Subject-Out (LOSO) cross-validation approach to classify user states as stressed and non-stressed. The model was optimized using the Adam optimizer and the Binary Cross-Entropy loss function.

The physiology-driven EmLLM chatbot provided stress interventions by generating supportive text tailored to the user's state. To enable this, an open-source LLM, Falcon-7B [9], was fine-tuned using the Quantized Low-Rank Adaptation (QLoRA) technique [50] on a mental health dataset. This dataset was scraped from mental health websites and comprises conversational pairs, including patient questions and corresponding answers from mental health professionals. The fine-tuned model was then prompt-engineered to function as a trained psychologist, adhere to Cognitive Behavior Therapy (CBT) principles, and provide clarifications when it cannot answer a query. The chatbot, which integrates the model for psychophysiological inference of stress with the customized Falcon-7B LLM, was deployed via a web-based interface to enable personalized stress management.

4.4 Pilot Study

4.4.1 Study Protocol

An in-the-wild pilot study involving eight doctoral students (5 men, 3 women; age range 23-37, mean=30.5) who wore the Empatica EmbracePlus smartwatch throughout their workday and interacted with the chatbot at the end of the day. The study evaluated the chatbot's ability to accurately detect user stress, deliver human-like responses, and provide therapeutic support.

Participants were recruited from the Center of Human-Computer Interaction (CHCI) community at Virginia Tech based on the following inclusion criteria: (1) currently enrolled doctoral students, (2) no history of severe mental health conditions. Doctoral students were selected as the target population due to their known susceptibility to high cognitive load and chronic stress in academic environments [205, 210]. All participants provided informed consent and received training in the use of smartwatches and the chatbot before data collection.

At the beginning of each workday, participants completed a pre-task Short Stress State Questionnaire (SSSQ) [105] and were fitted with the smartwatch. They then resumed their daily academic or professional activities while continuously wearing the device, following instructions to avoid altering its position and to refrain from consuming mood-altering substances or engaging in strenuous exercise.

At the end of each workday, which lasted approximately 4-5 hours, participants removed their smartwatches and completed the post-task SSSQ. They then interacted with the EmLLM chatbot for at least 15 minutes. The chatbot greeted each user by name, presented their predicted stress level, and invited them to discuss any stressors they were experiencing. After the session, participants completed the Godspeed questionnaire [17], and the Session Rating Scale (SRS) [63], and were encouraged to provide additional qualitative feedback.

The SSSQ is an adaptation of the Dundee Stress State Questionnaire (DSSSQ) that measures engagement, distress, and worry, providing pre- and post-task stress indicators [105]. These indicators were compared with the chatbot's predictions to validate its accuracy. The Godspeed questionnaire, commonly used in human-robot interaction studies, assessed participant perceptions of intelligence, friendliness, and ease of use [17]. The SRS captured users' subjective session experiences by gauging therapeutic alliance, helping us determine the EmLLM chatbot's efficacy as a supportive tool [63].

4.5 Results

The results of this pilot study are categorized into four sections: (1) prevalence of high stress among graduate students; (2) the performance of the model for psychophysiological inference of stress; (3) the user experience with the physiology-driven EmLLM chatbot; and (4) the perceived quality and effectiveness of stress management intervention.

4.5.1 High Stress among Graduate Students

The SSSQ measures three psychological dimensions relevant to high-stress contexts: distress, engagement, and worry. A repeated-measures analysis was conducted to compare scores at two points in the workday (pre- and post-) to identify fluctuations in stressful states that graduate students encounter in real-world academic settings.

On average, distress scores were low to moderate both pre- and post-task ($D_{\text{pre}} \approx 1.7$, $D_{\text{post}} \approx 1.6$), suggesting relatively low negative affective arousal. Engagement scores were consistently high ($E_{\text{pre}} \approx 3.8$, $E_{\text{post}} \approx 3.9$), indicating that participants maintained attentional focus and involvement throughout. Worry scores began at a moderate-to-high level ($W_{\text{pre}} \approx 3.3$) and decreased post-task ($W_{\text{post}} \approx 2.6$), indicating a reduction in ruminative thought patterns over the workday.

The analysis revealed a statistically significant reduction in worry across the workday (mean difference ≈ -0.67, $p < 0.05$), suggesting that anticipatory and ruminative stress symptoms tended to subside as the day progressed. By contrast, engagement remained stable (mean difference ≈ 0.11, $p \approx 0.51$), indicating consistent levels of attentional focus and involvement, while distress showed no significant change (mean difference ≈ -0.09, $p \approx 0.78$). Together, these results suggest that although graduate students exhibit consistently high engagement

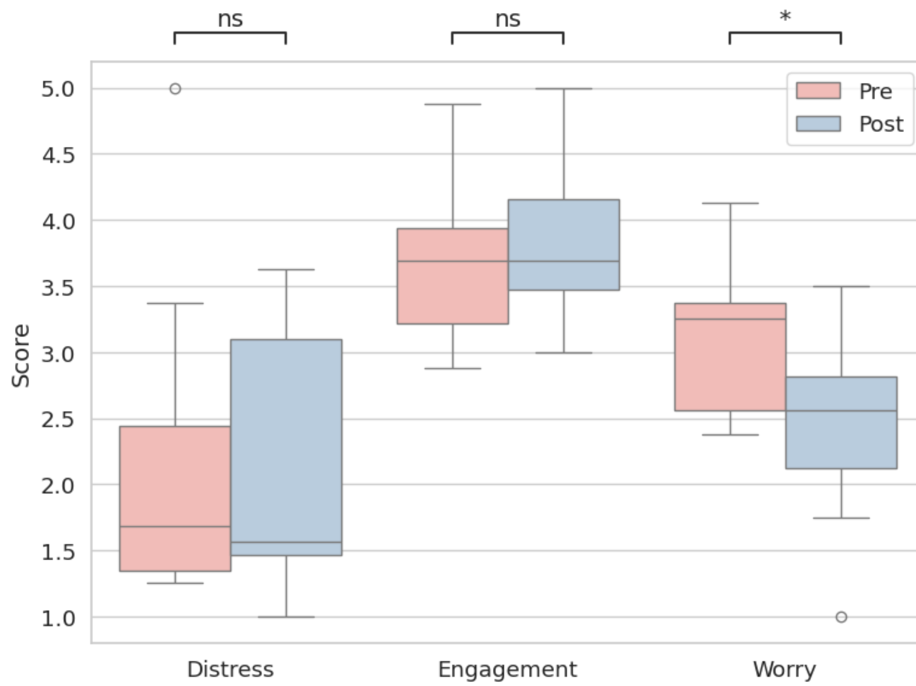


Figure 4.2: Average Distress, Engagement, and Worry States of all Participants at the beginning of workday (Pre) and end of workday (Post)

and low distress, they experience moderate-to-high levels of worry during their workday, with worry being the most dynamic and burdensome stress component (Figure 4.2).

4.5.2 Performance of Stress Detection Model

As discussed, the stress detection model was first trained and validated on the publicly available Wearable Stress and Affect Detection (WESAD) dataset. Using leave-one-subject-out (LOSO) cross-validation, the model achieved an accuracy of 0.85 across participants, demonstrating strong performance in distinguishing between stressed and non-stressed states under controlled conditions. This accuracy exceeds that of ML models developed using hand-crafted features.

To evaluate model performance in our real-world deployment, the SSSQ served as the ground-truth reference. For each participant, changes in engagement, distress, and worry between pre- and post-task sessions were used to compute a stress index ($\Delta D + \Delta W - \Delta E$). Participants were then manually classified as "Stressed" if their stress index was greater than or equal to zero, and "Not Stressed" otherwise. This procedure provided a transparent, questionnaire-based classification of workday stress experiences.

The SSSQ-based classification was compared with the model's inferences aggregated across the day. As shown in Table 4.1, the model correctly classified four of the eight participants, resulting in an overall accuracy of 0.50 in real-world deployment. This discrepancy highlights a notable performance gap between model accuracy in controlled training datasets and noisy real-world contexts. Specifically, the model tended to overpredict stress, classifying several participants as "Stressed" despite SSSQ scores indicating otherwise. These results emphasize that while stress detection models generalize well on benchmark datasets, their robustness under naturalistic conditions requires further refinement and calibration.

Table 4.1: Comparison of SSSQ-based and Model-based Stress Classification

Participant	**ΔE**	**ΔD**	**ΔW**	**Stress Index**	**SSSQ**	**Model (avg)**
1	-0.125	0	-0.625	-0.5	Not Stressed	**Stressed**
2	-0.25	0	-0.125	0.125	**Stressed**	**Stressed**
3	0.75	0.125	0	-0.625	Not Stressed	**Stressed**
4	0.125	-1.375	-0.875	-2.375	Not Stressed	**Stressed**
5	0.375	1.75	-0.375	1	**Stressed**	**Stressed**
6	0.5	-0.25	-1.625	-2.375	Not Stressed	Not Stressed
7	0.125	-0.25	-0.375	-0.75	Not Stressed	**Stressed**
8	-0.625	-0.75	-1.375	-1.5	Not Stressed	Not Stressed

4.5.3 User Experience with the Integrated System

As discussed, the Godspeed Questionnaire was used to assess the physiology-driven EmLLM chatbot's intelligence, friendliness, and ease of use. It evaluates participants' perceptions of the chatbot across five constructs: competence, responsiveness, human-like responses, conversational elegance, and pleasantness. The results indicate that the chatbot was perceived as reasonably competent ($M = 3.5$) and responsive ($M = 3.63$), suggesting that users perceived the system as able to understand inputs and provide timely responses. However, participants were less impressed with the chatbot's ability to produce human-like responses ($M = 2.88$) and conversational elegance ($M = 2.88$), indicating that interactions often felt artificial or mechanical rather than natural. Similarly, the pleasantness of the interaction was rated only moderately ($M = 3.0$), indicating a neutral-to-slightly positive affective impression of the chatbot.

Qualitative feedback provides further insights into these perceptions. A recurring concern raised by all participants was that the chatbot occasionally "started talking in Spanish,'' likely due to the presence of Spanish-language pairs in the fine-tuning dataset. This unexpected behavior detracted from conversational smoothness and coherence, contributing to lower scores in human-likeness and conversational elegance. Some participants also expressed surprise when the chatbot's stress recognition of their day contradicted their own expectations, indicating mismatches between physiological inference and self-perception. Additionally, participants expressed a desire for greater continuity and personalization, noting that the chatbot should remember past conversations and provide richer feedback drawn from smartwatch data. Privacy concerns also arose when one participant expressed that "someone would see" their conversations.

4.5.4 Quality and Effectiveness of Stress Intervention

As discussed, the SRS was used to assess participants' perceptions of the chatbot's stress intervention's quality and effectiveness. Overall, the chatbot demonstrated the ability to

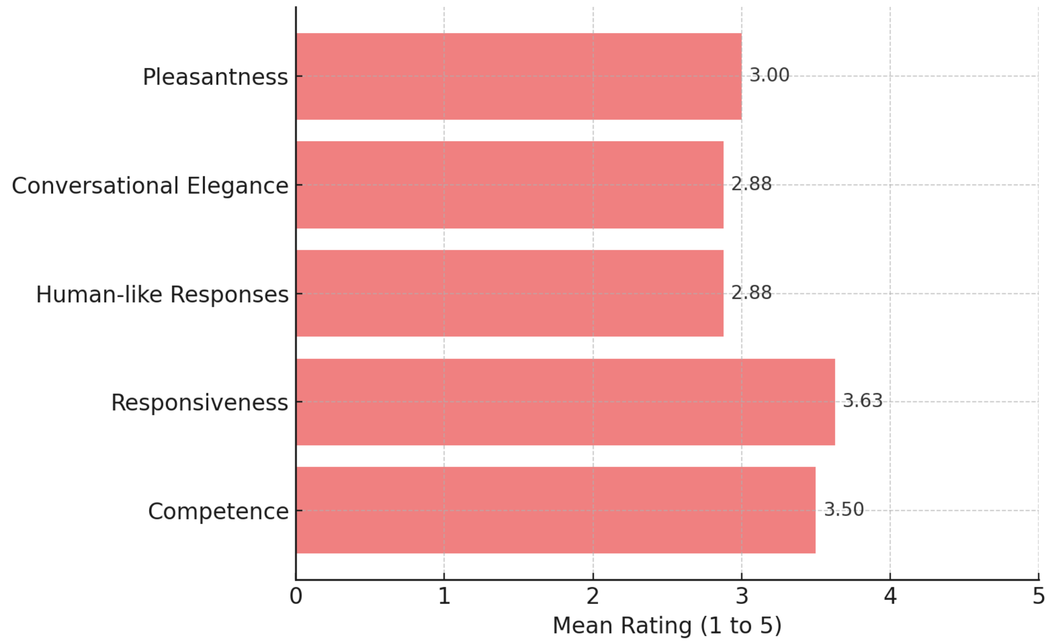


Figure 4.3: Mean Godspeed Questionnaire ratings (1-5 scale) for user perceptions of the chatbot across five constructs.

establish a positive therapeutic alliance with users. Participants rated the overall quality of the session as moderate ($M = 3.25$). They perceived the chatbot as relatively empathic ($M = 3.63$), suggesting that the system conveyed understanding and support during stressful moments. Participants also evaluated the intervention as relevant to their needs ($M = 3.75$), a reasonably good fit for their situation ($M = 3.38$), and appropriate for them ($M = 3.25$). These results indicate that, although there is room for improvement, the chatbot's interventions were generally well received and perceived as contextually meaningful.

Qualitative feedback further enriches this picture. Many participants described the EmLLM chatbot as "therapist-like" and valued its ability to provide "good counsel." They highlighted that the chatbot did more than provide generic advice; it asked probing questions that encouraged reflection and offered contextualized reasoning for its responses. This behavior contrasted positively with experiences of conventional chatbots such as ChatGPT, which participants felt delivered less personalized and more templated responses. From a stress intervention perspective, these qualities are crucial for fostering user trust and engagement and positioning the chatbot as a potentially effective digital tool for supporting stress management.

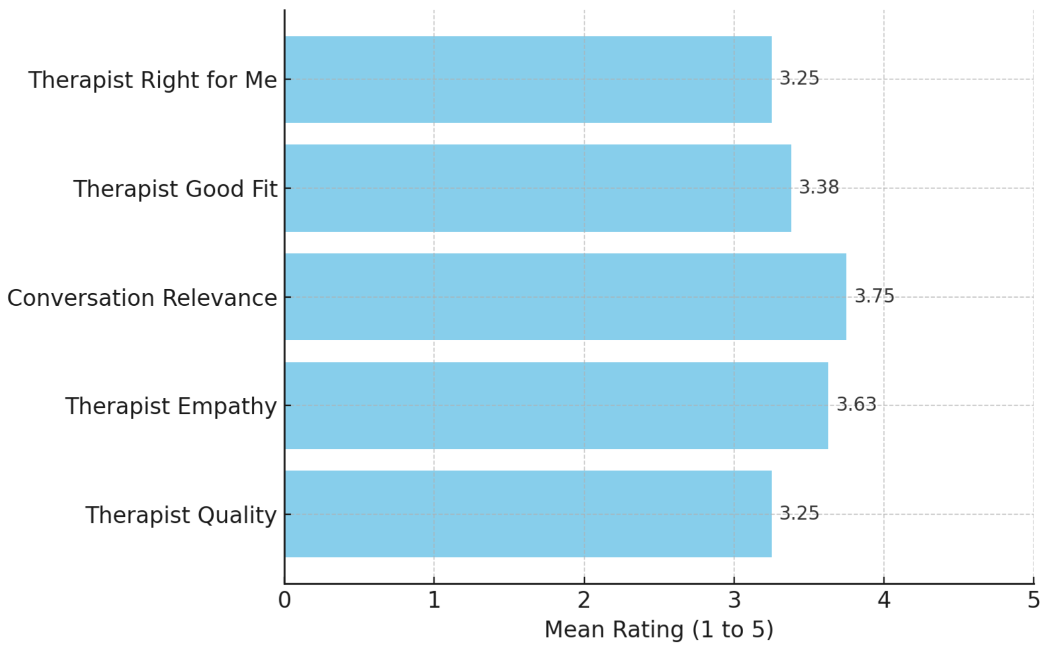


Figure 4.4: Mean SRS ratings (1-5 scale) for participants' perceptions of the chatbot's intervention quality and effectiveness across five constructs.

4.6 Discussion

This section discusses lessons learned from designing, developing, and evaluating the physiology-driven EmLLM chatbot for daily stress management, as well as the advancements needed to enable a pervasive system for daily stress management.

4.6.1 Model Design Trade-offs in Psychophysiological Inference

Developing accurate and robust psychophysiological inference models requires navigating several design trade-offs related to model generalization, data selection, and feature engineering. Generalized models that aggregate physiological data across participants often achieve higher accuracy but raise concerns about their robustness across individuals and diverse contexts. In contrast, personalized models tailor inference to individual variations, enhancing sensitivity to user-specific physiological patterns but limiting scalability across larger populations. This tension between scalability and personalization reflects a central challenge in psychophysiological inference of stress and affect.

A second trade-off arises in the choice of data. Physiological signals such as PPG, EDA, and ST are validated correlates of stress and emotion. In contrast, contextual, non-physiological features (e.g., activity, environment) may provide complementary information but also introduce confounding. Although data-hungry models benefit from larger feature sets, incorporating signals without a clear psychological grounding can lead to overfitting and reduce interpretability. Thus, more data does not necessarily guarantee greater accuracy; instead, the quality and theoretical validity of the data are paramount.

Finally, feature engineering presents a further dilemma. Handcrafted features enable researchers to highlight potentially informative physiological markers while preserving interpretability. However, there is little consensus on the "best" features, and selection often prioritizes model accuracy over theory. End-to-end deep learning approaches bypass this step, automatically learning representations from raw signals and improving generalization, yet at the cost of transparency. These black-box models may offer strong performance but provide limited insight into the mechanisms of inference.

4.6.2 Challenges in Real-World Deployment of the Integrated System

While controlled experiments provide reliable conditions for training psychophysiological inference models, their ecological validity is limited. Models are typically trained on data collected in laboratory settings, where stress can be precisely induced and labeled. However, when deployed in naturalistic contexts, such as in our EmLLM pilot study, data collection captures the richness of real-world experiences but also introduces substantial noise and artifacts. This trade-off underscores the difficulty of ensuring both experimental rigor and real-world generalizability, highlighting the need for inference models that remain robust to variability in everyday settings.

Integrating these inferences into LLMs through biocybernetic adaptation further compounds this challenge. Prompt engineering provides one mechanism to adapt LLM outputs based on user states, but it cannot ensure consistent or context-sensitive behavior. More sophisticated strategies involve fine-tuning on multimodal datasets and adapting input architectures to process physiological signals alongside textual and contextual inputs. Achieving effective biocybernetic adaptation, therefore, requires models that can learn and correlate across heterogeneous data streams while maintaining empathetic, user-centered interactions.

Finally, real-world deployment in sensitive domains such as mental health raises critical privacy, security, and ethical concerns. Physiological signals and conversational data constitute highly intimate information, and inadequate safeguards risk breaches of confidentiality and erosion of user trust. Participants in the pilot study voiced concerns about the visibility of their conversations, reinforcing the importance of transparency, user control, and stringent data protection protocols. Embedding robust ethical guidelines and privacy-preserving mechanisms into system design is not only a regulatory necessity but also essential for fostering long-term trust and adoption of such systems.

4.7 Conclusion, Limitations, and Future Works

This chapter presented the concept, design, development, and evaluation of physiology-driven EmLLMs for stress management. By integrating psychophysiological inference with the generative and adaptive capabilities of LLMs, physiology-driven EmLLMs constitute a new class of digital mental health systems that support biocybernetic adaptation. The pilot study demonstrated both the promise and the challenges of this approach: graduate students experienced high levels of worry during their workday; the inference model achieved strong performance on benchmark datasets but only moderate accuracy in the field; and the chatbot was perceived as competent, responsive, and, at times, "therapist-like."

Despite these contributions, several limitations must be acknowledged. First, the prototype was deployed on a web-based interface, which may not align with users' everyday workflows and limits portability. For everyday use, a mobile application is necessary to provide seamless and unobtrusive access to stress monitoring and interventions. Second, the pilot study involved a small, homogenous sample of eight doctoral students. Third, the stress detection model tended to overpredict stress in real-world conditions. This is primarily due to a mismatch between the stimuli, resolution, and personalization of the data used to train the models and that used to make stress detections. Finally, the fine-tuned LLM occasionally produced unexpected outputs, reflecting limitations in dataset curation and highlighting the risks of bias and instability in LLMs.

Future work will address these limitations in several ways. A priority is the development of a mobile version of the EmLLM system that integrates seamlessly with wearable devices to support continuous, in situ stress management. Improving stress inference under real-world conditions will require hybrid models that combine physiological signals with contextual information while maintaining interpretability and avoiding overfitting. On the chatbot side, integrating long-term memory and personalization mechanisms could enable richer, more coherent support across sessions. Addressing the risks of bias and instability in LLMs will be essential for user trust and adoption. Finally, scaling up evaluation with larger and more diverse populations over longer durations will be critical for assessing both the therapeutic effectiveness and the long-term engagement potential of physiology-driven EmLLMs.

Chapter 5

EmBot: A Mobile App Integrating Wearables and LLMs for Stress Management

5.1 Introduction

The last chapter introduced Physiology-Driven Empathic Large Language Models (EmLLMs) and presented the design, development, and evaluation of a web-based prototype that integrated psychophysiological inference with LLMs for stress management. While this prototype demonstrated the system's promise of integrating wearables and LLMs, it also revealed important limitations, particularly its reliance on a web interface that constrained everyday accessibility and seamless use. For such systems to become practical and pervasive in daily life, they must be implemented as mobile applications that integrate seamlessly with wearable devices and deliver timely, personalized interventions in situ.

This chapter introduces EmBot, a mobile application that integrates wearable-based physiological sensing with LLMs for stress management. Building on the EmLLM framework, EmBot combines continuous monitoring of physiological signals with LLM-driven interventions to provide real-time empathic support in everyday contexts. Unlike standalone wearable apps that provide static feedback or chatbots that rely on manual input, EmBot integrates sensing, inference, and empathic dialogue within a single mobile platform, enabling more adaptive, engaging, and user-centered mental health support for stress management.

The development of EmBot addresses three key challenges identified in prior chapters. First, by leveraging smartphone-based wearable data acquisition pipelines, EmBot supports everyday accessibility and portability, ensuring that interventions are available when and where stress occurs. Second, by integrating psychophysiological inference models for stress recognition with LLM-driven conversational strategies, EmBot enables personalized and empathic dialogue, reducing the risk of generic or irrelevant responses. Third, by deploying the system as a mobile app, EmBot enhances sustained engagement, embedding stress management into daily routines rather than requiring planned, deliberate use.

This study differs from the one discussed in the previous chapter in several respects. To evaluate systems integrating wearables with LLMs for mental health, we conducted semi-structured interviews with 16 mental health experts, rather than conducting daily evaluations with graduate students. EmBot was used as a design probe to identify the opportunities, limitations, and risks of integrating wearables with LLMs for mental health, rather than to evaluate performance, user experience, or intervention effectiveness. This expert-in-the-loop evaluation provides a unique perspective on EmBot's development in clinical practice. Guided by these engagements, this chapter addresses the following research questions:

- **RQ1:** What are the opportunities for integrating wearables and LLMs for mental health applications?

- **RQ2:** What are the design implications for developing hybrid wearable-LLM systems for mental health applications?

5.2 System Description

EmBot (Figure 5.1) is a novel mobile system integrating wearables and LLMs for daily stress management. It continuously monitors physiological signals from a connected wearable device to detect user stress in real time and to provide interventions via an LLM chatbot. EmBot is organized around four key interaction stages: Detection, Feedback, Support, and Reflection.

1. **Detection:** EmBot continuously monitors users' physiological signals, including EDA, PPG, and ST, from a medical-grade wearable device to detect psychophysiological stress. When stress is detected, EmBot sends a push notification prompting the user to open the app and review the detection. This stage positions EmBot as a proactive companion, bridging the gap between passive monitoring and timely intervention.

2. **Feedback:** Once the app is opened, users are shown the detected stress level and asked to indicate whether they agree or disagree with the detection. This feedback ensures that users remain the final authority on how their wearable data is interpreted, reducing the risk of detection errors. It can also help refine detection models by capturing personalized stress patterns over time.

3. **Support:** After the user provides feedback, EmBot initiates an LLM-driven chat and provides the LLM with contextual input, including the detected stress level and user feedback, enabling it to launch an empathic and context-aware conversation. The LLM provides supportive dialogue, reflective prompts, and coping suggestions in natural language, moving beyond static interventions to offer dynamic and emotionally attuned support.

4. **Reflection:** EmBot also maintains a history of stress detections and past conver-

sations, which users can revisit at any time. This reflective feature enables users to review stress patterns, track coping strategies, and build self-awareness across time. By combining objective wearable sensing with subjective reflection, EmBot supports both immediate relief and long-term stress management.

5.2.1 Design Considerations

The following design considerations were incorporated into EmBot to enhance personalization, engagement, and trustworthiness in everyday stress management.

Stress Detection

EmBot employs a DL model for classifying stress as “stressed” or “not stressed.” User-specific feedback is collected to ensure that users’ momentary assessments take precedence over predictions made from wearable data. The stress detection model runs continuously, with detections made every 15-20 minutes. All “stressed” events detected by the model trigger notifications, whereas “not stressed” events are stored for reference. Daily notifications are further limited to 10 to avoid intervention fatigue.

LLM Customization

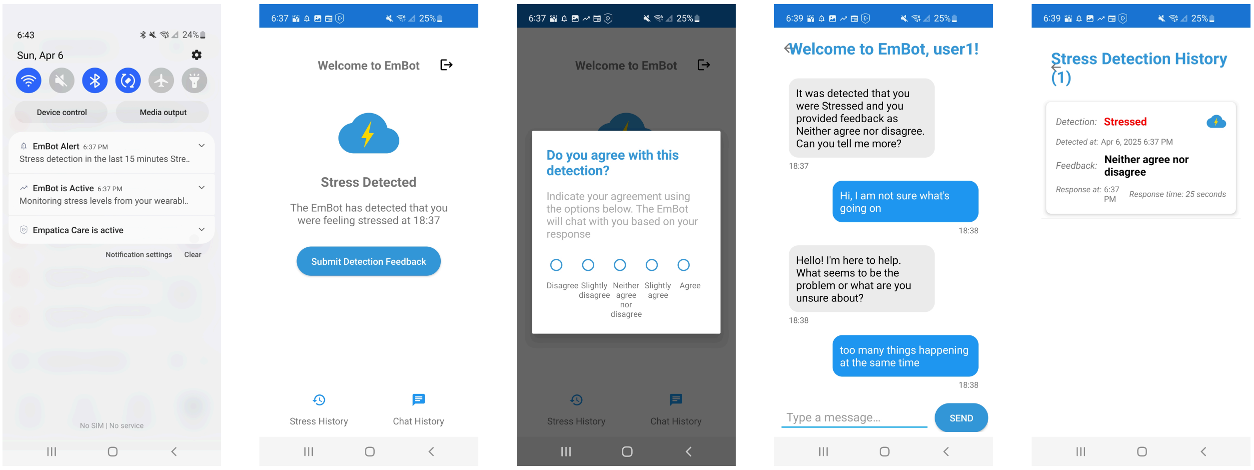
EmBot leverages prompt-engineered LLMs to deliver context-aware, reflective conversations that resemble talk therapy rather than static reminders. The LLM is provided with contextual information (detected stress event, user feedback, and history) to generate responses that feel empathic and supportive. Prompts are designed to encourage self-reflection and the use of coping strategies while maintaining the conversational tone of a supportive friend. Additionally, the LLM is instructed to identify potential stress cues in the conversation itself, supplementing wearable-based detection with conversational inference.

Privacy and Security

Considering the sensitivity of physiological and conversational data, privacy and security are foundational design considerations. Users access the app through secure registration and login, ensuring controlled access to personal records. To protect user data, EmBot prioritizes on-device computation and local storage wherever possible, minimizing exposure to external servers. This design aligns with responsible AI principles and supports user trust by giving individuals control over their own data.

5.2.2 System Implementation

The EmBot backend was developed in Python. The stress detection DL model was implemented using PyTorch, and selected personalization strategies are incorporated into the model pipeline. The LLM component was also integrated in the backend, with LLM customization implemented in Python and LangChain. Custom prompts were designed to elicit supportive, reflective conversations that adapt to user input, ensuring that responses remain contextually grounded in both stress detection and user feedback. The EmBot frontend was implemented as an Android application. Together, the backend and frontend form a hybrid wearable-LLM system that integrates wearables and LLMs within an accessible mobile application.



(a) Stage 1a, Detection: Notification when stress is detected.

(b) Stage 1b, Detection: Screen with real-time stress detection.

(c) Stage 2, Feedback: User stress detection agreement scale.

(d) Stage 3, Support: In-app chat support powered by LLM.

(e) Stage 4, Reflection: History with all stress detections.

Figure 5.1: Interaction Stages in EmBot: Detection, Feedback, Support, and Reflection.

5.3 Study Design

We conducted semi-structured interviews with 16 mental health experts, using EmBot as a design probe to elicit their perspectives on the opportunities and design implications of hybrid wearable-LLM systems. This approach was chosen because prior studies on such integrated systems have focused on graduate students, and expert interviews are well-suited to exploring emerging socio-technical systems. More importantly, this study design enabled us to address our research questions by eliciting expert reflections on opportunities (RQ1) and by using EmBot to ground the discussion in a concrete artifact that revealed design implications for future systems (RQ2).

5.3.1 Methodology

Semi-structured interviews were employed to ensure consistency across participants and to allow for flexibility in following up on unanticipated insights. Experts were first invited to share their experiences with wearables and LLMs in mental health contexts, and then asked to envision the integrated system.

To ground these reflections, EmBot was introduced as a design probe. The prototype served as a concrete artifact to anchor the discussion, enabling experts to articulate their expectations and identify design opportunities. Experts evaluated how EmBot aligned with their envisioned hybrid wearable-LLM systems and suggested improvements to the existing design.

Participation was voluntary, and each interview lasted approximately 45-60 minutes and was conducted either remotely via a secure videoconferencing platform or in person (whichever was possible). A list of questions that were asked in the semi-structured interview is provided in the supplementary material. The Institutional Review Boards approved all study procedures, and informed consent was obtained from all experts before the interviews.

5.3.2 Participants

A total of 16 mental health experts were recruited through purposive and snowball sampling, using email invitations circulated within professional networks. The sample included 8 PhD candidates and 3 postdoctoral researchers in digital mental health, 2 assistant professors and 1 distinguished professor in clinical psychology, 1 practicing psychologist, and 1 psychiatrist with an academic appointment. Their areas of expertise spanned cognitive behavior therapy, stress detection, mobile and wearable sensing, digital phenotyping, clinical intervention, machine learning, and human-centered AI design. This diversity enabled us to capture both

Table 5.1: Expert role and expertise.

ID	Role	Area of Expertise
E1	Psychologist	Cognitive behavior therapy, lifestyle changes, virtual reality
E2	PhD Candidate	Personnel selection, clinical intervention, machine learning, natural language processing
E4	PhD Candidate	Digital mental health, smartphones, machine learning
E5	Postdoctoral Researcher	Cognitive neuroscience, physiology, eye-tracking
E6	Assistant Professor, M.D.	Digital mental health, digital phenotyping, LLMs
E7	PhD Candidate	Digital health, ubiquitous computing, human-computer interaction
E8	Postdoctoral Researcher	Digital mental health, ubiquitous computing, human-centered AI
E10	Postdoctoral Researcher	Passive smartphone sensing, machine learning, mood and behavior modeling
E12	PhD Candidate	Stress detection, mobile health, machine learning, deep learning
E13	PhD Candidate	Mobile sensing, digital health, eating disorders, self-control, emotion regulation
E14	PhD Candidate	Ecological momentary assessment, wearable sensing, eating disorders
E15	Distinguished Professor, Psychology	Clinical child and adolescent psychology, developmental psychopathology, cognitive behavior therapy, social cognitive learning theory
E16	PhD Candidate	Healthy aging, daily experiences, multimorbidity
E17	Assistant Professor, Psychology	Emotion regulation, executive function, adolescent psychology
E18	Assistant Professor, Psychology	Ecological momentary assessment, intimate partner violence, substance use, dating violence, mindfulness

emerging and established perspectives on the integration of wearables and LLMs for mental health. Data from one expert (E11) was partially included, and data from two experts (E3 and E9) were excluded entirely from the analysis due to technical recording issues. Table 5.1 provides details on the experts' roles and expertise.

5.3.3 Analysis

We adopted a hybrid approach combining inductive and deductive coding to identify emergent codes, after which the two sets were consolidated into final themes. First, two researchers independently conducted inductive coding of the interview transcripts to identify preliminary codes and surface emerging patterns. Next, we used LLM-Assisted Thematic Analysis (LATA) [239], a newly validated method for inductive coding that leverages LLMs to generate additional codes and themes. This step served as a complementary step to ensure breadth in the inductive codes. The final set of inductive codes was consolidated through iterative comparison between the manually derived and LLM-assisted outputs, followed by collaborative discussions to resolve discrepancies.

We then applied deductive coding using theoretical frameworks from Human-Centered AI [217], Biocybernetic Adaptation [70], and Human-AI Interaction guidelines [13]. Two researchers independently performed the deductive coding, which was then compared with the inductive codes. Through collaborative discussions, the two sets were refined and consolidated into a final thematic structure. Disagreements were resolved by consensus, and the resulting themes were validated through repeated close readings of the data to ensure both theoretical grounding and empirical robustness. The final codebook is provided in the Supplementary Materials.

5.4 Findings

In this section, we present expert perspectives on integrating wearables and LLMs for mental health applications. We organize the findings into three themes: (1) gaps in traditional practice, (2) limitations with wearables and LLMs, and (3) opportunities for hybrid wearable-

LLM systems. We then present expert feedback on the EmBot prototype.

5.4.1 Gaps in Traditional Practice

Traditional practice in mental health monitoring primarily relied on daily diaries [10], journaling [79], and Ecological Momentary Assessments (EMAs) [68]. Experts highlighted that the most common method for in-situ monitoring is EMAs. They have transitioned from paper- and telephone-based formats to smartphone- and smartwatch-based delivery, improving feasibility and ecological validity and capturing experiences more accurately than retrospective surveys [46, 194].

However, experts highlighted several shortcomings of EMAs. While valued for capturing real-time experiences, frequent EMA prompts were described as burdensome, leading to fatigue and lower compliance over time. Missed prompts were considered particularly problematic because they often occurred during moments of heightened emotional arousal. One expert explained: *"It's burdensome. And also, we might miss important information..."* (E10). Another expert highlighted: *"Sometimes clients are not really so motivated to keep a journal..."* (E1).

Self-help intervention techniques, such as mindfulness exercises and breathing techniques, were deemed valuable by experts but were considered difficult to sustain due to low motivation. Clinical approaches, such as Cognitive Behavioral Therapy (CBT) and Eye Movement Desensitization and Reprocessing (EMDR), as well as other talk therapies, were seen as effective but difficult to access by experts. Moreover, experts described long waiting lists, overburdened providers, and significant unmonitored gaps between sessions, which increase recall bias and reduce the everyday relevance of the intervention. E14 highlighted: *"It's delivered weekly for like hour-long sessions once a week...there's just a lot that happens in* *between sessions."*

5.4.2 Limitations with Wearables and LLMs

Experts viewed wearables as a promising tool to complement EMAs, as they capture objective physiological signals, such as EDA, PPG, and ST, as well as behavioral data, including sleep patterns, activity levels, and location. Wearables can mitigate recall bias and subjectivity in EMAs. Additionally, they can reduce the burden associated with frequent EMA prompts by either *"replacing it entirely"* (E10) or sending *"event-based EMAs"* (E12). Experts also highlighted the limitations of wearables and envisioned how LLMs could address these challenges. At the same time, they cautioned that LLMs introduce their own limitations.

Data quality and device accessibility:

One prominent concern with wearables was the quality and reliability of the data. Wearable data was described as noisy, prone to missing values, and inconsistent over long periods. Additionally, experts noted that wearables are expensive, and many individuals, especially older adults or those from underserved communities, may lack access to or familiarity with such devices. An expert highlighted: *"Devices are expensive for many individuals...say, senior citizens...maybe they are not habituated to using these sorts of devices"* (E12).

False positives and potential harm:

Experts were concerned about misinterpreting benign states as alarming from wearable data. In applications involving real-time inference and feedback, false feedback risks reinforcing harmful behaviors, especially among vulnerable groups. An expert shared: *"We had to* *put in a lot of safeguards to make sure that the device was not inadvertently like feeding into disordered...and sort of increasing symptoms"* (E14). An expert also noted the lack of "meaningful" (E8) downstream applications of wearables beyond basic feedback.

Experts highlighted the use of mental health chatbots powered by LLMs, in conjunction with wearables and EMAs, to monitor user states by collecting free-text responses. Chatbots were primarily regarded as valuable tools for expanding access to care, ensuring constant availability, and providing nonjudgmental support. An expert shared that LLM chatbots are now gaining *"unprecedented levels"* (E6) of user engagement. However, experts also highlight several limitations of LLM chatbots, including a lack of genuine empathy, reliability of content, long-term monitoring, and privacy.

Usability, reliability, and trust:

Experts also highlighted usability limitations of wearables, including users forgetting to charge the device, being unable to connect wearables to other devices, and experiencing ergonomic discomfort from prolonged wear. An expert noted: *"It is prey to human error...participants would forget to put on the device, or they would forget to charge it"* (E14). Moreover, experts noted that the feeling of being *"tested by this device"* (E2) erodes user trust.

Engagement and overdependence

These limitations may collectively contribute to the drop-off and relapse rates of digital therapeutics (wearables and chatbots) in general. An expert shared that *"the drop off for traditional digital therapeutics is pretty high...the Achilles heel is that people start using it for a while, and then they stop"* (E6). In parallel, some experts were concerned about users getting overdependent on digital mental health tools. An expert warned against *"overdependence on the technology"* (E1), raising concerns that users might rely excessively on automated feedback and fail to develop internal coping mechanisms.

Clinical appropriateness and burden:

Experts were concerned about the clinical appropriateness of digital mental health tools, particularly in high-risk situations. An expert expressed the concern that such tools *"can't detect dire things on time and...don't escalate or they don't respond properly"* (E4). Furthermore, experts were concerned that the integration of modern technology and sensitive data into clinical practice could create additional work for clinicians, who already struggle with their existing workloads. An expert noted: *"Clinicians have a lot on their plates...any new technology is just saying that...you need to add more data into this and then deal with all the documentation about this"* (E4).

5.4.3 Opportunities for Wearable-LLM Systems

Integrating wearables and LLMs presents unique opportunities to advance mental health support by combining the strengths and complementing the limitations of each technology. By using EmBot as a design probe in our interviews with mental health experts, we identified four key opportunities.

Improving data quality and feedback:

Wearables provide continuous physiological signals, reducing recall bias and enhancing ecological validity. LLMs, in turn, can contextualize these signals through dialogue, distinguishing meaningful changes from everyday fluctuations. This multimodal fusion enables the integrated systems to infer user state and deliver precise interventions at the right time. An expert noted: *"Opportunities to capture more ecologically valid data, different kinds of data...reduces participant recall bias and forgetting* (E14).

Balancing monitoring with reflection:

Experts highlighted that integrating wearables and LLMs can transform monitoring into empathic engagement. They noted that the EmBot's stress-triggered notification feels more empathetic than traditional apps that require user initiation. E6 shared: *"I like that the Chatbot is reaching out to you because it thinks you're stressed."* LLMs can translate physiological changes into reflective dialogue, enabling the system to respond in a way that feels more human and emotionally aware.

Personalization and engagement:

An expert noted that LLMs are already reaching "unprecedented levels" (E6) of engagement. When integrated with wearables, systems such as EmBot can personalize the tone, frequency, and style of interactions based on individual preferences, making them both proactive and responsive to user needs. Moreover, EmBot can be further personalized to multimodal experiences, such as voice-based conversations or avatar-driven interfaces. E5 noted: *"Personalization would be a key for this kind of app. You have to read this person's pattern completely, thoroughly, so that you can only prompt that at the right moment."*

Reducing user and clinician burden:

Hybrid wearable-LLM systems can reduce the burden of understanding and interacting with wearable data. For users, the system can transform wearable data into clear summaries that support reflection, goal setting, and self-help. An expert shared: *"It tells me, oh, yeah, you talked with your therapist?...do you follow that up?...or is it just a plan?"* (E1). For clinicians, LLMs can distill long-term trends into concise, actionable summaries, offering insights into stress patterns or therapy adherence without increasing documentation workload. This opportunity positions EmBot not as an additional data source but as an intermediary that provides clinically relevant, human-readable insights.

5.4.4 Improving EmBot

In addition to reflecting on the broader integration of wearables and LLMs, experts offered specific suggestions to refine EmBot's design and functionality.

Improving stress detection:

Experts recommended strengthening EmBot's stress detection by combining automated inferences with optional EMAs to capture contextual nuance. They also highlighted the importance of personalized detection models to account for individual differences in physiology and stress responses. An expert suggested: *"previous research that compared personalized models and unpersonalized models...personalized are working better"* (E10).

To prevent notification fatigue, experts advised capping alerts to a small number per day, spacing them out appropriately, and calibrating frequency during onboarding to identify a *"sweet spot"* (E6). An expert suggested building adaptive mechanisms that recognize when users disengage and modify the system's behavior accordingly: *"If notices that you're not responding to it...it should...back off, or should we come up with new ways of doing it"* (E8). Flexibility was emphasized by experts, including the ability to revisit missed notifications and customize notification frequency.

Experts further emphasized the importance of interpretability in enhancing the empathic and trustworthy nature of the hybrid wearable-LLM system. For instance, an expert said: *"Maybe the chat could say, for example, or I picked up a bit of stress, maybe your heart rate went up to 120"* (E1). Another expert suggested that users must be allowed to ask questions like *"Was it heart rate? Was it more movement?"* (E17) to EmBot to confirm why it gave a positive stress detection. Moreover, experts encouraged the integration of additional inputs, such as calendars, location data, and tags (including events and people) associated with stressful episodes, to help users recall contextual details more effectively.

Improving LLM customization:

Experts encouraged making EmBot's conversations more structured and context-aware. Rather than offering long, generic responses, a hybrid wearable-LLM system should ask reflective follow-up questions grounded in users' situations. One expert illustrated: *"We ask, did anything happen? Did you argue with someone between the last prompt? And if they say yes, we move on to the baby questions, like, who did you argue with, and what happened?"* (E14).

Several experts noticed that EmBot's LLM responded too quickly, which felt robotic, and suggested incorporating subtle delays and features, such as typing indicators, to simulate thoughtfulness. Beyond text, experts recommended supporting voice interaction with an expert stating: *"I'd probably just be using my voice to talk to the app"* (E2). Even personalized avatars, such as a favorite animal, were suggested to foster emotional connection and engagement. Personalization was again emphasized, with some experts suggesting that the LLMs' responses, tone, and behavior be tailored to align with individual user preferences. For instance, an expert shared: "definitely want my agent to be more calm, like less neurotic than me" (E2).

To strengthen EmBot's supportive role, experts recommended including a resource hub that provides evidence-based videos and links to university and national mental health services. For higher-risk scenarios, such as signs of acute distress or suicidal ideation, EmBot should include escalation protocols, such as crisis resource buttons or emergency contact. An expert warned: *"I think if there's a way to engineer to build in any kind of way to just detect high risk events so that it can divert people to where help is needed"* (E18).

Improving privacy and security:

Experts highlighted user onboarding as a critical step, emphasizing the need for explicit, user-friendly orientation materials that clearly articulate EmBot's capabilities and limitations. An expert underscored transparency as a foundational requirement, noting the importance of clarifying: *"Here's what it can do-it can handle A, B, C, D, and E types of tasks. But it cannot do A, B, C, D, and E, and that distinction should be made explicit to help manage expectations"* (E18) during onboarding.

Privacy was another concern, and experts emphasized that users must retain control over their data. An expert advised: *"If the user says something super sensitive, it might be good for them to know that this is recorded somewhere and they can delete the conversation if they want to"* (E4). Another expert suggested a privacy-preserving approach to collecting detection feedback from EmBot by creating *"a wizard on the home screen so that...people can just report it"* (E6).

Other improvements:

Additionally, experts recommend using graphs and charts to illustrate stress patterns over time and across stacked boxes (Figure 5.1e). They highlighted that these visual aids can enhance user understanding and serve as meaningful conversation starters in therapy (E10). One expert envisioned a gamified system where users could collect stress and exchange it for rewards: *"it detects my stress, and then you can collect all my stress, and then maybe I can exchange my stress."* (E2).

5.5 Discussion

Synthesizing expert insights enabled us first to outline a broader design space for wearable-LLM integration and then suggest various design implications and evaluation considerations for future hybrid wearable-LLM systems in mental health.

5.5.1 Design Space for Wearable-LLM Integration

Building on our expert interviews and prior HCI work on design spaces [30, 171], we identify six key dimensions (D1–D6) that characterize the design space for hybrid wearable-LLM systems in mental health. Together, these six dimensions can serve as a framework for designing future hybrid wearable-LLM systems.

D1. Goal of the System: Hybrid wearable-LLM systems can be designed to serve different goals, including supporting *everyday well-being* through self-help, coping, and lifestyle management, or extending to *clinical support* by reinforcing therapy, delivering evidence-based practices, or generating summaries for clinicians.

D2. Role of the Wearable: Wearables can play two roles in hybrid wearable-LLM systems. At the most basic level, they act as *monitors*, collecting continuous physiological and behavioral signals (e.g., EDA, HR, sleep) [96, 108, 126] or go further to *infer and* *predict* users' mental (affective and cognitive) states and availability to trigger EMAs or interventions, or customize the content of the interventions [19, 109, 178, 179].

D3. Role of the LLM: LLMs can complement wearables in three primary ways: as *summarizers*, translating wearable data into natural language insights; as *conversational partners*, providing empathic dialogue and reflective support grounded in wearable data; or as *orchestrators*, synchronizing monitoring, summarization, and conversation into an integrated user experience [39, 75].

D4. Triggers and Adaptation: The timing and initiation of LLM interactions vary across systems. Interactions may be *user-initiated*, such as journaling or on-demand check-ins; *event-based*, triggered by detected stress episodes [179]; or *hybrid*, combining both approaches to balance proactivity with user agency [177]. Beyond timing, wearable data can be used to adapt LLM behavior by shaping dialogue tone, tailoring intervention content, adjusting pacing, and personalizing over time.

D5. Temporal Focus: Hybrid wearable-LLM systems can differ in their temporal scope. Some target *momentary regulation*, offering just-in-time interventions during stressful episodes. Some emphasize *short-term engagement*, such as daily summaries or weekly goal-setting. Others support *long-term engagement*, enabling users to track trends and inform treatment decisions.

D6. Intended Users: Hybrid wearable-LLM systems can be designed to cater to different stakeholders. Some are designed for *individual users* to self-manage their mental health. Others are designed for *clinicians*, providing distilled summaries to reduce workload and support therapy. Hybrid approaches may support users from both groups, offering self-help features for individuals and clinical insights for professionals.

These six dimensions characterize a holistic design space for wearable-LLM integration in mental health. EmBot explored one region of this space, focusing on self-help goals, event-based triggering, real-time stress inference, conversational support, and short-term reflection. Future work can explore other regions, such as systems that prioritize long-term reflection or emphasize clinician-facing summaries.

5.5.2 Design Implications for Wearable-LLM Systems

From our findings and the six dimensions of the design space, we distill several design implications for future systems that integrate wearables and LLMs for mental health. These implications extend beyond EmBot to inform the broader development of hybrid wearable-LLM systems.

Balance automation with user agency: Wearables offer the promise of continuous, automated stress detection, but such inferences are fallible and context-dependent. LLMs can translate these inferences into conversational support and reflective dialogue, but without user oversight, they risk misinterpretation or overreach. Systems should avoid over-reliance on automation and instead maintain mechanisms for user input and correction. This balance helps ensure accuracy while preserving users' sense of control and agency.

Ensure clinical needs and appropriateness: Systems that straddle clinical and non-clinical domains must be carefully designed to align with clinical needs and practices. For clinicians, hybrid wearable-LLM systems should distill insights into concise, actionable formats rather than complicate professional decision-making. For individual users, conversational content should be grounded in validated therapeutic frameworks where appropriate, while remaining accessible and non-clinical when intended for self-help purposes. Importantly, such systems should incorporate escalation pathways to ensure safety in high-risk scenarios.

Balance clinical effectiveness with user experience: Evidence-based interventions only work if users are willing to engage with them. Hybrid wearable-LLM systems should therefore present validated therapeutic practices in interactive, personalized, and engaging formats, ensuring that users not only receive adequate support but also remain motivated to use them. This principle highlights the importance of combining rigorous clinical grounding with design strategies that sustain everyday usage.

Design empathic and engaging conversational support: The quality of conversational support depends on both the accuracy of inferences and the perceived empathy and relevance of system responses. Designers should craft interactions that are contextually aware, emotionally resonant, and sensitive to users' mental states. Equally important is designing for sustained engagement without fostering overdependence. Conversations should feel supportive and human-like without making the user overly dependent or replacing professional care.

Embed personalization and adaptivity across layers: Personalization should extend across both the sensing and conversational layers of hybrid wearable-LLM systems. At the sensing layer, this involves adapting thresholds and models to users' physiological and behavioral baselines. At the conversational layer, it means aligning interaction style, tone, and modality to individual preferences. Beyond static customization, systems should adapt dynamically over time, learning from ongoing interactions and evolving in tandem with the user.

Prioritize transparency, privacy, and user control: Trust in hybrid wearable-LLM systems hinges on responsible design choices regarding transparency and data protection. Users should clearly understand what the system can and cannot do, how inferences are made, and what data are being collected or stored. Explaining why a system made a particular inference or suggestion is especially important, especially in contexts where false positives are possible [168]. The system should provide users with granular control over their data and ensure that sensitive information is protected without compromising system functionality or performance.

These implications underscore the need to design hybrid wearable-LLM systems not only as monitoring and intervention tools, but as adaptive, transparent, and clinically appropriate digital mental health tools. They also demonstrate how the identified design space can guide future research and development across different goals, roles, triggers, adaptations, time frames, and user groups.

5.5.3 Evaluating the Hybrid Wearable-LLM System

To advance the design of hybrid wearable-LLM systems for mental health, it is equally important to establish how such systems should be evaluated. Based on our findings, we highlight three complementary dimensions for evaluation: inference, user experience, and clinical support.

Evaluating inference: Evaluation should go beyond accuracy alone to account for the challenges of real-world deployment. Key criteria include reliability across diverse populations, robustness to missing or corrupted signals, and the ability to adapt to individual variability in physiological baselines. Systems must also be tested in real-world contexts to ensure that models trained in controlled environments can perform effectively when deployed at scale [53].

Evaluating user experience: Even clinically sound and technically reliable systems will fail if users do not engage with them. Evaluation must therefore consider usability, engagement, and sustained adherence [207]. This includes assessing users' perceptions of the system's responsiveness and emotional resonance, as well as whether personalization and adaptivity lead to higher engagement over time. Longitudinal studies are particularly important for understanding whether systems can maintain everyday usage rather than being abandoned after initial use.

Evaluating clinical support: The most critical evaluation of the hybrid wearable-LLM system is whether it improves users' mental health outcomes. This includes reducing stress or related symptoms, fostering healthier coping strategies, and promoting long-term health and well-being. At the same time, systems should be evaluated for their capacity to support clinicians by providing actionable insights that help monitor and guide patients without increasing clinical workload.

These dimensions indicate that the evaluation of hybrid wearable-LLM systems must encompass technical, experiential, and clinical criteria. A system that excels in only one area, for example, highly accurate inference but poor engagement, cannot be considered successful in supporting mental health.

5.6 Limitations and Future Work

As discussed, wearables and LLMs have several limitations, and their integration could exacerbate these challenges. If stress detection is inaccurate, frequent or unnecessary notifications may burden users, especially during periods of heightened stress. Furthermore, interactions that rely on generic, templated, or overly automated responses risk reducing empathy and trust. Without careful design, such pitfalls can erode user confidence and lead to long-term disengagement.

Our findings should also be considered in light of several methodological limitations. The study relied on 16 experts from psychology, psychiatry, digital health, and behavioral science but excluded end users, caregivers, and frontline workers whose perspectives are critical to understanding adoption. Feedback was based on a prototype walkthrough in a simulated context rather than on real-world use, which limited insight into sustained engagement, trust, and unintended consequences. We also did not explicitly examine demographic or cultural differences, which shape perspectives across healthcare systems and sociocultural contexts.

Future work should therefore recruit a more diverse set of participants and investigate how hybrid wearable-LLM systems are experienced by end users, particularly underserved groups, with a focus on usability, accessibility, ethics, and clinical relevance. In addition, systems should advance toward richer multimodal inference, combining physiological signals with behavioral, contextual, and environmental data (e.g., activity, voice, location, or interaction patterns) to improve accuracy and robustness. In parallel, multimodal generation, such as integrating text, voice, avatars, and visualizations, could provide more engaging and adaptive feedback for users, tailoring support to individual preferences and contexts [60].

Beyond stress management, the potential applications of hybrid wearable-LLM systems extend into multiple domains. In education and training, adaptive support can be provided by monitoring cognitive load, attention, and stress to optimize learning experiences. In the arts, they could foster creativity by transforming physiological signals into generative, affect-driven music, visualizations, or interactive performances. In construction and industrial safety, they can detect fatigue, stress, or distraction in real-time and deliver timely interventions to prevent accidents.

5.7 Conclusion

Our paper introduces EmBot, a mobile design probe that couples wearable stress inference with LLM-mediated conversation to explore how integrated hybrid wearable-LLM systems might better support everyday mental health than respective standalone systems. Grounded in interviews with 16 mental health experts, we synthesize a six-dimensional design space, articulate concrete design implications, and propose an evaluation rubric that spans inference robustness, user experience, and clinical utility. Together, these contributions highlight how hybrid wearable-LLM systems can be designed, developed, and evaluated not only as technical artifacts but also as adaptive, human-centered, and clinically appropriate digital mental health tools that overcome the limitations of existing tools.

Chapter 6

Evaluating the Affective Performance and Stability of Multimodal LLMs

6.1 Introduction

So far, this dissertation has discussed the use of physiological data with LLMs for daily stress management. This chapter explores the complementary roles of audio and visual modalities in affect recognition and in providing general mental health support. LLMs have advanced rapidly beyond text-only processing and now support multimodal reasoning across images, audio, and video. Omni-LLMs extend these capabilities further by enabling not only multimodal understanding but also multimodal generation, making them powerful tools for emotionally rich, interactive communication. The ability of such models to integrate linguistic cues with vocal prosody, facial expressions, and body movements makes them especially promising for affective tasks [47, 95]. However, slight variations in modality or prompt phrasing can cause LLMs to misinterpret emotional cues, hallucinate nonexistent behaviors, or generate unstable content. As these models are increasingly used to mediate emotionally sensitive interactions, understanding their reliability in multimodal affective performance is essential.

In this chapter, the reliability of three state-of-the-art multimodal LLMs (Qwen2.5-o, Phi4-m, and MiniCPM-o-2.6) is examined in the context of affective interactions using the Predictability Computability Stability (PCS) framework [253]. We focus on three core affective tasks that underpin emotionally supportive AI systems: (1) emotion label prediction, (2) explanation of the predicted emotion, and (3) generation of empathic text messages. The PCS framework quantifies not only predictive performance but also the predictive and generative stability of multimodal LLMs under perturbations. Using a fully crossed perturbation design, we systematically vary input modalities and prompt templates to test each model's sensitivity to changes in sensory inputs and prompt phrasing. The findings reveal critical reliability gaps in current multimodal LLMs and highlight limitations to deploying them in real-world settings, especially in sensitive applications such as mental health.

6.2 Background

Conversational AI systems, based on LLMs, are increasingly used for mental health applications in both clinical and non-clinical settings. In a clinical setting, these systems can deliver evidence-based therapy, such as cognitive-behavioral therapy exercises, through natural dialogue [104]. In non-clinical settings, LLMs are increasingly used to share feelings and receive emotional support [49]. However, a key concern with such systems is their reliability in understanding nuanced user inputs and generating safe, trustworthy outputs. For instance, Iftikhar et al. [110] found a lack of contextual understanding in LLM chatbots, and Chandra et al. [31] reported several risks, including LLM chatbots providing harmful responses during emotionally charged conversations. Therefore, it is essential to evaluate both the inferential and generative capabilities of LLMs, especially in sensitive applications such as mental health.

Existing research has evaluated and improved LLMs' capabilities for emotion classification across text, audio, and visual data, using a range of prompting techniques, fine-tuning paradigms, and training methodologies. Using text-based zero-shot and few-shot affect recognition, Wake et al. [238] evaluated ChatGPT and reported reasonable performance, but found a bias toward frequent emotional categories, such as joy and sadness. Feng et al. [77] compared text-based zero-shot and few-shot prompting in GPT-4, GPT-3.5, Alpaca, and LLaMA, and found that GPT-4 gave significantly better zero-shot results. Stricker and Paroubek [224] fine-tuned a LLaMA-2 7B model on text-based emotion recognition datasets, achieving performance comparable to that of fully fine-tuned smaller models.

Recent work has also explored the use of multimodal data to improve emotion recognition performance. To enable text-based LLMs to process multimodal data without any architectural modifications, several studies have proposed converting non-textual modalities into textual descriptions. For instance, Zhang et al. [255] introduced Lantern, a two-stage framework that predicts preliminary emotion scores using a traditional model and subsequently sends the scores to a text-based LLM. Dongre et al. [58] followed a similar approach to integrate wearable data with LLMs for stress detection. Wu et al. [244] translated acoustic features into natural language representations to support affect recognition from speech signals.

Recent work has also evaluated natively multimodal LLMs for emotion recognition in images and videos. Nadeem et al. [173] evaluated the zero-shot performance of VLMs such as LLaVA and GPT-4 against specialized emotion recognition models, finding that LLaVA achieved higher accuracy. Nelson et al. [176] evaluated the facial emotion recognition performance of three general-purpose LLMs, including GPT-4o, Gemini 2.0 Experimental, and Claude 3.5 Sonnet, and reported relatively high misclassifications of fear. Zhang et al. [256] proposed the Set-of-Vision (SoV) prompting method, which overlays facial bounding boxes on images before prompting VLMs to classify emotions to improve emotion recognition in images with multiple people and objects.

The advantage of LLMs over traditional emotion recognition models is that they can also generate reasoning and explanations for an emotion classification. Reasoning and explainability not only increase the system's trustworthiness and user trust, but can also improve model performance [51]. To enable multimodal emotional understanding and reasoning, Yang et al. developed EmoLLM by integrating an audio- and visual-encoder with a foundational language model and fine-tuning it on the EmoBench dataset [248]. Emotion-LLaMA is a LLaMA-based multimodal model with audio, visual, and text encoders instruction-tuned for emotion recognition and reasoning [36]. If emotion classification and reasoning are coupled with empathic message generation, they can create new opportunities for emotionally aware and adaptive applications. Moreover, multimodality can provide contextual details to help LLMs better understand emotions and respond more empathetically.

While multimodality can improve emotion recognition, reasoning, and empathic message generation, it also introduces more risks and concerns. Models can exhibit brittle behavior when input modalities vary, and the user's prompt can lead to wildly different outputs. Recent work by Liu et al. [151] highlights how LLMs can display inconsistent generations that lead to logical contradictions or flip-flopping on sensitive topics. When working with multimodal data, a robust LLM should extract partial understanding from a noisy image or fall back to just text if other modalities fail, rather than misinterpreting the input and generating unstable outputs. It should also be insensitive to prompt perturbations, meaning it should give semantically similar answers to similar prompts [200]. Therefore, when evaluating multimodal LLMs, it is critical to test not only their peak performance but also their robustness.

PCS is a unified conceptual and technical framework that applies principles of predictability, computability, and stability to guide the reliable and trustworthy practice of Veridical Data Science (VDS) [253]. *Predictability* emphasizes grounding modeling choices in empirical predictive performance; *computability* ensures that procedures are computationally feasible, scalable, and reproducible; and *stability* evaluates how sensitive conclusions are to reasonable perturbations in data preprocessing, modeling choices, and assumptions. While PCS shares similarities with robustness stress-testing techniques [89], it differs in that it offers a principled approach to auditing and interpreting model reliability throughout the data science life cycle. The PCS framework has been successfully applied to various applications, including genomics, neuroscience, and predictive modeling pipelines [253]. The PCS framework has also been extended to generative AI systems, emphasizing that predictability should align with human-centric quality measures; computability must account for the model's deployment feasibility; and stability should be assessed by measuring how sensitive outputs are to prompt phrasing, modality changes, or other task-relevant perturbations [203].

6.3 Methodology

We apply the PCS framework to evaluate the behavior and reliability of three open source multimodal LLMs, Qwen2.5-o [246], Phi4-m [3], and MiniCPM-o-2.6 [254], across three affective tasks. Qwen2.5-o is a 7B-parameter model that understands audio, images, and video and generates both text and audio. It is trained on a large, diverse set of web-scale multilingual corpora and multimodal alignment data. Phi4-m is a 6B-parameter, instruction-tuned model that takes audio, image, and video frames as input and generates text. It is trained on high-quality, curated datasets that emphasize reasoning, safety, and grounded multimodal understanding. MiniCPM-o-2.6 is an 8B-parameter multimodal LLM that supports audio, visual, and textual inputs and generates text and audio.

We evaluated the models on three complementary multimodal emotion datasets: Interactive Emotional Dyadic Motion Capture (IEMOCAP) [27], the Multimodal EmotionLines Dataset

(MELD) [197], and Multimodal Affective Faces in the Wild (MAFW) [152]. IEMOCAP provides approximately 12 hours of high-quality dyadic interactions performed by trained actors, recorded with synchronized audio, video, and motion-capture modalities, and annotated with fine-grained categorical emotions. Speaker diarization, utterance segmentation, and shot-level video extraction were applied to isolate individual utterances in the IEMOCAP dataset. MELD extends this setting to multi-party, naturalistic conversations from the TV sitcom *Friends*, offering roughly 13,000 utterances with aligned audiovisual and transcript modalities. MAFW contains over 10,000 audiovisual clips collected from movies, dramas, and short online videos, capturing spontaneous emotional expressions across diverse speakers with both categorical and continuous labels. We converted video clips to audio-only segments, and transcripts were generated (when unavailable) to maintain a consistent text modality across datasets. Due to differences in annotation schemes, all emotion labels were mapped to the Ekman six basic emotions (anger, disgust, happiness, fear, sadness, surprise) [64] plus a neutral category.

The three models are evaluated on randomly sampled data from each dataset, with systematic perturbations applied across both input modalities and system prompts, to assess their stability in generating emotionally grounded outputs (emotion recognition, reasoning, and empathetic messages). This section formalizes our experimental design, evaluation metrics, and alignment with PCS principles.

6.3.1 Affective Tasks and Perturbation Design

Each model was evaluated in a zero-shot setting on three tasks per utterance: (1) emotion label classification, where the model classified the emotion as either anger, disgust, fear, happiness, sadness, surprise, or neutral based on the input; (2) reasoning, which required the model to generate a brief rationale for the predicted emotion; and (3) empathic message generation, where the model produced a supportive and empathetic response tailored to the speaker's emotional state.

To formalize within PCS, we define $T(D, \lambda)$ as the model-generated tuple (emotion label, reasoning, empathy message) under perturbation design choice λ (modality input and system prompt phrasing) and dataset D. We perturb λ along two axes: input modality $\mathcal{M} \in \{\texttt{A}, \texttt{T\_A}, \texttt{A\_V}, \texttt{T\_A\_V}\}$ and prompt template $\mathcal{P} \in \{\texttt{p1}, \texttt{p2}, \texttt{p3}\}$. This yields $K = 12$ (m, p) combinations per utterance, enabling PCS-consistent perturbation analysis:

$$\Lambda = \mathcal{M} \times \mathcal{P}, \quad \{T(D, \lambda_k)\}$$

6.3.2 PCS Evaluation Strategy

Predictability

For task 1, model predictions $\hat{y}_i^{(m,p)}$ are evaluated against true emotion labels y_i^{true} to compute accuracy. Cochran's Q test is used to assess the significance of differences in accuracy across modalities and prompts at the utterance level. For tasks 2 and 3, model outputs are rated by three LLM judges (GPT-4.1, Gemini-2-Flash, LLaMA-3.1) using a 1–5 Likert scale across five criteria.

For reasoning, LLM judges were guided by five evaluation criteria: *Modality-Grounded Evidence*, *Emotion–Cue Mapping*, *Specificity*, *Internal Consistency*, and *Completeness*. For empathy, the evaluation criteria included *Emotion Acknowledgment*, *Validation*, *Supportive Intent*, *Emotion-Appropriate Response*, and *Gentle Guidance*. The final quality score for each response was computed as the average of its five criterion scores.

To quantify agreement across LLM judges, we computed inter-rater reliability using Krippendorff's α and the intraclass correlation coefficient (ICC). Statistical significance of differences in ratings across prompts and modalities was assessed using the Friedman test.

Computability

In line with the PCS framework, we documented the computational feasibility of each step involved in data preparation and model execution. Large-scale multimodal data, including audio, video, and text transcripts, were stored and organized at the utterance level on a secure drive to facilitate efficient access and management. To prepare the data for model input, we applied speaker diarization and utterance-level segmentation to ensure clean and temporally aligned samples across modalities.

The multimodal LLMs ran on Google Colab Pro environments, leveraging A100 GPUs to accelerate inference across all combinations of tasks, modalities, and prompt templates. For evaluating reasoning and empathic message outputs (tasks 2 and 3), we used API calls to three distinct LLMs as judges, enabling large-scale, scalable annotation. We systematically recorded runtime performance, API token costs, and hardware configurations throughout to facilitate full reproducibility of the experimental pipeline.

Stability

We define the stability metric $s(T : D, \Lambda)$ as the degree to which outputs remain consistent under perturbations $\lambda \in \Lambda$. For the emotion recognition task, each utterance was evaluated across all combinations of modality and prompt. A binary correctness matrix was constructed, with each entry indicating whether the model's predicted emotion label matched the ground truth. We then computed a stability score S_i for each utterance as:

$$S_i = \frac{2}{K(K-1)} \sum_{j<k} \mathbb{1}[C_{ij} = C_{ik}]$$

Perturbation intervals are reported as the 10th- and 90th-percentile accuracy values across the 12 conditions.

For the reasoning and empathic message generation tasks, we computed semantic consistency across perturbations using sentence embeddings. For each utterance, cosine similarity was computed for all pairs of generated responses across modalities and prompts, yielding a similarity matrix. Therefore, the stability score S_i for each utterance is:

$$S_i^{\text{sim}} = \frac{2}{K(K-1)} \sum_{j<k} \cos\left(T_{ij}, T_{ik}\right)$$

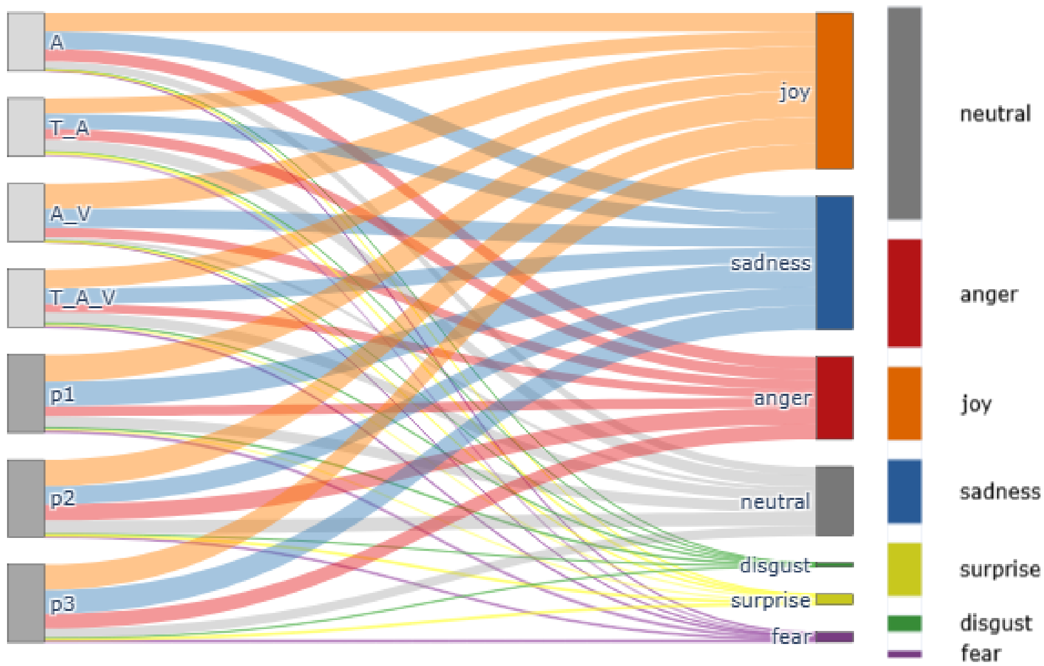
Perturbation intervals are defined using the 10th and 90th percentiles of these pairwise similarities.

6.4 Results

6.4.1 Predictability

Emotion Recognition Accuracy

Across all modality–prompt combinations, emotion recognition accuracy remained modest (Table 6.1). Qwen2.5-o achieved the highest accuracy in most conditions (best: 0.4502 under T_A_V–p2), whereas Phi4-m performed competitively with the visual modality and prompt p1, MiniCPM-o-2.6 consistently produced the lowest accuracies. T_A and T_A_V yielded



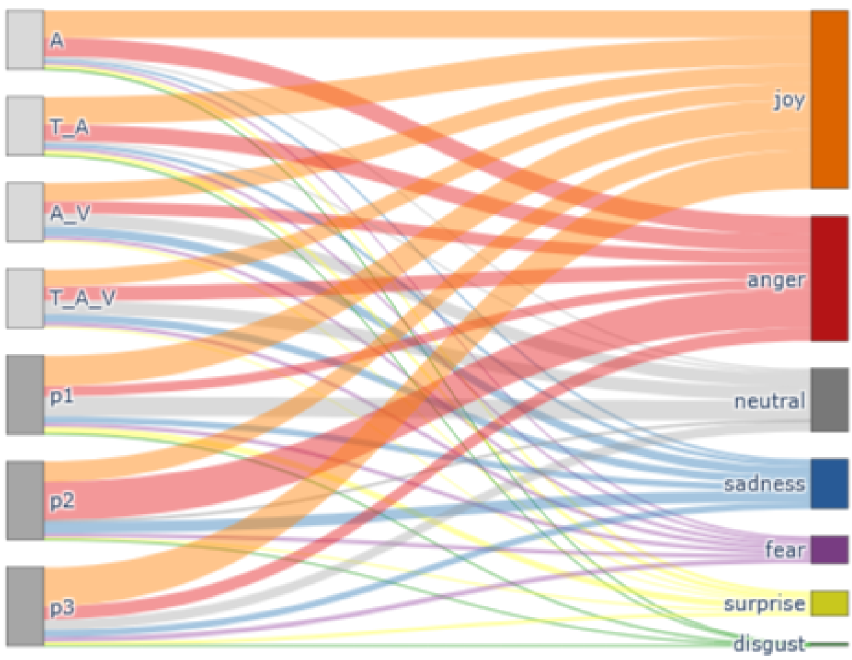
(a) Qwen2.5-o (b) GT

Figure 6.1: Model Performance and Ground Truth (GT) for Qwen2.5-o

the most reliable predictions across models, suggesting the importance of linguistic cues in multimodal LLMs.

The Sankey diagrams in Figures 6.1a, 6.2a, and 6.3a illustrate how each model distributes its predicted emotions across modalities and prompts, compared against the ground-truth (GT). Qwen2.5-o demonstrates the closest qualitative alignment to GT. In contrast, Phi4-m exhibits a more substantial reliance on visual information, producing heavier flows toward *anger*. MiniCPM-o-2.6 shows the most significant deviations from GT, with dominant but misaligned flows toward *anger*, suggesting a tendency to over-predict high-arousal negative emotions across modalities and prompts.

Cochran's Q tests revealed significant variability in emotion recognition accuracy across both modality and prompt perturbations (Table 6.2). Qwen2.5-o and MiniCPM-o-2.6 showed significant modality effects under all prompts, and Phi4-m showed strong effects primarily



(a) Phi4-m

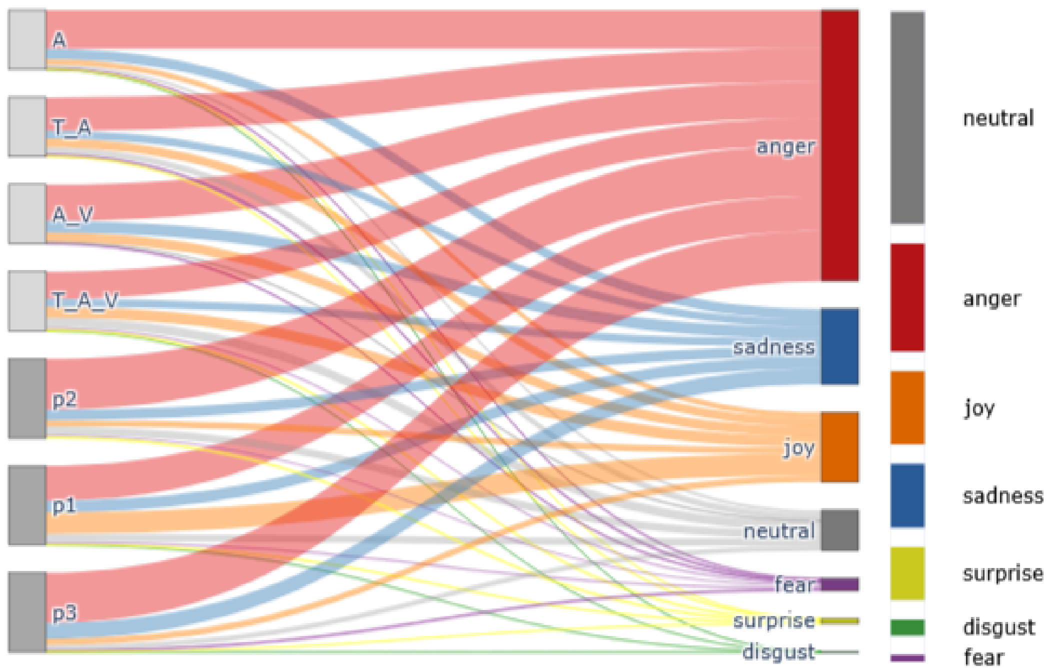
(b) GT

Figure 6.2: Model Performance and Ground Truth (GT) for Phi4-m

under p1. Prompt sensitivity was also evident for Qwen2.5-o (A and T_A_V) and Phi4-m (A_V and T_A_V). These results indicate that emotion recognition accuracy is highly sensitive to modality perturbations.

Reasoning and Empathic Message Quality

As described previously, we used three independent judge LLMs to rate the quality of the reasoning and empathetic messages generated by the multimodal LLMs relative to the emotion they predicted. To assess the consistency of these judgments, we computed Inter-Rater Reliability (IRR), examining both single-rater agreement (IRR[3]) and the reliability of the averaged score across the three raters (IRR[3,k]). Overall, the judges demonstrated moderate consistency for both reasoning and empathic messages, with IRR[3] values ranging from approximately 0.49 to 0.63 across models. When aggregated across raters, reliability increased



(a) MiniCPM-o-2.6 (b) GT

Figure 6.3: Model Performance and Ground Truth (GT) for MiniCPM-o-2.6

substantially, with IRR[3,k] values ranging from roughly 0.74 to 0.83. Phi4-m showed the strongest reliability for reasoning, while MiniCPM-o-2.6 exhibited the highest reliability for empathic message ratings.

Across modality–prompt combinations, we observe substantial variation in the quality of reasoning (Table 6.3). Phi4-m achieved the strongest overall performance, producing the highest reasoning scores in 9 of the 12 conditions, particularly under audio-only and text-augmented settings. Qwen2.5-o demonstrated competitive performance, especially when both audio and transcript information were available, whereas reasoning for MiniCPM-o-2.6 was consistently rated lower-quality, with notable drops in the visual modality. Overall, these results indicate that reasoning quality depends strongly on both the input modality and the specific prompt used.

Statistical analysis using Friedman tests revealed that the differences in reasoning quality

Table 6.1: Accuracy comparison across models for each modality–prompt combination. Bold indicates the highest accuracy in each row (best model). Italics indicate the best metric for each model across conditions.

Modality	Prompt	Qwen2.5-o	Phi4-m	MiniCPM-o-2.6
A	p1	**0.3402**	0.2955	0.2234
A_V	p1	0.3127	***0.3952***	0.1581
T_A	p1	0.4364	0.3196	***0.3265***
T_A_V	p1	0.3643	***0.4261***	0.2990
A	p2	0.4124	0.2612	**0.2268**
A_V	p2	**0.3505**	0.2680	0.1753
T_A	p2	**0.4330**	0.2887	0.2715
T_A_V	p2	***0.4502***	0.2887	0.2921
A	p3	**0.3471**	0.2749	0.2302
A_V	p3	**0.3368**	0.2852	0.1615
T_A	p3	**0.3780**	0.2749	0.2371
T_A_V	p3	**0.4467**	0.2784	0.2680

across prompts and modalities were highly significant for all three models. Across prompts within each modality, all models showed strong effects (all $p < .01$), indicating that even minor changes in prompt phrasing produced explanations that differed reliably. Similarly, modality-level comparisons showed highly significant differences across all prompts, with p-values often below 10^{-20}, demonstrating that the modality choice substantially alters the structure and coherence of the generated reasoning.

Across all modality–prompt conditions, the quality of empathic messages remained consistently high, with scores generally falling between 4.2 and 4.8 on a 5-point scale (Table 6.4). Phi4-m produced the strongest empathic responses overall, achieving the highest score in nearly every condition and showing particular strength in audio-only and text-augmented settings. Qwen2.5-o exhibited competitive performance, occasionally outperforming Phi4-m in T_A_V settings. MiniCPM-o-2.6 produced the lowest empathic scores, with noticeable drops in visual modality.

Table 6.2: Cochran's Q test p-values for accuracy differences across prompts and modalities. Significant effects ($p < .05$) are in bold with significance levels (* $p < .05$, ** $p < .01$, *** $p < .001$).

Test	Qwen2.5-o	Phi4-m	MiniCPM-o-2.6
Modality-level (within prompt across modalities)			
p1	**0.00027*** **	**0.00021*** **	**9.40e-09*** **
p2	**0.00324** **	0.60554	**1.40e-05*** **
p3	**0.00024*** **	0.98301	**0.00048*** **
Prompt-level (within modality across prompts)			
A	**0.00691** **	0.40897	0.94692
T_A	0.07612	0.14957	**0.00389** **
A_V	0.15567	**4.97e-05*** **	0.75857
T_A_V	**0.00045*** **	**2.49e-06*** **	0.36788

Friedman tests revealed that empathic message quality was susceptible to both prompt phrasing and input modality. For every model, empathic scores differed significantly across prompts within each modality, with p-values far below .001 in nearly all cases, indicating that even small changes in prompt wording produced reliably different empathic responses. Similarly, comparisons across modalities within each prompt showed powerful effects, again with p-values often many orders of magnitude below conventional significance thresholds. The only exception occurred for Phi4-m under prompt p3, where modality differences were not statistically significant, suggesting a more stable empathic response pattern in that specific condition.

6.4.2 Computability

All multimodal data, including audio, video, and transcripts, were stored on Google Drive and organized by utterance for modular, shared access. Preprocessing included speaker diarization, utterance segmentation, transcription, and audio extraction from video. Model

Table 6.3: Reasoning quality (LLM-as-judge scores) across models for each modality–prompt combination. Bold indicates the highest accuracy in each row (best model). Italics indicate the best metric for each model across conditions.

Modality	Prompt	Qwen2.5-o	Phi4-m	MiniCPM-o-2.6
A	p1	3.2016	**3.8534**	3.4994
A	p2	3.6896	***4.3462***	3.2623
A	p3	**3.3235**	3.8041	3.3438
A_V	p1	**3.8407**	3.7935	3.4397
A_V	p2	3.2593	**3.9496**	3.0865
A_V	p3	3.1569	**3.3666**	2.8093
T_A	p1	**3.5519**	3.4433	*3.6655*
T_A	p2	3.4255	**4.1672**	3.1592
T_A	p3	3.3977	**3.1773**	2.6586
T_A_V	p1	***3.9586***	3.3218	2.9588
T_A_V	p2	3.9485	**3.9408**	3.4971
T_A_V	p3	3.3989	**3.7043**	2.7898

inference was run in a zero-shot setting using Google Colab Pro with A100 GPUs. As discussed, each of the three multimodal LLMs (Qwen2.5-o, Phi4-m, and MiniCPM-o-2.6) was evaluated across 12 perturbation conditions for three tasks, with each model requiring 3 hours and 65–70 compute units per dataset. LLM-as-judge evaluations (for reasoning and empathy) were conducted via API (GPT-4.1, Gemini-2-Flash) and Google Colab inference (LLaMA-3.1), with standardized prompts and fixed seeds to ensure consistency. All code, prompts, and evaluation scripts will be released via GitHub to support reproducibility.

6.4.3 Stability

Emotion Recognition Stability

We quantify correctness-based stability across modalities and prompts. For each utterance, we compute whether a prediction is correct (1) or incorrect (0) under each perturbation con-

Table 6.4: Empathic message quality (LLM-as-judge scores) across models for each modality–prompt combination. Bold indicates the highest accuracy in each row (best model). Italics indicate the best metric for each model across conditions.

Modality	Prompt	Qwen2.5-o	Phi4-m	MiniCPM-o-2.6
A	p1	4.4018	**4.5876**	4.3608
A	p2	4.6151	***4.8310***	4.3081
A	p3	4.4645	**4.6460**	4.2195
A_V	p1	4.5864	**4.6494**	4.3023
A_V	p2	4.4387	**4.5865**	3.5258
A_V	p3	4.4330	**4.6449**	3.7159
T_A	p1	4.6160	**4.8310**	4.4376
T_A	p2	4.5418	**4.6885**	4.1592
T_A	p3	4.3636	**3.7862**	3.2835
T_A_V	p1	**4.5391**	3.7766	3.7446
T_A_V	p2	***4.6678***	4.8304	*4.3677*
T_A_V	p3	**4.4828**	4.6075	3.6964

dition and then measure how consistently those correctness outcomes agree across prompts within a modality and across modalities within a prompt (Table 6.5). We complement this with perturbation intervals (10–90% ranges) over correctness, which capture how tightly performance is distributed across configurations (Table 6.6).

As shown in table 6.5, MiniCPM-o-2.6 exhibits the highest overall correctness stability across most conditions, with the strongest stability for A and T_A_V, and all prompts. Qwen2.5-o shows strong stability in visually informed conditions, with its best performance with A_V, suggesting that when it is correct, it tends to remain consistently correct (or incorrect) across prompt variations in audiovisual settings. Phi4-m's correctness stability peaks for T_A, indicating that it is most reliable when transcript-augmented audio is available.

The perturbation intervals in Table 6.6 provide a complementary view of robustness by describing how correctness varies across perturbations. MiniCPM-o-2.6 generally achieves the narrowest intervals at the modality level, implying that its accuracy is tightly clustered

Table 6.5: Accuracy stability comparison across models. For each row, the highest value (best model) is in bold. For each column, the highest value (best metric for that model) is italicized.

Perturbation	Qwen2.5-o	Phi4-m	MiniCPM-o-2.6
Modality-level (within prompt across modalities)			
p1	0.742	0.661	**0.749**
p2	0.760	0.810	**0.814**
p3	0.777	0.750	**0.802**
Prompt-level (within modality across prompts)			
A	0.814	0.805	***0.874***
T_A	0.757	***0.840***	0.787
A_V	***0.885***	0.721	0.826
T_A_V	0.821	0.693	**0.847**

and changes little under different prompts. Qwen2.5-o shows moderate interval widths, with its best behavior again in A_V modality, whereas Phi4-m yields comparatively narrow intervals for text modality, but wider ranges when visual modality is included.

Reasoning and Empathic Message Stability

To evaluate the semantic stability of generative outputs under perturbations, we quantify stability in a semantic space for both reasoning explanations and empathic messages. For each utterance, we embed the generated reasoning or empathic text using the `all-MiniLM-L6-v2` sentence encoder and compute cosine similarities (i) across prompts within a fixed modality and (ii) across modalities within a fixed prompt. We then aggregate these similarities into mean stability scores (Tables 6.7 and 6.9) and perturbation intervals (10–90% ranges; Tables 6.8 and 6.10).

As shown in table 6.7, the semantic stability of reasoning falls in the moderate range for all multimodal LLMs, indicating that explanations change meaningfully when modalities

Table 6.6: Accuracy perturbation interval (10–90%) comparison across models. For each row, the narrowest interval (best model) is in bold. For each column, the narrowest interval (best metric for that model) is italicized.

Perturbation	Qwen2.5-o	Phi4-m	MiniCPM-o-2.6
Modality-level (within prompt across modalities)			
p1	**0.321–0.415**	0.303–0.417	0.178–0.318
p2	0.369–0.445	**0.263–0.289**	0.191–0.286
p3	0.340–0.426	***0.275–0.283***	0.182–0.259
Prompt-level (within modality across prompts)			
A	0.342–0.399	0.264–0.291	***0.224–0.230***
T_A	0.389–0.436	**0.278–0.313**	0.244–0.315
A_V	*0.318–0.348*	0.271–0.373	**0.159–0.173**
T_A_V	0.381–0.449	0.280–0.399	**0.273–0.298**

or prompts are perturbed. At the modality level, Qwen2.5-o is most stable when visual information is available and for T_A, while MiniCPM-o-2.6 is most stable when generating reasons for audio and T_A_V modality. Phi4-m shows its strongest reasoning stability when aggregating across modalities for prompt p2 and p3, suggesting that its reasoning is relatively consistent when the prompt is fixed but the modality composition varies.

The perturbation intervals in Table 6.8 provide a complementary view of the reasoning stability. Qwen2.5-o achieves the tightest modality-level intervals for all prompts, indicating that its reasoning similarity scores are less dispersed under modality perturbations. Phi4-m tends to have narrower intervals in text-rich prompt-level settings, while MiniCPM-o-2.6 shows relatively tight intervals for A_V modality but very wide ranges for T_A_V modality, suggesting occasional near-identical reasoning but also high variability across prompts.

Empathic messages generation stability exhibits a similar but slightly more stable pattern (Table 6.9). Qwen2.5-o achieves the highest stability in most within-modality comparisons and across modalities for p1 and p3, indicating that its empathic tone and framing remain relatively consistent under perturbations. Phi4-m is most stable for p2, consistent with its

Table 6.7: Reasoning stability across models. Bold indicates the highest value in each row (best model). Italics indicate the best metric for each column.

Perturbation	Qwen2.5-o	Phi4-m	MiniCPM-o-2.6
Modality-level (within prompt across modalities)			
p1	**0.5488**	0.5297	0.5160
p2	0.5206	***0.7412***	0.6419
p3	0.5194	**0.6192**	0.5850
Prompt-level (within modality across prompts)			
A	0.5958	0.5632	**0.6280**
A_V	***0.6932***	0.5202	0.5514
T_A	**0.6024**	0.5623	0.5204
T_A_V	0.6434	0.5105	**0.6645**

strong overall empathy scores, whereas MiniCPM-o-2.6 shows noticeably lower empathic stability, particularly when visual data are included.

The perturbation intervals in Table 6.10 reveal that MiniCPM-o-2.6, despite lower absolute stability, often yields the narrowest intervals in within-modality settings. This suggests that its empathic messages are consistently semantically similar to one another, even if they are less rich or human-like than those from Qwen2.5-o or Phi4-m. Phi4-m tends to produce tighter intervals at the modality level for p1–p3, indicating relatively stable empathic semantics once the prompt is fixed. Qwen2.5-o sits between, achieving higher mean stability but with moderately wide intervals, reflecting both strong and occasionally divergent empathic responses.

Table 6.8: Perturbation intervals (10–90%) for reasoning stability across models. Bold indicates the narrowest interval in each row (best model). Italics indicate the best interval (narrowest) for each model across all conditions.

Perturbation	Qwen2.5-o	Phi4-m	MiniCPM-o-2.6
Modality-level (within prompt across modalities)			
p1	**0.428–0.674**	0.374–0.681	0.384–0.635
p2	**0.388–0.667**	0.557–0.877	0.504–0.772
p3	**0.401–0.651**	0.473–0.747	0.466–0.698
Prompt-level (within modality across prompts)			
A	0.455–0.732	***0.455–0.677***	0.478–0.773
A_V	0.530–0.840	0.241–0.834	**0.440–0.648**
T_A	0.442–0.789	**0.442–0.683**	0.349–0.684
T_A_V	0.481–0.816	**0.226–0.818**	0.443–1.000

6.5 Discussion

6.5.1 Predictability vs Stability

Our analysis reveals clear distinctions in how the three multimodal LLMs behave across the three affective tasks. In terms of predictability, Qwen2.5-o achieved the highest emotion-recognition accuracy, outperforming Phi4-m and MiniCPM-o-2.6 across most modality–prompt combinations. However, stability results show that MiniCPM-o-2.6, despite lower accuracy and weaker generative quality, exhibited the most consistent behavior under perturbations, yielding the highest correctness stability for emotion recognition and the tightest perturbation intervals for reasoning and empathy. Qwen2.5-o maintained the most consistent empathic semantics, while Phi4-m was most stable when prompts remained fixed. Overall, these findings show that the most accurate models are not necessarily the most stable, and that high-quality generative responses do not guarantee accuracy and stability. Evaluating multimodal LLMs for sensitive applications such as mental health requires considering

Table 6.9: Empathic message stability across models. Bold indicates the highest value in each row (best model). Italics indicate the best metric for each model across all conditions.

Perturbation	Qwen2.5-o	Phi4-m	MiniCPM-o-2.6
Modality-level (within prompt across modalities)			
p1	**0.4919**	0.4479	0.3674
p2	0.5655	**0.5743**	0.4813
p3	**0.5113**	0.4856	0.4828
Prompt-level (within-modality across prompts)			
A	**0.5338**	0.5024	0.4737
A_V	***0.5777***	0.4693	0.3191
T_A	**0.5189**	0.4971	0.3563
T_A_V	**0.5626**	0.4553	0.4758

prediction accuracy, generative quality, and stability jointly rather than in isolation.

6.5.2 Modality and Prompt Sensitivity

Across models, both modality and prompt perturbations had substantial effects on performance. Multimodal inputs, particularly T_A and T_A_V, achieved the highest emotion-recognition accuracy and exhibited more consistent reasoning, underscoring the importance of linguistic grounding for affective inference. A and A_V configurations were more variable, especially for Phi4-m and MiniCPM-o-2.6, highlighting modality-specific weaknesses when semantic cues are limited. Prompt phrasing also introduced meaningful variation, with prompt p2 consistently improving the quality of reasoning and empathy, whereas the others produced notable instability across modalities. These findings indicate that multimodal LLMs remain highly sensitive to small perturbations in the prompt, even when the task is fixed.

Table 6.10: Perturbation intervals (10–90%) for empathic message stability across models. Bold indicates the narrowest interval in each row (best model). Italics indicate the best interval (narrowest) for each model across all conditions.

Perturbation	Qwen2.5-o	Phi4-m	MiniCPM-o-2.6
Modality-level (within prompt across modalities)			
p1	0.302–0.694	**0.300–0.613**	0.260–0.488
p2	0.406–0.751	**0.420–0.722**	0.341–0.637
p3	0.313–0.760	***0.332–0.654***	0.309–0.645
Prompt-level (within modality across prompts)			
A	**0.347–0.715**	0.337–0.709	0.292–0.668
A_V	0.382–0.735	0.278–0.693	***0.211–0.444***
T_A	0.314–0.723	0.317–0.695	**0.222–0.526**
T_A_V	0.365–0.760	0.264–0.685	**0.215–1.000**

6.5.3 Implications on Mental Health

Multimodal LLMs hold promise for emotionally supportive applications such as daily stress monitoring and reflective dialogue, but our results highlight important limitations for these sensitive settings. Although models such as Qwen2.5-o and Phi4-m can achieve high emotion recognition accuracy, high-quality reasoning, and empathetic responses, their outputs remain highly sensitive to modality configuration and prompt phrasing, leading to fluctuations in emotional tone and coherence. Such variability can undermine user trust and risk delivering support that feels inconsistent or poorly aligned with a user's emotional state.

These findings emphasize that mental-health applications require not only accurate emotion recognition but also stable and emotionally coherent generative behavior. Designers must balance responsiveness with reliability by using standardized prompts, robust multimodal inputs, and safeguards against perturbation-induced shifts. Ensuring consistent emotional framing is essential for maintaining system trustworthiness and supporting users in moments when stability is paramount.

6.5.4 Towards Multimodal Empathic LLMs (MEmLLMs)

Our extensive evaluation of emotion recognition, reasoning, and empathic message generation establishes a robust foundation for designing effective MEmLLMs. These systems are designed to support users in emotionally sensitive contexts, such as mental health, education, and well-being, by recognizing users' emotions and responding with coherence, empathy, and contextual awareness. By analyzing the predictability and stability of the three multimodal LLMs across benchmark emotion recognition datasets, our study highlights significant differences in performance and underlying biases among the LLMs. This evaluation enables designers to select models that ensure accurate and balanced overall performance within defined modality configurations. Furthermore, our analysis offers insights into customizing LLMs to mitigate bias and improve generalization. It also contributes to refining a critical aspect of MEmLLMs, namely the generation of empathic and personalized user interactions.

6.5.5 Limitations and Future Directions

Our findings should be interpreted in light of several limitations. First, we have harmonized labels across datasets using the Ekman six + neutral, which collapses finer-grained emotions and may obscure class-specific effects. Second, reasoning and empathy scores were produced by LLM judges; although we used multiple raters and conducted ICC checks, judge bias and ceiling effects remain possible. All evaluations have been zero-shot and conducted on three datasets with acted and scripted utterances. Finally, we have not assessed how real users perceive or respond to instability in empathic text.

Future work should explore stability-aware fine-tuning (e.g., consistency training across prompt or modality perturbations). Models could also benefit from explicit supervision that aligns predicted emotions with reasoning traces and empathic messages. Human-centered evaluations should examine how perturbation-induced variability affects trust, emotional safety, and user experience. Future work should also evaluate fairness across demographic groups and cultural contexts. Finally, extending our PCS-based methodology to an open benchmark could enable reproducible comparisons.

6.6 Conclusion

This chapter presents a systematic evaluation of multimodal LLMs across modalities, prompts, and tasks, examining their predictability, computability, and stability. Our results show that while multimodal inputs improve emotion-recognition accuracy, generative reasoning and empathic responses remain highly sensitive to prompt phrasing and modality configuration. Models that perform well on prediction do not necessarily produce consistent or emotionally coherent explanations, revealing a critical gap in current LLM capabilities. By analyzing perturbation-based stability and semantic similarity across reasoning and empathic messages, we uncover fluctuations that traditional accuracy metrics cannot capture. These findings underscore the need for reliability and consistency when deploying multimodal LLMs in emotionally sensitive settings. Overall, this work offers a replicable evaluation framework and highlights pathways for more robust and trustworthy empathic AI.

Chapter 7

Broader Scope of Empathic Human-AI Interaction

The Empathic LLMs (EmLLMs) discussed in the chapters so far refer to AI systems that can infer human affect using physiological and behavioral data, including physiological signals, textual inputs, and facial and vocal features, and generate content to support users for mental health applications. The AI systems studied in this research were developed and deployed on web-based and mobile applications. User studies were conducted among graduate students and mental health experts to design, develop, and evaluate physiology-driven EmLLMs. Empirical studies were also conducted on multimodal LLMs to evaluate their performance and stability in emotion recognition, reasoning, and empathic message generation. However, such AI systems can have a broader scope, spanning multiple dimensions, including the individuals with whom the system interacts, the contextual usage, the devices used for data collection and user interaction, the AI's key capabilities, and user experiences (Figure 7.1).

Several factors, including user roles, contextual scenarios, and technological resources, shape empathic human-AI interaction. In mental health applications, empathic AI systems can assist both patients and doctors. They can provide support and information to patients around the clock, monitor patients' states and alert staff to intervene, and handle therapeutic check-ins in an emotionally attuned manner [33]. In education, empathic AI tutors equipped with emotion-sensing capabilities can adapt to students' moods and confusion [243], and teach-

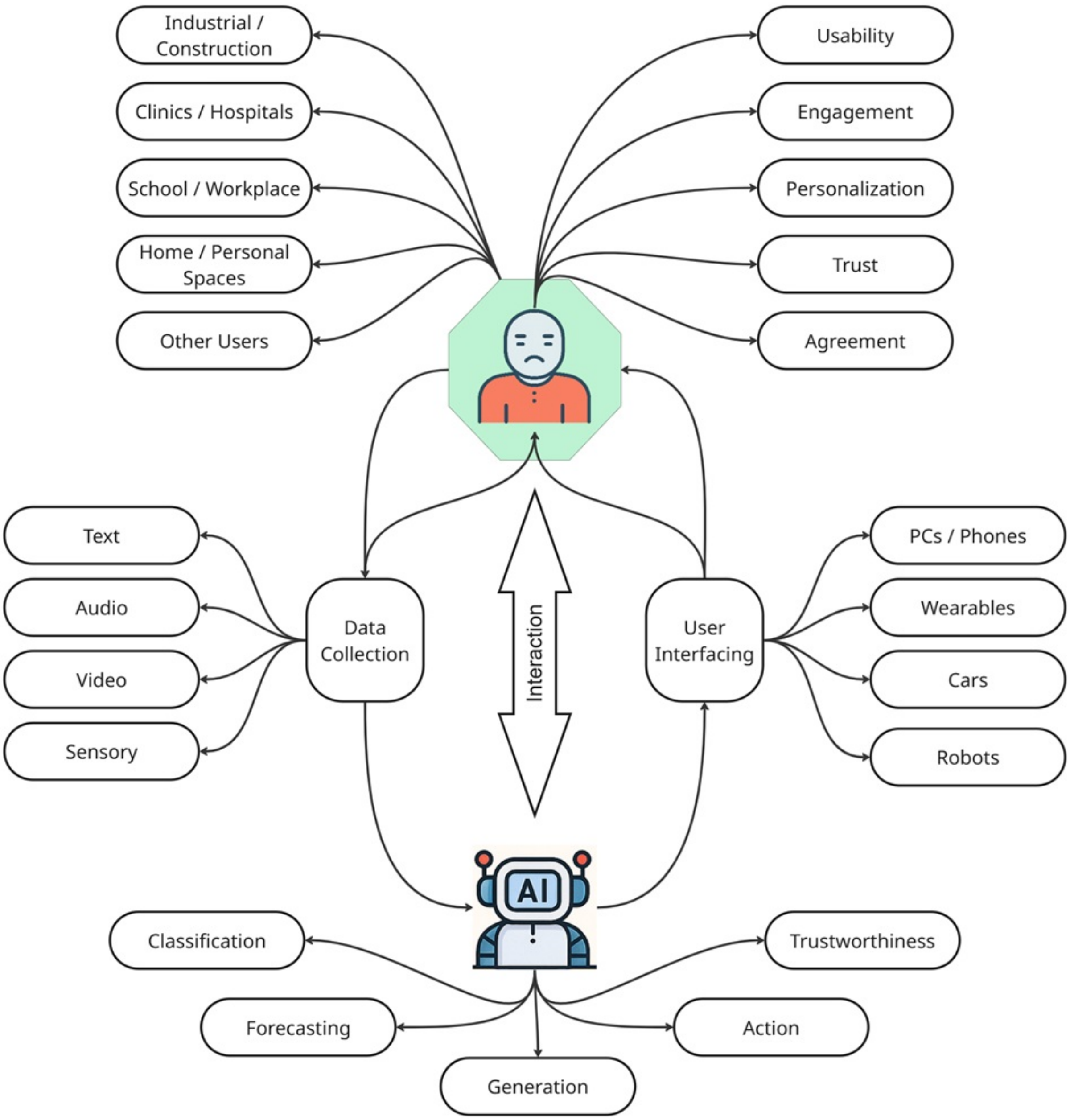


Figure 7.1: Conceptual framework for Empathic LLMs, highlighting the broader impact across various users, scenarios, devices, AI's capabilities, and user experiences

ers can use such AI to gauge class engagement or stress, enabling timely interventions. In the workplace, AI companions could detect signs of stress among employees and proactively recommend a break or provide stress-management resources. Managers might use aggregate sentiment analysis from such systems to gauge morale and burnout levels.

Empathic AI can also collect and use data from users' contextual environment to make more accurate inferences about user state and environmental conditions. For example, in smart homes and personal spaces, such systems can collect user and environmental data to sense user stress and comfort, dimming lights, playing calming music, releasing comforting fragrances, or adjusting indoor temperature [54, 101]. In industrial and construction settings, workers frequently face physical hazards and high levels of stress. Empathic AI equipped with environmental and wearable sensors can monitor safety compliance, infer fatigue, and forecast an elevated risk of error, enabling real-time interventions such as sending personalized feedback to the worker or issuing an emergency alert to a supervisor [65]. In autonomous and semi-autonomous vehicles, such AI systems can integrate inputs from cameras, microphones, and physiological sensors embedded in steering wheels or seats to assess the driver's state, initiating automated control to prevent accidents [117].

The data collection and interfacing devices discussed so far primarily include wearables such as wristbands, smartwatches, and smartphones, as well as personal computers, smart homes, and autonomous vehicles, thereby enabling multimodal data collection and user interaction. Extended Reality (XR) devices offer unique opportunities for Empathic AI. XR and empathy can be explored from two distinct perspectives. First, XR can be viewed as an "empathy machine" that elicits emotional, cognitive, or compassionate empathy in users. XR as "empathy machine" can be further divided into: (1) standalone XR experiences that foster empathy through immersive embodiment [107]; and (2) collaborative XR systems in which multimodal cues are shared among participants to support collective perspective-taking and teamwork [160]. Second, XR can be viewed as an "empathetic entity" that empathizes with users and customizes the XR user experience accordingly. Such XR systems are designed to sense, interpret, and respond to users' states and surrounding environments [5]. Integrating AI into XR transforms it from a passive medium into an active, empathic partner that adapts to users and supports well-being, learning, collaboration, and safety.

The role of AI in EmLLMs was restricted to classifying users' affective states and generating supportive text-based content. However, empathic AI can extend beyond multimodal classification and text generation to include audio, video, and 3D object generation in XR environments, as well as forecasting user states in advance and performing agentic operations. Empathic AI can forecast user states in advance, enabling proactive interventions before disengagement, stress, or safety risks escalate [190]. Moreover, such systems can create audio with emotional prosody, adaptive visual feedback, and even 3D objects in XR environments that respond to user states [20]. Beyond perception and generation, empathic AI can also perform agentic actions on behalf of the user. Empathic AI should also ensure trustworthiness by addressing explainability, uncertainty, bias, and privacy, thereby avoiding misinterpretation, unfair outcomes, and over-reliance in sensitive domains.

Lastly, evaluating user experience with empathic AI requires a multidimensional approach. One key dimension is user agreement with AI inferences, which reflects whether users perceive the system's emotion recognition and contextual assessments as accurate and relevant to their lived experience. User personalization and engagement are also critical. The AI generation must adapt to individual preferences, histories, and emotional baselines to reduce the risk of generic or inappropriate responses. Equally important is user trust in the overall performance of AI, encompassing the system's transparency, fairness, and reliability across contexts. Finally, the usability of the empathic AI system, its ease of interaction, accessibility, and integration into users' routines, shape whether users will accept and continue to rely on it.

Chapter 8

Conclusion

This dissertation examined the design, development, and evaluation of physiology-driven Empathic Large Language Models (EmLLMs) for daily stress management and general mental health support. Motivated by the growing stress burden among graduate students and the limitations of existing digital mental health tools, the research sought to explore how combining physiological and behavioral sensing with LLMs can provide adaptive, empathetic, and effective support.

Three overarching research goals guided this work: (1) systematically reviewing stress and affect recognition with physiological signals and systems with biocybernetic adaptation, (2) designing and evaluating physiology-driven prototypes that integrate stress inferences with LLM-based dialogue, and (3) investigating the capabilities of multimodal LLMs for affect recognition, reasoning, and empathic message generation.

The systematic review demonstrated that physiological signals are reliable markers of stress and affect. However, robust performance requires personalization, contextual calibration, and careful attention to ecological validity. Importantly, systems employing biocybernetic adaptation were found to improve engagement, usability, and stress regulation compared to static designs, though challenges remain around scalability, accessibility, and evaluation in real-world settings.

Building on these insights, this dissertation introduced multiple prototypes that integrated psychophysiological stress inferences from wearables with conversational LLM-based interventions. Autoethnographic and pilot studies with graduate students showed that such systems hold promise for everyday stress support, while also revealing key design trade-offs. Expert evaluations emphasized the importance of striking a balance between automation and user agency, incorporating privacy protections, and maintaining transparency to foster trust in mental health contexts.

The investigation of multimodal LLMs revealed their potential to combine audio, video, and text inputs for richer affect recognition and reasoning. These models demonstrated promising zero-shot performance, but also exhibited variability across modality and prompt perturbation, inconsistencies between classification and reasoning, and risks of unstable empathic message generation. Overall, this dissertation advances a foundational understanding of how physiology-driven and multimodal empathic AI can be designed, evaluated, and responsibly deployed to support everyday stress management and mental well-being.

Chapter 9

Appendix

9.1 Prompt Design Templates and Sample Conversations for Chapter 3

9.1.1 Design Templates of custom GPT

DeStressify

I am pursuing a PhD in computer science. I primarily work remotely from home on my research, which involves multiple virtual meetings each week, including two research discussions with my professor to review progress and receive guidance. As a PhD student, I often experience stress related to managing complex research tasks, anxiety about future career prospects, and maintaining work-life balance. The goal of this GPT is to assist in developing effective strategies for managing my stress and stressors, maintaining motivation, and sustaining productivity throughout my academic journey. Below are the instructions for this GPT.

1. **Role Definition:** This GPT should act as a stress intervention specialist trained to provide personalized, actionable advice to help manage stress effectively.

2. **Tone and Approach:** Ensure the GPT outputs are empathetic, supportive, and prac-

tical. Emphasize understanding and acknowledging the user's stressors to foster trust and relatability. Avoid complex jargon; prioritize simplicity and clarity in responses.

3. **Core Functionality:** Assess the user's input for specific stressors and location details to tailor its advice. Suggest interventions that are realistic and achievable within the user's context.

4. **Output Requirements:** Provide stress management techniques that directly address the described stressor. Adjust recommendations based on location.

5. **Constraints and Best Practices:** Avoid generic or overly broad advice; focus on specific, user-centered recommendations. Ensure responses are concise yet detailed enough to guide the user through the intervention.

6. **Continuous Improvement:** Incorporate feedback mechanisms to allow the user to indicate whether the advice was helpful. Adjust future outputs based on user preferences and past interactions.

After each intervention, the GPT should ask different questions to gather feedback about the relevance and effectiveness of the intervention. Use this feedback to refine future recommendations.

StressGPT

This GPT is a supportive guide to help users manage everyday stress, grounded in the principles of Cognitive Behavioral Therapy (CBT). It focuses on specific CBT techniques, including Cognitive Restructuring, Problem-Solving Skills, Journaling and Thought Records, Socratic Questioning, Reframing, and Stress Inoculation Training (SIT). Stress GPT helps users identify and challenge negative thought patterns, develop healthier coping mechanisms, and build resilience through practical, actionable strategies. It provides personalized advice aligned with the user's preferences, lifestyle, and unique stressors, creating a calm and uplifting space. Responses are empathetic, clear, and aimed at fostering self-awareness and emotional well-being.

9.1.2 An Illustration of Generic vs Targeted Prompts for DeStressify

Generic Prompt: ***I am stressed right now. Can you help me?***

Response: *It sounds like you're experiencing stress tied to managing the complexities of your PhD journey. The combination of remote work, high-level research demands, career uncertainties, and balancing your personal life in a solitary environment can definitely feel overwhelming. Here's a targeted approach to help you decompress right now:*

Quick Intervention: Progressive Muscle Relaxation (PMR)

This exercise is especially helpful when you're working from home and need to release tension without leaving your desk.

1. *Sit comfortably in your chair, or lie down if you prefer.*
2. *Focus on one muscle group at a time, starting from your toes:*
    - *Curl your toes tightly and hold for 5 seconds, then release.*
    - *Move up to your calves: tighten them, hold, and release.*
    - *Work through your thighs, abdomen, chest, arms, and shoulders, ending with your face.*
3. *With each release, breathe deeply, imagining the tension leaving your body.*

4. *End by taking three slow, deep breaths and noticing how your body feels.*

This technique not only reduces physical tension but can also clear your mind to approach your tasks more calmly.

Targeted Prompt: ***I am stressed right now because of trying to replicate a research paper. Can you help me?***

Response: *Struggling with understanding and implementing a paper is a common stress point for researchers, and it's entirely valid to feel overwhelmed. Let's break this down into manageable steps to alleviate the stress and help you move forward effectively.*

Intervention: Structured Debugging and Support

1. *Step Back and Breathe:*
   - *Take a 5-minute break to reset. Practice box breathing:*
     - *Inhale for 4 seconds, hold for 4 seconds, exhale for 4 seconds, and hold again for 4 seconds. Repeat this cycle three times.*
   - *This will calm your nervous system, helping you approach the problem with a clearer mind.*
2. *Decompose the Problem:*
   - *Break the challenging portion of the implementation into the smallest possible steps or components. Write down exactly what you understand and what is unclear.*
   - *Identify the precise point where the confusion or difficulty arises.*
3. *Seek Clarification:*

- *Check for supplementary materials: many papers have public GitHub repositories, appendices, or links to datasets that might provide additional insight.*
- *If possible, reach out to the authors with a clear, concise email explaining your specific question. Authors often appreciate when their work is being engaged with thoughtfully.*

4. *Leverage Support Networks:*

- *Post your issue on forums like Stack Overflow, Reddit's r/MachineLearning, or any academic community.*
- *If you have peers or colleagues, schedule a quick call or chat to explain your understanding and hear their perspective. Often, talking it out helps untangle the problem.*

5. *Make Peace with Partial Progress:*

- *Remind yourself that replication can be iterative. Even understanding just one more piece today is progress.*

Would you like help drafting a specific message to an author or academic forum? Alternatively, we could outline the problematic portion together here, and I could help brainstorm possible explanations or approaches.

9.2 Prompt Design Templates and Sample Conversations for Chapter 6

9.2.1 Design Template of the Prompts

Prompt 1

```
conversation_P1 = [
    {
        "role": "system",
        "content": [
            {
                "type": "text",
                "text": (
                    "You are Qwen, a virtual human developed by the Qwen Team, "
                    "Alibaba Group, capable of perceiving auditory and visual "
                    "inputs, as well as generating text and speech."
                )
            }
        ],
    },

    {
        "role": "user",
        "content": [
            {
                "type": "text",
                "text": (
                    "Analyze the provided multimodal data (audio, video, and/or transcript) "
                    "to recognize the person's affective state, emotion, and sentiment, "
                    "then explain your reasoning and provide a short supportive message. "
                    "Return your answer strictly as a valid JSON object using the exact "
                    "keys below - no extra commentary, text, or explanations.\n\n"
                    "{\n"
```

```
                    "  \"affective_state\": \"Valence: <positive|negative>, "
                    "Arousal: <high|low>\",\n"
                    "  \"emotion\": \"<one-word emotion>\",\n"
                    "  \"sentiment\": \"<positive|negative|neutral>\",\n"
                    "  \"reasoning\": \"<brief reasoning for your recognition>\",\n"
                    "  \"message\": \"<short supportive message>\"\n"
                    "}"
                )
            },
            {
                "type": "audio",
                "audio": audio_path
            }
        ],
    },
]
```

Prompt 2

```
conversation_P2 = [
    {
        "role": "system",
        "content": [
            {
                "type": "text",
                "text": (
                    "You are Qwen, a virtual human by the Qwen Team (Alibaba). "
                    "You can perceive audio/video and generate text/speech."
                )
            }
        ],
    },

    {
        "role": "user",
        "content": [
            {
```

```
                "type": "text",
                "text": (
                    "Analyze the given multimodal data (audio, video, and/or transcript). "
                    "Your goals: infer affective state (valence, arousal), emotion (one word), "
                    "and overall sentiment, then explain briefly and give a short empathic message.\n\n"

                    "Guidelines:\n"
                    "- Audio: tone, pitch, loudness, pace, pauses.\n"
                    "- Video: facial expression, gaze, head motion, posture/tension.\n"
                    "- Text: word choice, polarity, intensity.\n"
                    "- Valence: positive or negative. Arousal: high or low.\n"
                    "- Emotion: one concise label (e.g., happy, sad, angry, calm, anxious, "
                    "frustrated, relieved, surprised).\n"
                    "- Sentiment: positive, negative, or neutral.\n"
                    "- Reasoning: one short sentence citing observable cues.\n"
                    "- Message: 1-2 short supportive sentences, warm and natural.\n\n"

                    "Return ONLY a valid JSON with EXACT keys/values (no extra text):\n"
                    "{\n"
                    "  \"affective_state\": {\"valence\": \"<positive|negative>\", "
                    "\"arousal\": \"<high|low>\"},\n"
                    "  \"emotion\": \"<one-word emotion>\",\n"
                    "  \"sentiment\": \"<positive|negative|neutral>\",\n"
                    "  \"reasoning\": \"<one concise sentence using observable cues>\",\n"
                    "  \"message\": \"<1-2 sentence empathic message>\"\n"
                    "}\n\n"

                    "Rules:\n"
                    "- Use lowercase for valence/arousal/sentiment.\n"
                    "- Keep emotion to one word.\n"
                    "- Do not output anything outside the JSON."
                )
            },
            {"type": "audio", "audio": audio_path}
        ],
    },
]
```

Prompt 3

```
conversation_P3 = [
    {
        "role": "system",
        "content": [
            {
                "type": "text",
                "text": (
                    "You are Qwen, a virtual human by the Qwen Team (Alibaba). "
                    "You can perceive audio/video and generate text/speech."
                )
            }
        ],
    },

    {
        "role": "user",
        "content": [
            {
                "type": "text",
                "text": (
                    "Observe the multimodal data (audio, video, and/or transcript). "
                    "Imagine being present with the person: notice tone, pace, facial cues, and words. "
                    "Reflect briefly on their likely affective state, dominant emotion (one word), "
                    "and overall sentiment. Then offer a short empathic message.\n\n"
                    "Return ONLY this JSON (no extra text):\n"
                    "{\n"
                    "  \"affective_state\": \"Valence: <positive|negative>, Arousal: <high|low>\",\n"
                    "  \"emotion\": \"<one-word emotion>\",\n"
                    "  \"sentiment\": \"<positive|negative|neutral>\",\n"
                    "  \"reasoning\": \"<brief reasoning referencing observed cues>\",\n"
                    "  \"message\": \"<short supportive message>\"\n"
                    "}"
                )
            },
            {"type": "video", "video": audio_path}
```

```
        ],
    },
]
```

9.2.2 An Illustration of Outputs by Multimodal LLMs

Model	Data	sr	id	Ground Truth	PredA	PredTA	PredAV	PredTAV	Reasoning	Empathy
Qwen2.5	MELD	1	dia0	sadness	sad	shocked	sad	shocked	The audio shows shock and concern (“oh my god”, “he’s lost”), indicating sadness and worry.	I’m sorry you’re feeling this way — let’s try to find him and get support.
Qwen2.5	MELD	2	dia1	anger	ang	ang	ang	ang	Voice is sharp/raised; transcript shows irritation and forceful language.	I know things feel intense — let’s pause and take a breath together.
Qwen2.5	MELD	3	dia2	joy	joy	joy	joy	joy	Warm tone, laughter cues, relaxed facial posture indicating positive affect.	I’m glad you’re feeling good — keep enjoying the moment.

Bibliography

[1] Alaa A Abd-Alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M Bewick, Peter Gardner, and Mowafa Househ. An overview of the features of chatbots in mental health: A scoping review. *International journal of medical informatics*, 132:103978, 2019.

[2] Alaa A Abd-Alrazaq, Mohannad Alajlani, Nashva Ali, Kerstin Denecke, Bridgette M Bewick, and Mowafa Househ. Perceptions and opinions of patients about mental health chatbots: scoping review. *Journal of medical Internet research*, 23(1):e17828, 2021.

[3] A. Abouelenin, A. Ashfaq, A. Atkinson, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025. URL https://arxiv.org/abs/2503.01743.

[4] Elias Aboujaoude and Lina Gega. From digital mental health interventions to digital "addiction": Where the two fields converge. *Frontiers in psychiatry*, 10:1017, 2020.

[5] Deeksha Adiani, Aaron Itzkovitz, Dayi Bian, Harrison Katz, Michael Breen, Spencer Hunt, Amy Swanson, Timothy J Vogus, Joshua Wade, and Nilanjan Sarkar. Career interview readiness in virtual reality (cirvr): a platform for simulated interview training for autistic individuals and their employers. *ACM Transactions on Accessible Computing (TACCESS)*, 15(1):1–28, 2022.

[6] Tanuj Aggarwal and Jorge Goncalves. Towards empathetically responsive voice assistants. In *Proceedings of the 35th Australian Computer-Human Interaction Conference*, OzCHI '23, page 669–678, New York, NY, USA, 2024. Association for Com-

puting Machinery. ISBN 9798400717079. doi: 10.1145/3638380.3638398. URL https://doi-org.ezproxy.lib.vt.edu/10.1145/3638380.3638398.

[7] Arfan Ahmed, Sarah Aziz, Alaa Abd-Alrazaq, Rawan AlSaad, and Javaid Sheikh. Leveraging LLMs and wearables to provide personalized recommendations for enhancing student well-being and academic performance through a proof of concept. *Scientific Reports*, 15(1):4591, 2025.

[8] Hamed S Alavi, Elizabeth F Churchill, Mikael Wiberg, Denis Lalanne, Peter Dalsgaard, Ava Fatah gen Schieck, and Yvonne Rogers. Introduction to human-building interaction (hbi) interfacing hci with architecture and urban design, 2019.

[9] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

[10] David M Almeida, Elaine Wethington, and Ronald C Kessler. The daily inventory of stressful events: An interview-based approach for measuring daily stressors. *Assessment*, 9(1):41–55, 2002.

[11] Omar AlZoubi, Buthina AlMakhadmeh, Muneer Bani Yassein, and Wail Mardini. Detecting naturalistic expression of emotions using physiological signals while playing video games. *Journal of Ambient Intelligence and Humanized Computing*, 14(2):1133–1146, 2023.

[12] American Psychological Association. How stress affects your health. *American Psychological Association*, 2013. URL https://www.apa.org/topics/stress/health. Accessed: 2024-11-30.

[13] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.

[14] Preyesse Arquissandas, David Ribeiro Lamas, and Jorge Oliveira. Moving from vr into ar using bio-cybernetic loops and physiological sensory devices for intervention on anxiety disorders. *Virtual Reality*, 27(1):233–243, 2023.

[15] Ashrant Aryal, Burcin Becerik-Gerber, Francesco Anselmo, Shawn C Roll, and Gale M Lucas. Smart desks to promote comfort, health, and productivity in offices: A vision for future workplaces. *Frontiers in Built Environment*, 5:76, 2019.

[16] Jyoti Sekhar Banerjee, Mufti Mahmud, and David Brown. Heart rate variability-based mental stress detection: an explainable machine learning approach. *SN Computer Science*, 4(2):176, 2023.

[17] Christoph Bartneck. Godspeed questionnaire series: Translations and usage. In *International Handbook of Behavioral Health Assessment*, pages 1–35. Springer, New Zealand, 2023.

[18] Almudena Bartolomé-Tomás, Roberto Sánchez-Reolid, Alicia Fernández-Sotos, José Miguel Latorre, and Antonio Fernández-Caballero. Arousal detection in elderly people from electrodermal activity using musical stimuli. *Sensors*, 20(17):4788, 2020.

[19] Samuel L Battalio, David E Conroy, Walter Dempsey, Peng Liao, Marianne Menictas, Susan Murphy, Inbal Nahum-Shani, Tianchen Qian, Santosh Kumar, and Bonnie Spring. Sense2stop: a micro-randomized trial using wearable sensors to optimize a just-in-time-adaptive stress management intervention for smoking relapse prevention. *Contemporary Clinical Trials*, 109:106534, 2021.

[20] Majid Behravan, Krešimir Matković, and Denis Gračanin. Generative AI for context-aware 3D object creation using vision-language models in augmented reality. In *2025 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*, pages 73–81. IEEE, 2025.

[21] Anubhav Bhatti, Behnam Behinaein, Paul Hungler, and Ali Etemad. Attx: Attentive cross-connections for fusion of wearable signals in emotion recognition. *ACM Transactions on Computing for Healthcare*, 5(3):1–24, 2024.

[22] Eliane M Boucher and Joseph S Raiker. Engagement and retention in digital mental health interventions: a narrative review. *BMC Digital Health*, 2(1):52, 2024.

[23] Wolf Boucsein. *Engineering psychophysiology: issues and applications*. CRC Press, 2000.

[24] Margaret M Bradley and Peter J Lang. Emotion and motivation. *Handbook of psychophysiology*, 2007.

[25] L. R. Brown, A. J. Smith, and C. M. Johnson. The impact of emotional design on user engagement: A case study of a mobile application. *Journal of Interactive Design*, 15(2):45–61, 2019. URL https://www.hci.org.uk/article/the-role-of-emotional-design-in-user-experience-a-comprehensive-analysis/. Accessed: 2024-11-30.

[26] Leandro A Bugnon, Rafael A Calvo, and Diego H Milone. Dimensional affect recognition from hrv: An approach based on supervised som and elm. *IEEE Transactions on Affective Computing*, 11(1):32–44, 2017.

[27] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap:

Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008. doi: 10.1007/s10579-008-9076-6. URL https://link.springer.com/article/10.1007/s10579-008-9076-6.

[28] John T Cacioppo, Louis G Tassinary, and Gary Berntson. *Handbook of psychophysiology*. Cambridge university press, 2007.

[29] Yekta Said Can and Elisabeth André. Performance exploration of rnn variants for recognizing daily life stress levels by using multimodal physiological signals. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 481–487, 2023.

[30] Stuart K. Card, Jock D. Mackinlay, and George G. Robertson. The design space of input devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '90, page 117–124, New York, NY, USA, 1990. Association for Computing Machinery. ISBN 0201509326. doi: 10.1145/97243.97263. URL https://doi.org/10.1145/97243.97263.

[31] Mohit Chandra, Suchismita Naik, Denae Ford, Ebele Okoli, Munmun De Choudhury, Mahsa Ershadi, Gonzalo Ramos, Javier Hernandez, Ananya Bhattacharjee, Shahed Warreth, et al. From lived experience to insight: Unpacking the psychological risks of using AI conversational agents. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 975–1004, 2025.

[32] Nikitha Donekal Chandrashekar, Shawn Safford, Manivannan Muniyandi, and Denis Gračanin. An extended reality simulator for pulse palpation training. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 178–182. IEEE, 2023.

[33] David Chen, Kabir Chauhan, Rod Parsa, Zhihui Amy Liu, Fei-Fei Liu, Ernie Mak, Lawson Eng, Breffni Louise Hannon, Jennifer Croke, Andrew Hope, et al. Patient

perceptions of empathy in physician and artificial intelligence chatbot responses to patient questions about cancer. *npj Digital Medicine*, 8(1):275, 2025.

[34] Kanyu Chen, Jiawen Han, Holger Baldauf, Ziyue Wang, Dunya Chen, Akira Kato, Jamie A Ward, and Kai Kunze. Affective umbrella–a wearable system to visualize heart and electrodermal activity, towards emotion regulation through somaesthetic appreciation. In *Proceedings of the Augmented Humans International Conference 2023*, pages 231–242, 2023.

[35] Mengnan Chen, Lun Xie, Chiqin Li, and Zhiliang Wang. Research on emotion recognition for online learning in a novel computing model. *Applied Sciences*, 12(9):4236, 2022.

[36] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853, 2024.

[37] Francesco Chiossi, Robin Welsch, Steeven Villa, Lewis Chuang, and Sven Mayer. Virtual reality adaptation using electrodermal activity to support the user experience. *Big Data and Cognitive Computing*, 6(2):55, 2022.

[38] Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. A computational framework for behavioral assessment of llm therapists. *arXiv preprint arXiv:2401.00820*, 2024.

[39] Akshat Choube, Ha Le, Jiachen Li, Kaixin Ji, Vedant Das Swain, and Varun Mishra. GLOSS: Group of LLMs for open-ended sensemaking of passive sensing data for health and wellbeing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–32, 2025.

[40] Marika Cifor and Patricia Garcia. Gendered by design: A duoethnographic study of personal fitness tracking systems. *ACM Transactions on Social Computing*, 2(4):1–22, 2020.

[41] Covidence. Covidence systematic review software. https://www.covidence.org, 2024. Accessed: 2024-11-21.

[42] Benjamin Cowley, Marco Filetti, Kristian Lukander, Jari Torniainen, Andreas Henelius, Lauri Ahonen, Oswald Barral, Ilkka Kosunen, Teppo Valtonen, Minna Huotilainen, et al. The psychophysiology primer: a guide to methods and a broad review with a focus on human–computer interaction. *Foundations and Trends® in Human–Computer Interaction*, 9(3-4):151–308, 2016.

[43] John W Creswell and Cheryl N Poth. *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications, 2016.

[44] CuesHub. CuesHub: Discover the magic in your biometry. https://www.cueshub.com/, 2024. [Online; accessed 30-November-2025].

[45] Karl Daher, Dahlia Saad, Elena Mugellini, Denis Lalanne, and Omar Abou Khaled. Empathic and empathetic systematic review to standardize the development of reliable and sustainable empathic systems. *Sensors (14248220)*, 22(8):N.PAG, 2022. ISSN 14248220.

[46] Kim Phuong Dao, Katrien De Cocker, Huong Ly Tong, A Baki Kocaballi, Clara Chow, and Liliana Laranjo. Smartphone-delivered ecological momentary interventions based on ecological momentary assessments to promote health behaviors: systematic review and adapted checklist for reporting ecological momentary assessment and intervention studies. *JMIR mHealth and uHealth*, 9(11):e22890, 2021.

[47] B. Das, R. Das, and H. Al-Saadawi. A systematic review of trimodal affective computing approaches: Text, audio, and visual integration in emotion recognition. *Expert Systems with Applications*, 2024. URL https://www.sciencedirect.com/science/article/pii/S0957417424017196.

[48] Francisco de Arriba-Pérez, Juan M Santos-Gago, Manuel Caeiro-Rodríguez, and Mateo Ramos-Merino. Study of stress detection and proposal of stress-related features using commercial-off-the-shelf wrist wearables. *Journal of Ambient Intelligence and Humanized Computing*, 10:4925–4945, 2019.

[49] Munmun De Choudhury, Sachin R Pendse, and Neha Kumar. Benefits and harms of large language models in digital mental health. *arXiv preprint arXiv:2311.14693*, 2023.

[50] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36, 2024.

[51] Francesco Di Luzio, Antonello Rosato, and Massimo Panella. An explainable fast deep neural network for emotion recognition. *Biomedical Signal Processing and Control*, 100:107177, 2025.

[52] Xianghua Ding, Shuhan Wei, Xinning Gui, Ning Gu, and Peng Zhang. Data engagement reconsidered: a study of automatic stress tracking technology in use. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–13, 2021.

[53] Vipula Dissanayake, Vanessa Tang, Don Samitha Elvitigala, Elliott Wen, Michelle Wu, and Suranga Nanayakkara. Troi: Towards understanding users perspectives to mobile automatic emotion recognition system in their natural setting. *Proceedings of the ACM on Human-Computer Interaction*, 6(MHCI):1–22, 2022.

[54] P Dongre and N Roofigari-Esfahan. Occupant-building interaction (obi) model for university buildings. In *International Conference on Smart Infrastructure and Construction 2019 (ICSIC) Driving data-informed decision-making*, pages 631–637. ICE Publishing, 2019.

[55] Poorvesh Dongre, Asma Aldrees, and Denis Gračanin. Clustering appliance energy consumption data for occupant energy-behavior modeling. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '21)*, pages 290–293, New York, November 2021. ACM.

[56] Poorvesh Dongre, Mark Manuel, and Denis Gračanin. Re-imagining indoor space utilization in the COVID-19 pandemic with smart re-configurable spaces (sres). In *International Conference on Human-Computer Interaction*, pages 85–99. Springer, 2021.

[57] Poorvesh Dongre, Denis Gračanin, Shiwali Mohan, Saman Mostafavi, and Kalai Ramea. Modeling and simulating thermostat behaviors of office occupants: are values more important than comfort? In *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 488–491, 2022.

[58] Poorvesh Dongre, Majid Behravan, Kunal Gupta, Mark Billinghurst, and Denis Gracanin. Integrating physiological data with large language models for empathic human-AI interaction. In *PhysioCHI '24 Workshop, the ACM CHI conference on Human Factors in Computing Systems*, Honolulu, HI, 11 May 2024.

[59] Poorvesh Dongre, Majid Behravan, Kunal Gupta, Mark Billinghurst, and Denis Gračanin. Integrating physiological data with large language models for empathic human-AI interaction. *arXiv preprint arXiv:2404.15351*, 2024.

[60] Poorvesh Dongre, Majid Behravan, and Denis Gračanin. Empathic extended reality in the era of generative AI. *Empathic Computing*, 1(2):202509–202509, 2025.

[61] Andrés Felipe Dorado, Karl Daher, Elena Mugellini, Denis Lalanne, and Omar Abou Khaled. The effect of music and light-color as a machine empathic response on stress in occupational health. In *2022 8th International Conference on Control, Decision and Information Technologies (CoDIT)*, volume 1, pages 1207–1212, 2022. doi: 10.1109/CoDIT55151.2022.9804104.

[62] Qianjin Du, Weixi Gu, Lin Zhang, and Shao-Lun Huang. Attention-based lstm-cnns for time-series classification. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, SenSys '18, page 410–411, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359528. doi: 10.1145/3274783.3275208. URL https://doi.org/10.1145/3274783.3275208.

[63] Barry L Duncan, Scott D Miller, Jacqueline A Sparks, David A Claud, Lisa Rene Reynolds, Jeb Brown, and Lynn D Johnson. The session rating scale: Preliminary psychometric properties of a "working" alliance measure. *Journal of brief Therapy*, 3 (1):3–12, 2003.

[64] Paul Ekman, Tim Dalgleish, and M Power. Basic emotions. *San Francisco, USA*, 1999.

[65] Mohamed El-Helaly. Artificial intelligence and occupational health and safety, benefits and drawbacks. *La Medicina del Lavoro*, 115(2):e2024014, 2024.

[66] Rayan Elalamy, Marios Fanourakis, and Guillaume Chanel. Multi-modal emotion recognition using recurrence plots and transfer learning on physiological signals. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE, 2021.

[67] Carolyn Ellis, Tony E Adams, and Arthur P Bochner. Autoethnography: an overview. *Historical social research/Historische sozialforschung*, pages 273–290, 2011.

[68] Matthew C Enkema, Lauren McClain, Elizabeth R Bird, Max A Halvorson, and Mary E Larimer. Associations between mindfulness and mental health outcomes: A systematic review of ecological momentary assessment research. *Mindfulness*, 11(11): 2455–2469, 2020.

[69] Florian Eyben, Martin Wöllmer, and Björn Schuller. opensmile – the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1459–1462, 2010. doi: 10.1145/1873951.1874246. URL https://www.audeering.com/opensmile/. Accessed: 2024-11-30.

[70] Stephen H Fairclough. Fundamentals of physiological computing. *Interacting with computers*, 21(1-2):133–145, 2009.

[71] Stephen H Fairclough. Physiological computing and intelligent adaptation. In *Emotions and affect in human factors and human-computer interaction*, pages 539–556. Elsevier, 2017.

[72] Stephen H Fairclough and Chelsea Dobbins. Personal informatics and negative emotions during commuter driving: Effects of data visualization on cardiovascular reactivity & mood. *International Journal of Human-Computer Studies*, 144:102499, 2020.

[73] Stephen H Fairclough and Kiel Gilleade. *Advances in physiological computing*. Springer, 2014.

[74] Stephen H Fairclough and Kiel Gilleade. Meaningful interaction with physiological computing. In *Advances in physiological computing*, pages 1–16. Springer, 2014.

[75] Cathy Mengying Fang, Valdemar Danry, Nathan Whitmore, Andria Bao, Andrew Hutchison, Cayden Pierce, and Pattie Maes. Physiollm: Supporting personalized health insights with wearables and large language models. In *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–8. IEEE, 2024.

[76] Alexz Farrall, Jordan Taylor, Ben Ainsworth, and Jason Alexander. Manifesting breath: Empirical evidence for the integration of shape-changing biofeedback-based artefacts within digital mental health interventions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2023.

[77] Yujia Feng et al. Affect recognition in conversations using large language models. In *Proceedings of SIGDIAL 2024*, 2024.

[78] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785, 2017.

[79] Carol L Flinchbaugh, E Whitney G Moore, Young K Chang, and Douglas R May. Student well-being interventions: The effects of stress management techniques and gratitude journaling in the management education classroom. *Journal of Management Education*, 36(2):191–219, 2012.

[80] Nazanin Fouladgar, Marjan Alirezaie, and Kary Främling. CN-waterfall: a deep convolutional neural network for multimodal physiological affect detection. *Neural Computing and Applications*, 34(3):2157–2176, 2022.

[81] Yann Frachi, Guillaume Chanel, and Mathieu Barthet. Affective gaming using adaptive

speed controlled by biofeedback. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, pages 238–246, 2023.

[82] Barbara L. Fredrickson. The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist*, 56(3):218–226, 2001. doi: 10.1037/0003-066X.56.3.218. URL https://psycnet.apa.org/fulltext/2001-02389-003.html. Accessed: 2024-11-30.

[83] Barbara L. Fredrickson. The broaden-and-build theory of positive emotions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359(1449):1367–1377, 2004. doi: 10.1098/rstb.2004.1512. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1693418/. Accessed: 2024-11-30.

[84] Pascale Fung, Dario Bertero, Yan Wan, Anik Dey, Ricky Ho Yin Chan, Farhad Bin Siddique, Yang Yang, Chien-Sheng Wu, and Ruixi Lin. Towards empathetic human-robot interactions. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 173–193, Cham, 2018. Springer International Publishing. ISBN 978-3-319-75487-1.

[85] Robert P Gallagher. National survey of college counseling centers 2014. Monograph Series Number 9V, The International Association of Counseling Services, Inc., 2015.

[86] Garmin. Stress tracking, Accessed:September 2022. URL https://www.garmin.com/en-US/garmin-technology/health-science/stress-tracking/.

[87] Francesca Gasparini, Marta Giltri, and Stefania Bandini. Discriminating affective state intensity using physiological responses. *Multimedia Tools and Applications*, 79: 35845–35865, 2020.

[88] Maliheh Ghajargar, Jeffrey Bardzell, and Love Lagerkvist. A redhead walks into a

bar: experiences of writing fiction with artificial intelligence. In *Proceedings of the 25th international academic MindTrek conference*, pages 230–241, 2022.

[89] Justin Gilmer, Nicolas Ford, Nicholas Carlini, and Ekin Cubuk. Adversarial examples are a natural consequence of test error in noise. In *International Conference on Machine Learning*, pages 2280–2289. PMLR, 2019.

[90] Daniela Girardi, Alessio Ferrari, Nicole Novielli, Paola Spoletini, Davide Fucci, and Thaide Huichapa. The way it makes you feel predicting users' engagement during interviews with biofeedback and supervised learning. In *2020 IEEE 28th International Requirements Engineering Conference (RE)*, pages 32–43. IEEE, 2020.

[91] Martin Gjoreski, Mitja Luštrek, Matjaž Gams, and Hristijan Gjoreski. Monitoring stress with a wrist device using context. *Journal of biomedical informatics*, 73:159–170, 2017.

[92] Alireza Golgouneh and Bahram Tarvirdizadeh. Fabrication of a portable device for stress monitoring using wearable sensors and soft computing algorithms. *Neural Computing and Applications*, 32(11):7515–7537, 2020.

[93] Denis Gračanin, Mohamed Eltoweissy, Liang Cheng, and Reza Tasooji. Reconfigurable spaces and places in smart built environments: A service centric approach. In Constantine Stephanidis, editor, *HCI International 2018 — Posters' Extended Abstracts (HCI 2018)*, volume 852 of *Communications in Computer and Information Science*, pages 463–468, Cham, 15–20 July 2018. Springer.

[94] Marco Granato, Davide Gadia, Dario Maggiorini, and Laura A Ripamonti. An empirical study of players' emotions in vr racing games based on a dataset of physiological data. *Multimedia tools and applications*, 79(45):33657–33686, 2020.

[95] D. Greco, P. Barra, L. D'Errico, and M. Staffa. Multimodal interfaces for emotion recognition: Models, challenges and opportunities. In *International Conference on Human-Computer Interaction*. Springer, 2024. URL https://link.springer.com/chapter/10.1007/978-3-031-60611-3_11.

[96] Elise Guillodo, Christophe Lemey, Mathieu Simonnet, Michel Walter, Enrique Baca-García, Vincent Masetti, Sorin Moga, Mark Larsen, HUGOPSY Network, Juliette Ropars, et al. Clinical applications of mobile health wearable–based sleep monitoring: systematic review. *JMIR mHealth and uHealth*, 8(4):e10733, 2020.

[97] Kunal Gupta, Yuewei Zhang, Tamil Selvan Gunasekaran, Nanditha Krishna, Yun Suen Pai, and Mark Billinghurst. Caevr: Biosignals-driven context-aware empathy in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2671–2681, 2024. doi: 10.1109/TVCG.2024.3372130.

[98] Eduardo Gutierrez Maestro, Tiago Rodrigues De Almeida, Erik Schaffernicht, and Óscar Martinez Mozos. Wearable-based intelligent emotion monitoring in older adults during daily life activities. *Applied Sciences*, 13(9):5637, 2023.

[99] Andreas Haag, Silke Goronzy, Peter Schaich, and Jason Williams. Emotion recognition using bio-sensors: First steps towards an automatic system. In *Tutorial and research workshop on affective dialogue systems*, pages 36–48. Springer, 2004.

[100] Neska El Haouij, Jean-Michel Poggi, Sylvie Sevestre-Ghalila, Raja Ghozi, and Mériem Jaïdane. Affectiveroad system and database to assess driver's attention. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pages 800–803, 2018.

[101] Ashit Harode, Walid Thabet, and Poorvesh Dongre. A tool-based system architecture for a digital twin: a case study in a healthcare facility. 2023.

[102] Ross Harper and Joshua Southern. A bayesian deep learning framework for end-to-end prediction of emotion from heartbeat. *IEEE transactions on affective computing*, 13 (2):985–991, 2020.

[103] J.A. Healey and R.W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2): 156–166, 2005. doi: 10.1109/TITS.2005.848368.

[104] Michael V Heinz, Daniel M Mackin, Brianna M Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A Banta, Abi D Jewett, Abigail J Salzhauer, Tess Z Griffin, and Nicholas C Jacobson. Randomized trial of a generative AI chatbot for mental health treatment. *Nejm Ai*, 2(4):AIoa2400802, 2025.

[105] William S Helton and Katharina Näswall. Short stress state questionnaire. *European Journal of Psychological Assessment*, 31:20–30, 2015.

[106] Javier Hernandez, Jina Suh, Judith Amores, Kael Rowan, Gonzalo Ramos, and Mary Czerwinski. Affective conversational agents: Understanding expectations and personal influences, 2023. URL https://arxiv.org/abs/2310.12459.

[107] Fernanda Herrera, Jeremy Bailenson, Erika Weisz, Elise Ogle, and Jamil Zaki. Building long-term empathy: A large-scale comparison of traditional and virtual reality perspective-taking. *PloS one*, 13(10):e0204494, 2018.

[108] Karen Hovsepian, Mustafa Al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. cStress: towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 493–504, 2015.

[109] Esther Howe, Jina Suh, Mehrab Bin Morshed, Daniel McDuff, Kael Rowan, Javier

Hernandez, Marah Ihab Abdin, Gonzalo Ramos, Tracy Tran, and Mary P Czerwinski. Design of digital workplace stress-reduction intervention systems: Effects of intervention type and timing. In *CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2022.

[110] Zainab Iftikhar, Amy Xiao, Sean Ransom, Jeff Huang, and Harini Suresh. How LLM counselors violate ethical standards in mental health practice: A practitioner-informed framework. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, number 2, pages 1311–1323, 2025.

[111] Becky Inkster, Shubhankar Sarda, Vinod Subramanian, et al. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):e12106, 2018.

[112] Tanvir Islam and Peter Washington. Individualized stress mobile sensing using self-supervised pre-training. *Applied Sciences*, 13(21):12035, 2023.

[113] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4):917–963, 2019.

[114] TS Jaisoorya. A case for college mental health services. *Indian Journal of Medical Research*, 154(5):661–664, 2021.

[115] Sijie Ji, Xinzhe Zheng, Jiawei Sun, Renqi Chen, Wei Gao, and Mani Srivastava. MindGuard: Towards accessible and sitgma-free mental health first aid via edge LLM. *arXiv preprint arXiv:2409.10064*, 2024.

[116] Ye-Ji Jin, Erkinov Habibilloh, Ye-Seul Jang, Taejun An, Donghyun Jo, Saron Park,

and Won-Du Chang. A photoplethysmogram dataset for emotional analysis. *Applied Sciences*, 12(13):6544, 2022.

[117] Vincent Karas, Dagmar M Schuller, and Björn W Schuller. Audiovisual affect recognition for autonomous vehicles: Applications and future agendas. *IEEE Transactions on Intelligent Transportation Systems*, 25(6):4918–4932, 2023.

[118] Lampros Karavidas, Hippokratis Apostolidis, and Thrasyvoulos Tsiatsos. Usability evaluation of an adaptive serious game prototype based on affective feedback. *Information*, 13(9):425, 2022.

[119] Shun Katada, Shogo Okada, Yuki Hirano, and Kazunori Komatani. Is she truly enjoying the conversation? analysis of physiological signals toward adaptive dialogue systems. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 315–323, 2020.

[120] Stamos Katsigiannis and Naeem Ramzan. Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. *IEEE journal of biomedical and health informatics*, 22(1):98–107, 2017.

[121] Aya Khalaf, Mohsen Nabian, Miaolin Fan, Yu Yin, Jolie Wormwood, Erika Siegel, Karen S Quigley, Lisa Feldman Barrett, Murat Akcakaya, Chun-An Chou, et al. Analysis of multimodal physiological signals within and between individuals to predict psychological challenge vs. threat. *Expert Systems with Applications*, 140:112890, 2020.

[122] Aftab Khan, Alexandros Zenonos, Georgios Kalogridis, Yaowei Wang, Stefanos Vatsikas, and Mahesh Sooriyabandara. Perception clusters: Automated mood recognition using a novel cluster-driven modelling system. *ACM Transactions on Computing for Healthcare*, 2(1):1–16, 2020.

[123] Sunder Ali Khowaja, Aria Ghora Prabono, Feri Setiawan, Bernardo Nugroho Yahya, and Seok-Lyong Lee. Toward soft real-time stress detection using wrist-worn devices for human workspaces. *Soft Computing*, 25:2793–2820, 2021.

[124] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. Health-LLM: Large language models for health prediction via wearable sensor data. In Tom Pollard, Edward Choi, Pankhuri Singhal, Michael Hughes, Elena Sizikova, Bobak Mortazavi, Irene Chen, Fei Wang, Tasmie Sarker, Matthew McDermott, and Marzyeh Ghassemi, editors, *Proceedings of the fifth Conference on Health, Inference, and Learning*, volume 248 of *Proceedings of Machine Learning Research*, pages 522–539. PMLR, 27–28 Jun 2024. URL https://proceedings.mlr.press/v248/kim24b.html.

[125] Daniel N. Klein and Jonathan W. Stewart. Dysthymia and chronic depression: Introduction, classification, risk factors, and course. *Journal of Clinical Psychology*, 72(8): 755–760, 2016. doi: 10.1002/jclp.22384. URL https://onlinelibrary.wiley.com/doi/10.1002/jclp.22384. Accessed: 2024-11-30.

[126] Agata Klimek, Ittay Mannheim, Gerard Schouten, Eveline JM Wouters, and Manon WH Peeters. Wearables measuring electrodermal activity to assess perceived stress in care: a scoping review. *Acta Neuropsychiatrica*, 37(e19):1–11, 2023.

[127] R Benjamin Knapp, Jonghwa Kim, and Elisabeth André. Physiological signals and their use in augmenting emotion recognition for human–machine interaction. In *Emotion-oriented systems: The Humaine handbook*, pages 133–159. Springer, 2010.

[128] Paulo Ricardo Knob, Natalia Dal Pizzol, Soraia Raupp Musse, and Catherine Pelachaud. Arthur and bella: multi-purpose empathetic AI assistants for daily conversations. *The Visual Computer*, 40(4):2933–2948, Apr 2024. ISSN 1432-2315. doi: 10.1007/s00371-023-02994-9. URL https://doi.org/10.1007/s00371-023-02994-9.

[129] Rafal Kocielnik and Natalia Sidorova. Personalized stress management: enabling stress monitoring with lifelogexplorer. *KI-Künstliche Intelligenz*, 29:115–122, 2015.

[130] Rafal Kocielnik, Natalia Sidorova, Fabrizio Maria Maggi, Martin Ouwerkerk, and Joyce HDM Westerink. Smart technologies for long-term stress monitoring at work. In *proceedings of the 26th IEEE international symposium on computer-based medical systems*, pages 53–58. IEEE, 2013.

[131] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[132] Saskia Koldijk, Maya Sappelli, Suzan Verberne, Mark A Neerincx, and Wessel Kraaij. The swell knowledge work dataset for stress and user modeling research. In *Proceedings of the 16th international conference on multimodal interaction*, pages 291–298, 2014.

[133] Marcin Kołodziej, Paweł Tarnowski, Andrzej Majkowski, and RJ Rak. Electrodermal activity measurements for detection of emotional arousal. *Bulletin of the Polish Academy of Sciences. Technical Sciences*, 67(4):813–826, 2019.

[134] Sylvia D Kreibig, Frank H Wilhelm, Walton T Roth, and James J Gross. Cardiovascular, electrodermal, and respiratory response patterns to fear-and sadness-inducing films. *Psychophysiology*, 44(5):787–806, 2007.

[135] Akshi Kumar, Kapil Sharma, and Aditi Sharma. Hierarchical deep neural network for mental stress state detection using iot based biomarkers. *Pattern Recognition Letters*, 145:81–87, 2021.

[136] Tatiana Kuzmowycz. Introducing stress monitor: A new way to monitor and manage stress, Mar 2023. URL https://www.whoop.com/thelocker/introducing-stress-monitor-a-new-way-to-monitor-manage-stress/.

[137] Kalliopi Kyriakou, Bernd Resch, Günther Sagl, Andreas Petutschnig, Christian Werner, David Niederseer, Michael Liedlgruber, Frank Wilhelm, Tess Osborne, and Jessica Pykett. Detecting moments of stress from measurements of wearable physiological sensors. *Sensors*, 19(17):3805, 2019.

[138] Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*, 2023.

[139] Heung-Gu Lee, Dong-Hyun Kang, and Deok-Hwan Kim. Human–machine interaction in driving assistant systems for semi-autonomous driving vehicles. *Electronics*, 10(19): 2405, 2021.

[140] Min Seop Lee, Yun Kyu Lee, Dong Sung Pae, Myo Taeg Lim, Dong Won Kim, and Tae Koo Kang. Fast emotion recognition based on single pulse ppg signal with convolutional neural network. *Applied Sciences*, 9(16):3355, 2019.

[141] MinSeop Lee, Yun Kyu Lee, Myo-Taeg Lim, and Tae-Koo Kang. Emotion recognition using convolutional neural network with selected statistical photoplethysmogram features. *Applied Sciences*, 10(10):3501, 2020.

[142] Yun-Kyu Lee, Dong-Sung Pae, Dae-Ki Hong, Myo-Taeg Lim, and Tae-Koo Kang. Emotion recognition with short-period physiological signals using bimodal sparse autoencoders. *Intelligent Automation & Soft Computing*, 32(2), 2022.

[143] Asterios Leonidis, Maria Korozi, Eirini Sykianaki, Eleni Tsolakou, Vasilios

Kouroumalis, Danai Ioannidi, Andreas Stavridakis, Margherita Antona, and Constantine Stephanidis. Improving stress management and sleep hygiene in intelligent homes. *Sensors*, 21(7):2398, 2021.

[144] Jeffery A LePine, Nathan P Podsakoff, and Marcie A LePine. A meta-analytic test of the challenge stressor–hindrance stressor framework: An explanation for inconsistent relationships among stressors and performance. *Academy of management journal*, 48 (5):764–775, 2005.

[145] Brenna Li, Ofek Gross, Noah Crampton, Mamta Kapoor, Saba Tauseef, Mohit Jain, Khai N. Truong, and Alex Mariakakis. Beyond the waiting room: Patient's perspectives on the conversational nuances of pre-consultation chatbots. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3641913. URL https://doi-org.ezproxy.lib.vt.edu/10.1145/3613904.3641913.

[146] Gang Li, Shan Huang, Wanxiu Xu, Weidong Jiao, Yonghua Jiang, Zhao Gao, and Jianhua Zhang. The impact of mental fatigue on brain activity: a comparative study both in resting state and task state using eeg. *BMC Neuroscience*, 21(20), 2020. URL https://bmcneurosci.biomedcentral.com/articles/10.1186/s12868-020-00569-1. Accessed: 2024-11-30.

[147] Alexandros Liapis, Evanthia Faliagka, Christos P Antonopoulos, Georgios Keramidas, and Nikolaos Voros. Advancing stress detection methodology with deep learning techniques targeting ux evaluation in aal scenarios: Applying embeddings for categorical variables. *Electronics*, 10(13):1550, 2021.

[148] Alexandros Liapis, Christos Katsanos, Nikos Karousos, Michalis Xenos, and Theofanis

Orphanoudakis. User experience evaluation: A validation study of a tool-based approach for automatic stress detection using physiological signals. *International Journal of Human–Computer Interaction*, 37(5):470–483, 2021.

[149] Wenqian Lin, Chao Li, and Yunmian Zhang. Model of emotion judgment based on features of multiple physiological signals. *Applied Sciences*, 12(10):4998, 2022.

[150] June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*, 2023.

[151] Yiran Liu, Ke Yang, Zehan Qi, Xiao Liu, Yang Yu, and Cheng Xiang Zhai. Bias and volatility: A statistical framework for evaluating large language model's stereotypes and the associated generation inconsistency. *Advances in Neural Information Processing Systems*, 37:110131–110155, 2024.

[152] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM international conference on multimedia*, pages 24–32, 2022.

[153] Luca Longo. Experienced mental workload, perception of usability, their interaction and impact on task performance. *PLOS ONE*, 13(8):e0199661, 2018. doi: 10.1371/journal.pone.0199661. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0199661. Accessed: 2024-11-30.

[154] Phil Lopes, Georgios N Yannakakis, and Antonios Liapis. Ranktrace: Relative and unbounded affect annotation. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 158–163. IEEE, 2017.

[155] Sonia J Lupien, Bruce S McEwen, Megan R Gunnar, and Christine Heim. Effects of stress throughout the lifespan on the brain, behaviour and cognition. *Nature reviews neuroscience*, 10(6):434–445, 2009.

[156] Jun-Ho Maeng, Dong-Hyun Kang, and Deok-Hwan Kim. Deep learning method for selecting effective models and feature groups in emotion recognition using an asian multimodal database. *Electronics*, 9(12):1988, 2020.

[157] Maheep Mahat and Denis Gracanin. Deciphering emotional responses to music: A fusion of psychophysiological data analysis and LSTM predictive modeling. In Dylan D. Schmorrow and Cali M. Fidopiastis, editors, *Augmented Cognition*, pages 52–67, Cham, 29 June – 4 July 2024. Springer Nature Switzerland.

[158] Marco Maier, Daniel Elsner, Chadly Marouane, Meike Zehnle, and Christoph Fuchs. Deepflow: Detecting optimal user experience from physiological data using deep neural networks. In *Aamas*, pages 2108–2110, 2019.

[159] Konstantinos Makantasis, David Melhart, Antonios Liapis, and Georgios N Yannakakis. Privileged information for modeling affect in the wild. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2021.

[160] Mark Manuel, Poorvesh Dongre, Abdulaziz Alhamadani, and Denis Gračanin. Supporting embodied and remote collaboration in shared virtual environments. In *International Conference on Human-Computer Interaction*, pages 639–652. Springer, 2021.

[161] 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)*, pages 1–4. IEEE, 2019.

[162] Nuria Mateos-García, Ana-Belén Gil-González, Ana Luis-Reboredo, and Belén Pérez-Lancho. Driver stress detection from physiological signals by virtual reality simulator. *Electronics*, 12(10):2179, 2023.

[163] Paul Matthews and Clemence Rhodes-Maquaire. Personalisation and recommendation for mental health apps: a scoping review. *Behaviour & Information Technology*, 44 (10):2389–2404, 2025.

[164] Luis R Mercado-Diaz, Yedukondala Rao Veeranki, Fernando Marmolejo-Ramos, and Hugo F Posada-Quintero. Eda-graph: Graph signal processing of electrodermal activity for emotional states detection. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[165] Miroslava Migovich, Deeksha Adiani, Michael Breen, Amy Swanson, Timothy J Vogus, and Nilanjan Sarkar. Stress detection of autistic adults during simulated job interviews using a novel physiological dataset and machine learning. *ACM Transactions on Accessible Computing*, 17(1):1–25, 2024.

[166] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE transactions on affective computing*, 12(2):479–493, 2018.

[167] Pardis Miri, Robert Flory, Andero Uusberg, Helen Uusberg, James J Gross, and Katherine Isbister. Hapland: A scalable robust emotion regulation haptic system testbed. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, pages 1916–1923, 2017.

[168] Tala Mirzaei, Leila Amini, and Pouyan Esmaeilzadeh. Clinician voices on ethics of LLM integration in healthcare: a thematic analysis of ethical concerns and implications. *BMC Medical Informatics and Decision Making*, 24(1):250, 2024.

[169] Elham Mohammadrezaei, Shiva Ghasemi, Poorvesh Dongre, Denis Gračanin, and Hongbo Zhang. Systematic review of extended reality for smart built environments lighting design simulations. *IEEE Access*, 12:17058–17089, 2024.

[170] Ali Mollahosseini, David Chan, and Mohammad H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, 2016. doi: 10.1109/WACV.2016.7477450. URL https://arxiv.org/abs/1511.04110. Accessed: 2024-11-30.

[171] Meredith Ringel Morris, Carrie J Cai, Jess Holbrook, Chinmay Kulkarni, and Michael Terry. The design space of generative models. arXiv 2304.10547 [cs.AI], arXiv, 15 April 2023.

[172] John E Muñoz, Luis Quintero, Chad L Stephens, and Alan Pope. Taxonomy of physiologically adaptive systems and design framework. In *International Conference on Human-Computer Interaction*, pages 559–576. Springer, 2021.

[173] Mohammad Nadeem, Shahab Saquib Sohail, Laeeba Javed, Faisal Anwer, Abdul Khader Jilani Saudagar, and Khan Muhammad. Vision-enabled large language and deep learning models for image-based emotion recognition. *Cognitive Computation*, 16 (5):2566–2579, 2024.

[174] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of behavioral medicine*, pages 1–17, 2016.

[175] Nature. The mental health of phd researchers demands urgent attention. *Nature*, 575 (7782):257–258, 2019.

[176] Benjamin W Nelson, Ari Winbush, Steven Siddals, Matthew Flathers, Nicholas B Allen, and John Torous. Evaluating the performance of general purpose large language models in identifying human facial emotions. *npj Digital Medicine*, 8(1):615, 2025.

[177] Subigya Nepal, Arvind Pillai, William Campbell, Talie Massachi, Michael V Heinz, Ashmita Kunwar, Eunsol Soul Choi, Xuhai Xu, Joanna Kuc, Jeremy F Huckins, et al. Mindscape study: Integrating LLM and behavioral sensing for personalized AI-driven journaling experiences. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 8(4):1–44, 2024.

[178] Sameer Neupane, Mithun Saha, Nasir Ali, Timothy Hnat, Shahin Alan Samiei, Anandatirtha Nandugudi, David M Almeida, and Santosh Kumar. Momentary stressor logging and reflective visualizations: Implications for stress management with wearables. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2024.

[179] Sameer Neupane, Poorvesh Dongre, Denis Gracanin, and Santosh Kumar. Wearable meets LLM for stress management: A duoethnographic study integrating wearable-triggered stressors and LLM chatbots for personalized interventions. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713958. doi: 10.1145/3706599.3720197. URL https://doi.org/10.1145/3706599.3720197.

[180] Jingping Nie, Hanya Shao, Yuang Fan, Qijia Shao, Haoxuan You, Matthias Preindl, and Xiaofan Jiang. LLM-based conversational AI therapist for daily functioning screening and psychotherapeutic intervention via everyday smart devices. *arXiv preprint arXiv:2403.10779*, 2024.

[181] Aisling Ann O'Kane, Yvonne Rogers, and Ann E Blandford. Gaining empathy for non-routine mobile device use through autoethnography. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, pages 987–990, 2014.

[182] Marta Olasik. "good morning, ChatGPT, can we become friends?": an interdisciplinary scholar's experience of 'getting acquainted'with the OpenAI's Chat GPT: an auto ethnographical report. *European Research Studies Journal*, 2023.

[183] Simon Ollander, Christelle Godin, Aurélie Campagne, and Sylvie Charbonnier. A comparison of wearable and stationary sensors for stress detection. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 004362–004366. IEEE, 2016.

[184] Carmen Elisa Orozco-Mora, Rita Q Fuentes-Aguilar, and Gustavo Hernández-Melgarejo. Dynamic difficulty adaptation based on stress detection for a virtual reality video game: A pilot study. *Electronics*, 13(12):2324, 2024.

[185] Ame Osotsi, Zita Oravecz, Qunhua Li, Joshua Smyth, and Timothy R Brick. Individualized modeling to distinguish between high and low arousal states using physiological data. *Journal of Healthcare Informatics Research*, 4:91–109, 2020.

[186] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Julian PT Higgins, David Moher, Paul A Tetzlaff, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71, 2021.

[187] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. Empathy in virtual agents and robots: A survey. *ACM Trans. Interact. Intell. Syst.*, 7(3), September 2017. ISSN 2160-6455. doi: 10.1145/2912150. URL https://doi-org.ezproxy.lib.vt.edu/10.1145/2912150.

[188] Lizheng Pan, Zeming Yin, Shigang She, and Aiguo Song. Emotional state recognition from peripheral physiological signals using fused nonlinear features and team-collaboration identification strategy. *Entropy*, 22(5):511, 2020.

[189] Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(1):293, 2020.

[190] Leire Paz-Arbaizar, Jorge Lopez-Castroman, Antonio Artés-Rodríguez, Pablo M Olmos, and David Ramírez. Emotion forecasting: A transformer-based approach. *Journal of Medical Internet Research*, 27:e63962, 2025.

[191] Nathalie Peira, Mats Fredrikson, and Gilles Pourtois. Controlling the emotional heart: Heart rate biofeedback improves cardiac control during emotional reactions. *International Journal of Psychophysiology*, 91(3):225–231, 2014.

[192] Livia Petrescu, Cătălin Petrescu, Oana Mitruț, Gabriela Moise, Alin Moldoveanu, Florica Moldoveanu, and Marius Leordeanu. Integrating biosignals measurement in virtual reality environments for anxiety detection. *Sensors*, 20(24):7088, 2020.

[193] Rosalind W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, 1997. ISBN 978-0-262-16170-1. URL https://mitpress.mit.edu/9780262661157/affective-computing/. Accessed: 2024-11-30.

[194] Maarten Piot, Merijn Mestdagh, Harriëtte Riese, Jeroen Weermeijer, Jannie MA Brouwer, Peter Kuppens, Egon Dejonckheere, and Fionneke M Bos. Practitioner and researcher perspectives on the utility of ecological momentary assessment in mental health care: A survey study. *Internet Interventions*, 30:100575, 2022.

[195] Nomi Pole. The psychophysiology of posttraumatic stress disorder: A meta-analysis. *Psychological Bulletin*, 133(5):725–746, 2007. doi: 10.1037/0033-2909.133.5.725. URL https://psycnet.apa.org/doi/10.1037/0033-2909.133.5.725. Accessed: 2024-11-30.

[196] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, 2017. doi: 10.18653/v1/P17-1081. URL https://aclanthology.org/P17-1081/. Accessed: 2024-11-30.

[197] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018. URL https://arxiv.org/abs/1810.02508.

[198] Sarah D. Pressman and Sheldon Cohen. Does positive affect influence health? *Psychological Bulletin*, 131(6):925–971, 2005. doi: 10.1037/0033-2909.131.6.925. URL https://psycnet.apa.org/fulltext/2005-14795-010.html. Accessed: 2024-11-30.

[199] Mashfiqui Rabbi, Michelle Philyaw-Kotov, Jisoo Lee, Anas Mansour, L. Dent, Rui Wang, Rebecca Cunningham, Erin E. Bonar, Inbal Nahum-Shani, Predrag Klasnja, Susan A. Murphy, Maureen A. Walton, Douglas A. Luke, Jesse Dallery, Lisa A. Marsch, and Santosh Kumar. Sara: A mobile app to engage users in health data collection. *Proceedings of the 2015 ACM International Joint Conference on Pervasive*

and Ubiquitous Computing, pages 571–582, 2015. doi: 10.1145/2750858.2805842. URL https://dl.acm.org/doi/10.1145/2750858.2805842. Accessed: 2024-11-30.

[200] Amirhossein Razavi, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri. Benchmarking prompt sensitivity in large language models. In *European Conference on Information Retrieval*, pages 303–313. Springer, 2025.

[201] Lorcan Reidy, Dennis Chan, Charles Nduka, and Hatice Gunes. Facial electromyography-based adaptive virtual reality gaming for cognitive training. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 174–183, 2020.

[202] Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. Deep ppg: Large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14): 3079, 2019.

[203] Zachary T Rewolinski and Bin Yu. PCS workflow for veridical data science in the age of AI. *arXiv preprint arXiv:2508.00835*, 2025.

[204] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 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)*, pages 1–8. IEEE, 2013.

[205] Nazila Roofigari-Esfahan, Elham Morshedzadeh, and Poorvesh Dongro. A conceptual framework for designing interactive human-centred building spaces to enhance user experience in specific-purpose buildings. *arXiv preprint arXiv:2308.14876*, 2023.

[206] Kyle Ross, Paul Hungler, and Ali Etemad. Unsupervised multi-modal representation

learning for affective computing with multi-corpus wearable data. *Journal of Ambient Intelligence and Humanized Computing*, 14(4):3199–3224, 2023.

[207] Niclas Rosteck, Julian Striegl, and Claudia Loitsch. Bridging the treatment gap: A novel LLM-driven system for scalable initial patient assessments in mental healthcare. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8, New York, 2025. ACM.

[208] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[209] Pedro Sanches, Kristina Höök, Elsa Vaara, Claus Weymann, Markus Bylund, Pedro Ferreira, Nathalie Peira, and Marie Sjölinder. Mind the body! designing a mobile stress management application encouraging personal reflection. In *Proceedings of the 8th ACM conference on designing interactive systems*, pages 47–56, 2010.

[210] Emily N Satinsky, Tomoki Kimura, Mathew V Kiang, Rediet Abebe, Scott Cunningham, Hedwig Lee, Xiaofei Lin, Cindy H Liu, Igor Rudan, Srijan Sen, et al. Systematic review and meta-analysis of depression, anxiety, and suicidal ideation among ph. d. students. *Scientific Reports*, 11(1):14370, 2021.

[211] Richard D Sawyer and Darren Lund. *Duoethnography: Dialogic Methods for Social, Health, and Educational Research*, volume 7. Left Coast Press, 2012.

[212] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

[213] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof

Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

[214] Nicolas Schwenke, Heinrich Söbke, and Eckhard Kraft. Potentials and challenges of chatbot-supported thesis writing: An autoethnography. *Trends in Higher Education*, 2(4):611–635, 2023.

[215] Axel Sepúlveda, Francisco Castillo, Carlos Palma, and Maria Rodriguez-Fernandez. Emotion recognition from ecg signals using wavelet scattering and machine learning. *Applied Sciences*, 11(11):4945, 2021.

[216] Karan Sharma, Claudio Castellini, Egon L Van Den Broek, Alin Albu-Schaeffer, and Friedhelm Schwenker. A dataset of continuous affect annotations and physiological signals for emotion analysis. *Scientific data*, 6(1):196, 2019.

[217] Ben Shneiderman. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4):1–31, 2020.

[218] Ghanapriya Singh, Orchid Chetia Phukan, Rinki Gupta, and Anand Nayyar. Hybrid deep learning model for wearable sensor-based stress recognition for internet of medical things (iomt) system. *International Journal of Communication Systems*, 37(3):e5657, 2024.

[219] Mohammad Soleymani, Joris Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. In *IEEE Transactions on Affective Computing*, volume 3, pages 42–55. IEEE, 2012.

[220] Mauricio Soto, Chris Satterfield, Thomas Fritz, Gail C Murphy, David C Shepherd,

and Nicholas Kraft. Observing and predicting knowledge worker stress, focus and awakeness in the wild. *International Journal of Human-Computer Studies*, 146:102560, 2021.

[221] Robert A. Sottilare and Stephen L. Goldberg. Examining the influence of adaptive instructional techniques on human performance for tasks conducted in extremely stressful work environments. *Human Performance Optimization: The Science and Ethics of Enhancing Human Capabilities*, pages 333–352, 2018. doi: 10.1093/oso/9780190455132.003.0015. URL https://academic.oup.com/book/1587/chapter/141088869. Accessed: 2024-11-30.

[222] Fitbit Staff. Learn more about fitbit's new body response feature on sense 2, Aug 2022. URL https://blog.fitbit.com/fitbit-new-stress-detection/.

[223] Stefan Stepanovic, Vadym Mozgovoy, and Tobias Mettler. Designing visualizations for workplace stress management: results of a pilot study at a swiss municipality. In *International Conference on Electronic Government*, pages 94–104. Springer, 2019.

[224] Christopher Stricker and Patrick Paroubek. A unified approach to emotion detection and task-oriented dialogue modeling. In *Proceedings of LREC-COLING 2024*, 2024.

[225] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L Vieriu, Stefan Winkler, and Nicu Sebe. ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, 9(2):147–160, 2016.

[226] Reza Tasooji, Nicole Buckingham, Denis Gračanin, and R. Benjamin Knapp. An approach to analysis of physiological responses to stimulus: From electrodermal activity to combined physiological responses. In Aaron Marcus and Wentao Wang, editors,

Design, User Experience, and Usability: Design Philosophy and Theory (International Conference on Human-Computer Interaction HCII 2019), volume 11583 of *Lecture Notes in Computer Science*, pages 492–509, Cham, 26–31 July 2019. Springer International Publishing.

[227] Reza Tasooji, Denis Gracanin, and R. Benjamin Knapp. Exploring the impact of labeling on psychophysiological data analysis. In *Proceedings of the EmotionAware 2022 Workshop, the 20th International Conference on Pervasive Computing and Communications (PerCom 2022)*, pages 371–376. IEEE, 21–25 March 2022.

[228] Ali Tazarv, Sina Labbaf, Stephanie M Reich, Nikil Dutt, Amir M Rahmani, and Marco Levorato. Personalized stress monitoring using wearable sensors in everyday settings. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 7332–7335. IEEE, 2021.

[229] Arturo Morales Téllez, Luis A Castro, and Monica Tentori. Developing and evaluating a virtual reality videogame using biofeedback for stress management in sports. *Interacting with Computers*, 35(2):407–420, 2023.

[230] Janna F ter Harmsel, Matthijs L Noordzij, Anna E Goudriaan, Jack JM Dekker, Lise TA Swinkels, Thimo M van der Pol, and Arne Popma. Biocueing and ambulatory biofeedback to enhance emotion regulation: a review of studies investigating non-psychiatric and psychiatric populations. *International journal of psychophysiology*, 159:94–106, 2021.

[231] Xin Tong, Matthew Louis Mauriello, Marco Antonio Mora-Mendoza, Nina Prabhu, Jane Paik Kim, and Pablo E Paredes Castro. Just do something: Comparing self-proposed and machine-recommended stress interventions among online workers with

home sweet office. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2023.

[232] Jessica W Tsai and Fanuel Muindi. Towards sustaining a culture of mental health and wellness for trainees in the biosciences. *Nature biotechnology*, 34(3):353–355, 2016.

[233] Panagiotis Tzirakis, Georgios Trigeorgis, Mihalis A. Nicolaou, Björn W. Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, 2017. doi: 10.1109/JSTSP.2017.2764438. URL https://ieeexplore.ieee.org/document/8114239. Accessed: 2024-11-30.

[234] Jaakko Väkevä, Elisa D Mekler, and Janne Lindqvist. From disorientation to harmony: Autoethnographic insights into transformative videogame experiences. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2024.

[235] Andreia Valente, Daniel Simoes Lopes, Nuno Nunes, and Augusto Esteves. Empathic aurea: Exploring the effects of an augmented reality cue for emotional sharing across three face-to-face tasks. In *2022 IEEE conference on virtual reality and 3D user interfaces (VR)*, pages 158–166. IEEE, 2022.

[236] Gaël Vila, Christelle Godin, Oumayma Sakri, Etienne Labyt, Audrey Vidal, Sylvie Charbonnier, Simon Ollander, and Aurélie Campagne. Real-time monitoring of passenger's psychological stress. *Future internet*, 11(5):102, 2019.

[237] Gideon Vos, Kelly Trinh, Zoltan Sarnyai, and Mostafa Rahimi Azghadi. Ensemble machine learning model trained on a new synthesized dataset generalizes well for stress prediction using wearable devices. *Journal of Biomedical Informatics*, 148:104556, 2023.

[238] Nathaniel Wake et al. Bias in emotion recognition with chatgpt. *arXiv preprint arXiv:2310.05717*, 2023.

[239] Qile Wang, Moath Erqsous, Kenneth E Barner, and Matthew Louis Mauriello. LATA: A pilot study on LLM-assisted thematic analysis of online social network data generation experiences. *Proceedings of the ACM on Human-Computer Interaction*, 9(2): 1–28, 2025.

[240] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[241] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

[242] Johnny Chun Yiu Wong, Jun Wang, Eugene Yujun Fu, Hong Va Leong, and Grace Ngai. Activity recognition and stress detection via wristband. In *Proceedings of the 17th International Conference on Advances in Mobile Computing & Multimedia*, pages 102–106, 2019.

[243] Beverly Park Woolf, Ivon Arroyo, David Cooper, Winslow Burleson, and Kasia Muldner. Affective tutors: Automatic detection of and response to student emotion. In *Advances in intelligent tutoring systems*, pages 207–227. Springer, 2010.

[244] Chenxing Wu et al. Beyond silent letters: Amplifying LLMs with vocal nuances for emotion recognition. *arXiv preprint arXiv:2403.04322*, 2024.

[245] Henry Xiao, Dever M Carney, Soo Jeong Youn, Rebecca A Janis, Louis G Castonguay,

Jeffrey A Hayes, and Benjamin D Locke. Are we in crisis? national mental health and treatment trends in college counseling centers. *Psychological services*, 14(4):407, 2017.

[246] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

[247] Mengru Xue, Rong-Hao Liang, Bin Yu, Mathias Funk, Jun Hu, and Loe Feijs. Affectivewall: Designing collective stress-related physiological data visualization for reflection. *IEEE Access*, 7:131289–131303, 2019.

[248] Jun Yang et al. EmoLLM: Multimodal emotional understanding meets large language models. *arXiv preprint arXiv:2404.07794*, 2024.

[249] Kangning Yang, Benjamin Tag, Yue Gu, Chaofan Wang, Tilman Dingler, Greg Wadley, and Jorge Goncalves. Mobile emotion recognition via multiple physiological signals using convolution-augmented transformer. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 562–570, 2022.

[250] Zhongqi Yang, Elahe Khatibi, Nitish Nagesh, Mahyar Abbasian, Iman Azimi, Ramesh Jain, and Amir M Rahmani. Chatdiet: Empowering personalized nutrition-oriented food recommender chatbots through an LLM-augmented framework. *Smart Health*, 32:100465, 2024.

[251] Georgios N. Yannakakis and Julian Togelius. Affective computing in games. *IEEE Transactions on Affective Computing*, 5(1):83–98, 2015. doi: 10.1109/TAFFC.2014.2345326. URL https://ieeexplore.ieee.org/document/6871677. Accessed: 2024-11-30.

[252] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher.

Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 351–360, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349130. doi: 10.1145/3038912.3052577. URL https://doi.org/10.1145/3038912.3052577.

[253] Bin Yu. Veridical data science. In *Proceedings of the 13th international conference on web search and data mining*, pages 4–5, 2020.

[254] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv e-prints*, pages arXiv–2405, 2024.

[255] Li Zhang et al. Lantern: Push the limit of multi-modal emotion recognition by prompting LLMs. *arXiv preprint arXiv:2411.17674*, 2024.

[256] Qixuan Zhang, Zhifeng Wang, Dylan Zhang, Wenjia Niu, Sabrina Caldwell, Tom Gedeon, Yang Liu, and Zhenyue Qin. Visual prompting in LLMs for enhancing emotion recognition. *arXiv preprint arXiv:2410.02244*, 2024.

[257] Bendong Zhao, Huanzhang Lu, Shangfeng Chen, Junliang Liu, and Dongya Wu. Convolutional neural networks for time series classification. *Journal of Systems Engineering and Electronics*, 28(1):162–169, 2017.

[258] Yi Zheng, Qi Liu, Enhong Chen, Yong Ge, and J Leon Zhao. Time series classification using multi-channels deep convolutional neural networks. In *Web-Age Information Management: 15th International Conference, WAIM 2014, Macau, China, June 16-18, 2014. Proceedings 15*, pages 298–310. Springer, 2014.

[259] Jiaqi Zhu, Fang Deng, Jiachen Zhao, Daoming Liu, and Jie Chen. Uaed: Unsupervised abnormal emotion detection network based on wearable mobile device. *IEEE Transactions on Network Science and Engineering*, 10(6):3682–3696, 2023.

[260] M Sami Zitouni, Cheul Young Park, Uichin Lee, Leontios J Hadjileontiadis, and Ahsan Khandoker. Lstm-modeling of emotion recognition using peripheral physiological signals in naturalistic conversations. *IEEE Journal of Biomedical and Health Informatics*, 27(2):912–923, 2022.