Emotion-Aware Digital Twin for a Large Language Model-Based Personalized Therapy Solution

by

Karim Al Ghoul

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for
the Ph.D. degree in Electrical and Computer Engineering

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Abstract

Mental health disorders are increasing worldwide, yet many individuals still face limited access to timely mental health support. At the same time, wearable devices such as smartwatches enable continuous collection of physiological signals that may provide useful indicators of affective states in everyday life. This thesis explores how wearable-based emotion recognition can be integrated with retrieval-augmented large language models (RAG-LLMs) and a digital twin framework to provide accessible, personalized, and context-aware mental well-being support anytime and anywhere.

First, the thesis introduces WARM-VR, a new affective computing dataset designed to address limitations of existing resources by combining immersive stimuli with wearable physiological sensing and self-report. The dataset includes recordings from 31 participants and integrates multimodal signals with arousal, valence, and relaxation annotations to support reproducible affect research in more realistic settings. Second, the thesis develops and evaluates multiple deep learning architectures for emotion recognition from Photoplethysmography (PPG)-based wearables, including hybrid CNN–LSTM–TCN models as well as Transformer-based and Mamba-based approaches, with attention to noise, subject variability, and class imbalance. Third, the thesis investigates the use of Retrieval-Augmented Generation (RAG) to reduce hallucination and improve reliability in mental-health-oriented conversational systems, and introduces an evaluation framework that distinguishes between retrieval-grounded factual accuracy and therapist-like conversational support across different LLMs.

Building on these components, the thesis proposes UbiMyTherapist (*You Be My Therapist*), a ubiquitous emotion-aware digital twin framework designed to operate continuously alongside the user by combining wearable-based emotion estimation, user history, and psychological knowledge bases, enabling reactive conversational support and proactive interventions. A proof-of-concept prototype is implemented using an agentic orchestration layer and a vector-based RAG pipeline grounded in Cognitive Behavioral Therapy (CBT) content. Finally, a user case study with 24 participants compares four system configurations: baseline GPT-4o, prompt-engineered GPT-4o, RAG-based GPT-4o, and the full digital twin prototype. The results from the pilot study show that integrating retrieval grounding with user context and emotional-state history improves perceived therapist-like conversation, empathy, personalization, and overall user experience.

Acknowledgment

First and foremost, I would like to thank God for granting me the strength, patience, and resilience to complete my PhD journey.

I would like to dedicate this thesis to my father, **Bahije Al Ghoul**, who devoted his life to ensuring that my siblings and I received the best possible education. Born in 1948, at a time when completing high school was itself a remarkable achievement, my father dreamed of finishing his studies and pursuing a university education. However, one year before graduation, he had to leave school in order to support his parents and siblings. Although he was unable to fulfill that dream himself, he carried it forward through his children. His sacrifices, guidance, and unwavering belief in us have shaped who I am today. I am also deeply grateful to my entire family, who have supported me from the very beginning, never doubted my abilities, and stood by me throughout this journey.

أودّ أن أهدي هذه الأطروحة إلى والدي، **بهيج الغول**، الذي كرّس حياته لضمان حصولي أنا وإخوتي على أفضل تعليم ممكن. وُلد والدي عام 1948، في زمنٍ كان فيه إكمال المرحلة الثانوية إنجازًا عظيمًا بحدّ ذاته، وكان يحلم بإتمام دراسته ومتابعة تعليمه الجامعي. إلا أنّه، قبل عام واحد من التخرّج، اضطر إلى ترك المدرسة لمساعدة والديه وإخوته. وعلى الرغم من أنّه لم يتمكّن من تحقيق هذا الحلم لنفسه، فقد حمله في قلبه وحقّقه من خلال أبنائه. لقد كان لتضحياته وتوجيهه وإيمانه الراسخ بنا أثرٌ كبيرٌ في تشكيل الشخص الذي أنا عليه اليوم. كما أعبّر عن عميق امتناني لعائلتي بأكملها، التي دعمتني منذ البداية، ولم تشكّ يومًا في قدراتي، ووقفت إلى جانبي طوال هذه الرحلة

This achievement would not have been possible without the unconditional love, patience, and support of my wife, Rayan. Her sacrifices, encouragement, and constant presence gave me the strength to keep moving forward. She made this dream possible, and for that I am forever grateful.

I would also like to express my deepest gratitude and appreciation to my supervisors, Dr. Abdulmotaleb El Saddik and Dr. Hussein Al Osman, from whom I have learned so much. Their guidance, support, and mentorship began even before my PhD, and have continued throughout this journey. I am truly thankful for their trust, encouragement, and invaluable contributions to my academic and personal growth.

Table of Contents

List of Figures

List of Tables

List of Definitions

ACC — Acceleration Signal

ANS — Autonomic Nervous System

AR — Augmented Reality

BERT — Bidirectional Encoder Representations from Transformers

Bi-LSTM — Bidirectional LSTM

CBT — Cognitive Behavioral Therapy

CNN — Convolutional Neural Network

EDA — Electrodermal Activity

ECG — Electrocardiogram

EEG — Electroencephalogram

EI — Emotional Intelligence

EMG — Electromyography

GRU — Gated Recurrent Unit

GSR — Galvanic Skin Response

HF — High-Frequency

HCI — Human-Computer Interaction

HRV — Heart Rate Variability

LLM — Large Language Model

LSTM — Long Short-Term Memory

NLP — Natural Language Processing

PPG — Photoplethysmography

QA — Question-Answering

RAG — Retrieval-Augmented Generation

ReLU — Rectified Linear Unit

RESP — Respiration

RNN — Recurrent Neural Network

ROUGE — Recall-Oriented Understudy for Gisting Evaluation

SAM — Self-Assessment Manikin

RRS — Relaxation Rating Scale

TCN — Temporal Convolutional Networks

TEMP — Temperature Signal

TSST — Trier Social Stress Test

WDT — Well-being Digital Twin

VR — Virtual Reality

Chapter 1. Introduction

1.1 Problem statement

Access to timely mental well-being is limited. People are continually exposed to a wide range of emotionally salient stimuli in their everyday lives. It is often challenging for them to distinguish between typical reactions and potential indicators of elevated distress or reduced well-being. To better understand their emotions and mental states and to learn how to respond to them, individuals would benefit immensely from timely feedback.

This problem affects a broad range of individuals, including people experiencing mild to moderate symptoms, people who are not yet engaged in formal care, and people who face barriers to accessing professional mental health services. Common barriers include long wait times, limited clinician availability, cost, and geographic constraints. As a result, support is often intermittent, and many individuals do not receive consistent, timely feedback that helps them better understand their emotions and mental states.

Existing digital mental health solutions partially address this gap, but they remain limited in practice. Many applications rely on generic content and do not adapt to the user's personal context over time. Conversational agents can provide immediate interaction, but they may produce inconsistent or ungrounded responses, and they often lack mechanisms for reliable personalization and safety. In parallel, while wearable devices are now widely used and can continuously capture physiological signals, most consumer solutions focus on physical health indicators and do not provide robust emotion-aware feedback that generalizes across users and real-life conditions.

Therefore, the problem addressed in this thesis is the lack of accessible, safe, and personalized mental well-being support that can operate outside clinical settings, reflect the user's evolving emotional state, provide meaningful feedback when needed, and complement professional mental health care when it is available.

This thesis focuses on a support tool that provides guidance and timely feedback to improve self-awareness and day-to-day well-being. The system is designed to be used independently in everyday life and to complement professional mental health care when available. It is not a replacement for clinical diagnosis or treatment.

1.2 Motivation

When timely mental well-being support is not available in daily life, many individuals receive help only after symptoms worsen or when problems begin to affect their functioning, relationships, or physical health. This is especially harmful for people with mild-to-moderate symptoms, people who are not yet engaged in formal care, and people who face barriers such as long wait times, cost, or limited local services. The result is that a large number of individuals experience persistent unmet or partially met needs, despite the growing demand for mental health support.

1.2.1 Increase in Mental Disorders

In 2022, more than 15% of Canadians aged 15 and older were reported to have a mental illness, totaling approximately 5 million individuals, thereby emphasizing the increasing prevalence of mental health issues [1]. According to the World Health Organization, one in every eight individuals worldwide lives with a mental disorder, and the majority lack access to effective care [2]. These numbers highlight that mental health challenges are common and affect a large portion of the population.

1.2.2 Mental Health Care System

A substantial proportion of individuals who need support do not receive it in a timely or sufficient manner. Over a third (36.6%) of Canadians with mental disorders have reported that their needs for healthcare and mental health services are either unmet or only partially met [3]. When support is delayed or incomplete, individuals may experience prolonged distress and reduced quality of life, and the likelihood of escalation to more severe conditions can increase.

1.2.3 Mental Health vs Physical Health

The consequences of unmet mental health needs are not limited to emotional well-being. Mental health is closely connected to physical health; for example, depression can significantly increase the risk of chronic physical conditions, including diabetes, heart disease, and stroke [4].

1.2.4 Growth in Wearable Technology

Wearable technologies have seen rapid adoption in recent years and are now widely used to monitor physical health indicators such as heart rate, sleep patterns, and physical activity [5]. These devices are primarily designed to support fitness and physical well-being, while their potential for monitoring mental and emotional states remains relatively underexplored. Although physiological signals

collected by wearables are closely linked to emotional responses, most consumer applications do not yet leverage this information to provide mental health–related feedback or support [6]. At the same time, the continuous nature of wearable sensing enables the capture of physiological responses in natural, everyday environments. This continuous monitoring enables the detection of changes in affective states beyond controlled laboratory settings. Several studies have shown that physiological data collected using wearable devices can achieve performance comparable to laboratory-grade equipment for affect recognition tasks, supporting the feasibility of wearable-based emotion detection in real-world scenarios [7].

However, continuous data collection alone is not sufficient to support mental well-being. Wearable signals must be interpreted reliably under real-world conditions, and the resulting feedback must adapt to the user over time. This requires modeling emotional states in a way that accounts for individual differences and evolving personal context. These requirements motivate the use of a digital twin-based approach, in which the user's emotional state is continuously updated and used to provide personalized support.

1.2.5 Digital Twins and Emotional Intelligence (EI)

Emotionally intelligent machines can significantly enhance Human-Computer Interaction (HCI) by enabling systems to interpret and respond to user needs more effectively [8]. In this context, digital twins can leverage EI to improve quality of life and well-being by continuously modeling the user's state over time [9]. This is especially relevant in consumer health settings, where wearable devices and smart clothing generate continuous personal data streams that can support real-time monitoring and feedback.

As mentioned in [10], digital twins have significant potential for applications in mental health. This direction aligns with the concept of a Well-being Digital Twin (WDT) [11], which is typically built by combining continuous wearable sensing with personal records and an AI-driven feedback loop to support monitoring and personalized recommendations over time.

While there has been considerable research on WDT solutions, important challenges persist in practice. In this thesis, I introduce three complementary solutions that address several of these challenges. The next section describes these challenges in detail.

1.3 Challenges and Research Problems

The main drivers of the digital twin solution to address our problem statement are the need for EI and grounded responses. Therefore, the main focus is on tackling challenges in detecting the individual's emotional state and making Large Language Models more accurate, personalized and grounded in psychology literature.

1.3.1 Emotion Detection

Detecting emotion from wearable physiological signals remains challenging because the signals are highly sensitive to noise and may be affected by motion artifacts and other recording conditions. Physiological measurements can be influenced by multiple factors beyond emotion, including physical movement, posture changes, and individual baseline differences.

This challenge becomes more significant when using wearable PPG signals. Compared to clinical-grade sensors, wearable PPG is more sensitive to motion artifacts, sensor placement, and contact quality, and it may vary across devices. In addition, physiological responses differ across individuals, so a model that performs well for some subjects may not generalize well to unseen users.

Another difficulty is label uncertainty. Emotion labels are typically obtained using self-reports, which are subjective and may not perfectly align with the timing of physiological changes. This creates noisy ground truth and increases the difficulty of supervised learning. Finally, affective datasets often include limited samples and class imbalance, which can lead to overfitting and misleading performance if evaluation is not carefully designed.

These challenges motivate the need for robust modeling and evaluation strategies for wearable-based emotion recognition, which is a key component of the emotion-aware digital twin proposed in this thesis.

1.3.2 Dataset limitations

To develop an effective model for emotion and affect recognition, a high-quality dataset is required for training.

Many current models are based on limited datasets because collecting data is both challenging and costly. As a result, the lack of data complicates emotion recognition, particularly for deep learning models that rely on relatively large training datasets.

The existing datasets for emotional analysis, particularly those utilizing PPG signals, exhibit several limitations:

- **Limited Number of Participants:** Many datasets include a small number of participants. For instance, the WESAD [12] dataset has only 15 participants, and the PPGE [13] dataset has only 18 participants. This limitation restricts the generalizability of the findings and poses challenges in effectively training large network models [12], [14].
- **Short Duration of signal recorded:** The experiments used to evoke emotions in participants are often short, typically around one minute. Such brief durations may be insufficient to elicit strong emotional responses, potentially compromising the reliability of the collected emotional data [13].
- **Dependence on Specific Stimuli:** The effectiveness of these datasets frequently depends on specific video clips or stimuli used to invoke emotions. These stimuli may not be universally effective across different populations or cultures, potentially limiting the applicability of the results [13].
- **Lack of Diversity:** Existing datasets often lack diversity in terms of age and gender [13], [14], and lack racial diversity [14], which can impact the generalizability and applicability of the models across varied populations.

1.3.3 LLM Reliability and Personalization

LLMs provide a promising interface for delivering mental well-being guidance through natural conversation. However, when used in isolation, they exhibit limitations that restrict their reliability in real-world mental health support applications. In this thesis, two challenges are particularly important: hallucination and insufficient personalization.

LLMs can generate responses that appear plausible but are factually incorrect or unsupported, a phenomenon commonly referred to as hallucination [15]. In a mental health context, this risk is especially concerning because the information shared with users can be sensitive, and inaccurate guidance may negatively affect users' well-being [16]. Therefore, reducing hallucination and improving grounding are critical requirements for building a trustworthy conversational support system.

In addition, personalization is essential for meaningful mental well-being support. A therapist-like system must consider the user's personal context and history to provide guidance aligned with

individual needs. Generic, context-agnostic responses may reduce perceived empathy and usefulness, and may limit the system's ability to support users over time. As a result, mechanisms that incorporate user-specific information and support context-aware responses are necessary for improving the usefulness and reliability of LLM-based mental well-being assistants.

1.4 Research Methodology

1.4.1 Overview

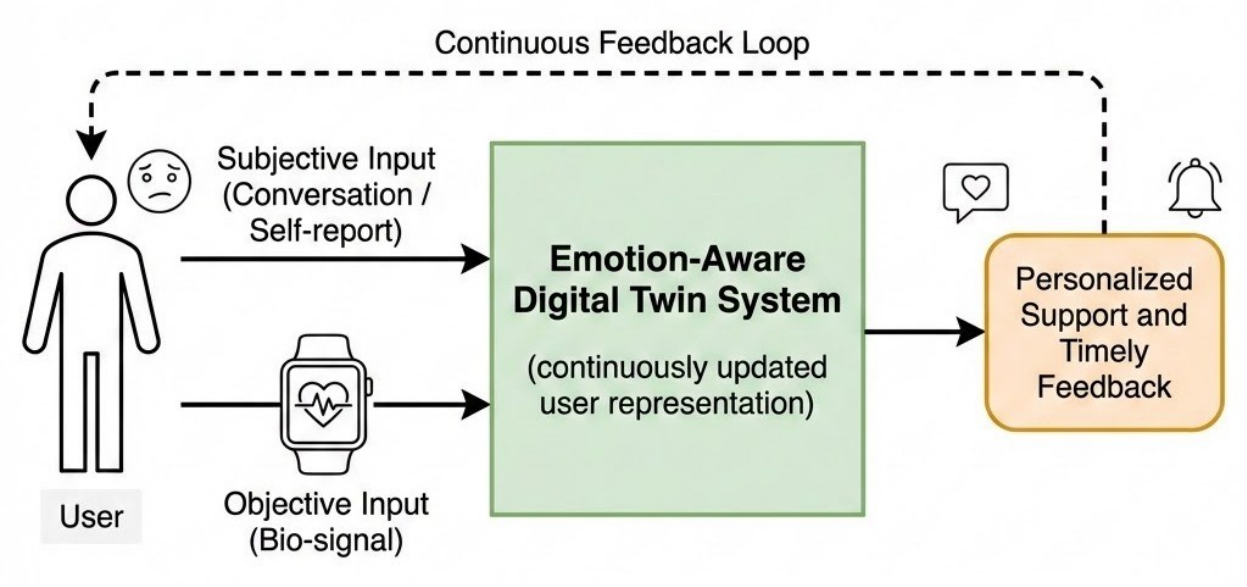


Figure 1.1. Overview of my solution

This thesis investigates an emotion-aware digital twin approach for providing accessible mental well-being support outside traditional clinical settings. The main idea is to combine two complementary sources of information: subjective input provided by the user through conversation or brief self-reports, and objective input derived from wearable physiological signals that reflect changes in affective state. These inputs are integrated within a continuously updated digital representation of the user, which maintains an evolving view of emotional history and personal context over time. Based on this evolving state, the system provides personalized support and timely feedback to improve self-awareness and support mental well-being. Figure 1.1 presents a high-level overview of the thesis concept and how the main components interact.

1.4.2 Proposed Approach

In the digital world, the digital twin can support self-monitoring, potentially enabling users to recognize early signs of changes in well-being. In this thesis, I adopt the definition in El Saddik et al. [17] that states that a digital twin refers to a continuously updated digital representation of an individual that supports monitoring and understanding of their state and enables personalized feedback to improve quality of life and well-being.

I address the lack of timely accessibility of mental well-being support mentioned in the problem statement by developing an emotion-aware WDT that (i) ubiquitously senses affect from wearable physiological signals, (ii) maintains a representation of the user's emotional state and personal context, and (iii) uses retrieval-grounded large language model responses to provide safe, personalized mental well-being support.

To enable this approach, the thesis first introduces WARM-VR, a new wearable-based affective dataset collected under a controlled yet realistic protocol that includes baseline, stress induction, and relaxation in immersive virtual reality. This dataset provides synchronized physiological signals and self-reported affect annotations to support reproducible emotion recognition models.

Building on this dataset, the thesis develops and evaluates wearable-based emotion recognition models aimed at improving robustness and generalization across users. Finally, these components are integrated into UbiMyTherapist, a digital-twin framework that maintains a representation of the user's emotional state and personal information, and uses RAG to ground LLM responses for more reliable and personalized support. A prototype of the UbiMyTherapist is implemented, and user case study is conducted, as a proof of concept to assess feasibility and user experience.

1.4.3 Objectives

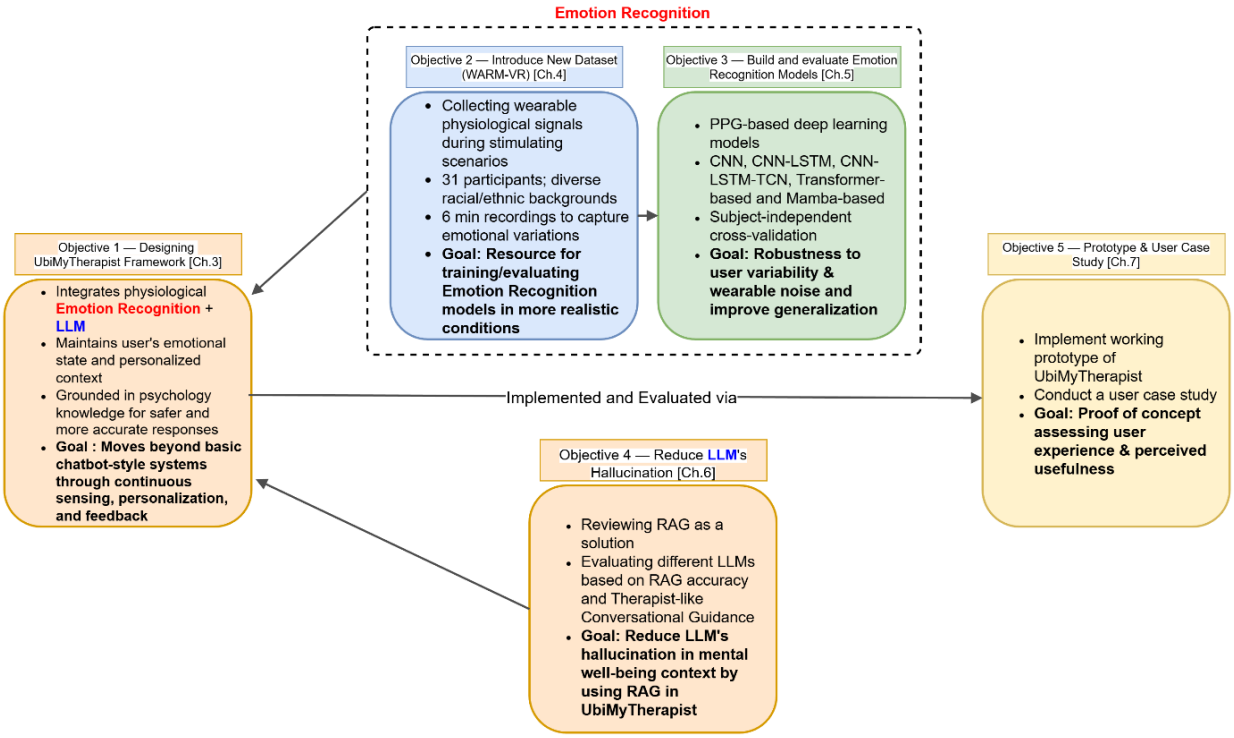


Figure 1.2. Overview of the Objectives

Figure 1.2 summarizes the thesis objectives and their relationship within the proposed UbiMyTherapist framework. The work begins by defining the system design requirements and identifying two core components that must be addressed: wearable-based emotion recognition and reliable generative AI support. To support emotion recognition, I first need a reliable dataset for building and evaluating emotion recognition models. In parallel, I investigate retrieval-augmented generation as a strategy to reduce hallucinations and improve the reliability of LLM-based mental well-being support. Finally, as a proof-of-concept, these components are implemented in a working prototype and evaluated through a user case study to assess feasibility, user experience, and perceived usefulness.

1.5 Contributions

1. The design and implementation of UbiMyTherapist: A framework under the umbrella of WDT. This system is a ubiquitous, emotion-aware digital twin framework for personalized mental well-being support. The framework integrates wearable-based emotion

recognition with retrieval-grounded large language model responses, maintains the user's emotional history and personalized context over time, and supports timely feedback aimed at improving self-awareness and well-being. As a proof of concept, the thesis also presents a prototype implementation and a user case study conducted with 24 participants to assess feasibility, user experience, and perceived usefulness.

2. The introduction of WARM-VR [18]: a wearable-based affective dataset collected during stimulating Virtual Reality (VR) scenarios. WARM-VR addresses key limitations in existing affect datasets by providing longer recordings and participants with diverse racial and ethnic backgrounds. The dataset is publicly available through IEEE DataPort [19], and it is intended to support reproducible training and evaluation of emotion recognition models under more realistic sensing conditions.
3. The development and evaluation of multiple deep learning architectures for PPG based wearable emotion recognition. This thesis benchmarks several model families, including CNN-based and sequence-based approaches, and extends them with hybrid and modern architectures, such as CNN–LSTM–TCN, Transformer-based models, and Mamba-based models. The models are evaluated using subject-independent cross-validation to assess robustness to user variability and wearable noise, with the goal of identifying modeling strategies that improve generalization and support real-world integration within emotion-aware systems.

1.6 Scholarly Achievements

In the process of completing this work, the following publications have been accepted or published:

1. K. Alghoul, H. Al Osman and A. El Saddik, "Enhancing Generalization in PPG-Based Emotion Measurement with a CNN-TCN-LSTM Model"– 2025 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Chemnitz, Germany, pp. 1-6, doi: 10.1109/I2MTC62753.2025.11079085. [20]
2. K. Alghoul, R. Sharma, H. Al Osman and A. El Saddik, "UbiMyTherapist: A Digital Twin MultiModal LLM-based System with Emotion Detection" – 2026 IEEE International Conference on Consumer Electronics (ICCE), doi: 10.1109/ICCE67443.2026.11449896. [21]

3. K. Alghoul, H. Al Osman and A. El Saddik, “PPG-Based Affect Recognition with Long-Range Deep Models: A Measurement-Driven Comparison of CNN, Transformer, and Mamba Architectures”– 2026 IEEE International Instrumentation and Measurement Technology Conference (I2MTC). [22]
4. Y. Valdivieso, M Faisal, K. Alghoul, M. Vahdati, K. Hamlabadi, H. Al Osman, A. El Saddik, “The Potential of Olfactory Stimuli in Stress Reduction through Virtual Reality” – 2025 IEEE Medical Measurements & Applications (MeMeA), Chania, Greece, pp. 1-6, doi: 10.1109/MeMeA65319.2025.11068102. [23]
5. H. Awad, K. Al Ghoul, H. Al Osman and N. Baddour, "Enhancing Nipple Positioning Accuracy in Chest Reconstruction Surgery: An Automated Machine Learning Approach," 2025 IEEE Medical Measurements & Applications (MeMeA), Chania, Greece, pp. 1-6, doi: 10.1109/MeMeA65319.2025.11068037. [24]

1.7 Thesis Outline

The remainder of this thesis is organized as follows.

Chapter 2 presents the background and related work on affective computing, physiological signals-based emotion recognition, affective datasets, LLM-based mental health assistants (including RAG systems), and digital twins for health and well-being, and concludes with a gap analysis that motivates the proposed approach.

Chapter 3 presents UbiMyTherapist, an emotion-aware digital twin framework that integrates wearable emotion recognition, use history, and RAG-based generation to support reactive and proactive assistance.

Chapter 4 introduces the WARM-VR dataset, including the study protocol, equipment used, ground truth, labeling strategy, and initial analyses.

Chapter 5 Introduce a new Hybrid deep learning model for PPG-based emotion recognition, and investigates multiple deep learning architectures. It also presents a structured evaluation of their performance in a subject-independent cross-validation setting.

Chapter 6 evaluates RAG-enabled LLM systems for mental health support through experiments that assess both context-grounded accuracy and therapist-like conversational guidance.

Chapter 7 describes the prototype implementation and the user case study conducted to evaluate feasibility and user experience.

Chapter 8 concludes the thesis by summarizing the main findings, discussing limitations, and outlining future research directions.

Chapter 2. Background and Related Work

2.1 Affective Computing

The development of affective computing has been marked by several key milestones, particularly in the integration of emotion recognition within HCI [25]. Early research and theoretical foundations began with the recognition of the importance of EI in human interactions, prompting researchers to explore how machines could recognize and respond to human emotions [8]. This laid the groundwork for affective computing. Advances in signal analysis for speech and facial expressions have been crucial, with techniques based on neural networks being extensively used to extract emotional features from speech and facial gestures. These methods have enabled a more nuanced understanding of emotional signals [26].

Machine learning has played a pivotal role in the advancement of affective computing. Algorithms have been developed to classify emotional states based on various inputs, such as speech patterns and facial expressions. The creation of comprehensive databases containing audiovisual material representing a wide range of emotional behaviors has been essential, as these databases facilitate the training of machine learning models, allowing them to learn from real-world emotional expressions rather than posed scenarios [14].

Continuous research into the physiological and neurological underpinnings of emotion has further informed the development of affective computing technologies. Understanding how emotions are processed in the brain can lead to more sophisticated models that mimic human emotional responses.

For instance, a machine capable of recognizing emotions can adapt its communication style and content based on the user's emotional state. This can be particularly beneficial in many applications and fields. For example, the AI-based education applications can benefit from this capability, where the machine can adjust its teaching style based on the student's level of understanding and frustration. Additionally, this could be valuable in healthcare fields, where machines could assist medical professionals in identifying patients who are in distress or require immediate attention [8].

The integration of wearable technology has significantly influenced the field of affective computing and machine learning by enabling the acquisition of physiological signals, such as Electroencephalogram (EEG), Electrocardiogram (ECG) and PPG, which are crucial for emotion

recognition [27]. These physiological signals provide more objective and reliable predictions of emotional states, as they are less susceptible to social masking compared to physical signals like facial expressions or voice tones [28][29][30]. Wearable devices facilitate the collection of real-time data, enabling the detection and recognition of subtle sentiments and complex emotions across various contexts. This capability enhances the robustness and accuracy of affective computing models, particularly in multimodal affective analysis, where physiological data can be combined with physical data (textual, audio, and visual) to improve overall performance.

Moreover, the use of wearable technology has led to the development of new algorithms and machine learning models that can process and analyze the rich data generated by these devices, thereby advancing affective computing and its applications in real-world scenarios, such as mental health monitoring and HCI [25].

2.1.1 Physiological Measurements

The most common modalities for Physiological Measurements-Based emotion recognition include EEG-measured brain activity signals and ECG [31].

Since the quality judgment process takes place in the brain and is influenced by a person's emotional characteristics, EEG signals are expected to provide more significant information about an individual's emotional state [32]. Several studies have proposed EEG-based models to enhance emotion detection [33][34][35]. While EEG is recognized as the most accurate and reliable physiological signal [36], it necessitates wearing an EEG head-mounted device, which can be inconvenient for the user.

Cardiac-related signals, such as ECG, are also closely associated with human emotions. ECG is commonly used as a unimodal signal for emotion recognition [37]. This is demonstrated by several studies that have achieved high accuracy in emotion classification using ECG [38] [39].

To date, PPG has been used less frequently as a unimodal signal for emotion recognition; however, Sayed Ismail et al. [40] suggest that PPG remains a viable alternative to ECG for developing emotion recognition systems. While PPG is commonly employed in multimodal emotion recognition systems (ERS) [40][41][42][43][44], its application as a standalone signal in ERS remains limited. Because PPG signals are highly sensitive to noise, researchers have traditionally favored ECG over PPG. However, recent studies have demonstrated that PPG can be used as the sole modality in ERS with results comparable to those obtained with ECG [37].

Compared with other physiological measurements, PPG offers several practical advantages that make it well-suited for real-world deployment. It can be acquired unobtrusively via compact, inexpensive wearables (e.g., smartwatches or smartphones), avoiding bulkier instrumentation such as ECG chest straps. Wrist-based PPG sensing is a non-invasive, user-friendly, and cost-effective method that enables continuous monitoring outside laboratory settings [45].

2.1.2 Feature Extraction: Traditional vs Machine Learning based Methods

In emotion recognition pipelines, feature extraction determines what information is provided to the learning model. There are two primary approaches for feature extraction from physiological measurements.

2.1.2.1 Hand-crafted Feature Extraction

Traditional physiological emotion recognition systems typically rely on domain-informed, hand-crafted features extracted from each modality. For example, EEG-based approaches often use spectral band power and asymmetry measures [46][47]. Wang et al. conducted a comparative study of EEG feature sets for emotion classification tasks. Jenke et al. provided a comprehensive review of EEG-based feature extraction techniques and identified features that are particularly informative for emotion recognition k[48].

In contrast, for cardiac-related signals (ECG/PPG), the most widely used handcrafted feature family is derived from inter-beat intervals. For ECG, inter-beat intervals are extracted between consecutive R-peaks (RR intervals), and for PPG they are extracted between consecutive pulse peaks (often referred to as pulse-to-pulse intervals). In both cases, the resulting interval series is used to compute Heart Rate Variability (HRV) features. HRV feature extraction is commonly grouped into three analysis domains: time-domain, frequency-domain, and nonlinear-domain [49].

a. **Time Domain:** These features reflect the overall variability of heartbeats using statistical methods [50], including the standard deviation of IBIs (SDNN), the square root of the mean squared differences between adjacent IBIs (RMSSD), the standard deviation of differences between adjacent IBIs (SDSD) and the count or percentage of successive beat lengths that differ by more than 50 ms (NN50, pNN50).
b. **Frequency Domain:** Spectrum estimation is performed on the IBI series using fast Fourier transform and Welch's periodogram techniques. The spectrum is divided into

very low frequency (VLF, 0-0.04 Hz), low frequency (LF, 0.04-0.15 Hz), and high frequency (HF, 0.15-0.4 Hz) bands. Since HRV features were calculated over a 2-minute signal, VLF was excluded due to inaccurate values. Total power (TF) and the LF/HF ratio were also calculated, with LF and HF represented in normalized units (nLF, nHF) to better capture the sympathovagal balance.

c. **Nonlinear Domain:** Time-delay embedding methods are employed to capture the nonlinear properties of the PPI time series. These methods estimate the complexity of the time series and its relationship with mental states. Nonlinear measurements include Poincaré plots [51] and correlation dimensions [52].

Despite their utility and interpretability, hand-crafted feature methods often suffer from practical limitations, including longer recognition windows and comparatively lower accuracy.[ ]

2.1.2.2 Machine Learning-based Representation Learning (Automatic Feature Extraction)

An alternative to handcrafted features is to use deep learning models that learn representations directly from the signal. In this approach, feature extraction is performed implicitly by using various neural network models [53][54]. As deep learning became more widely adopted, many studies began using deep neural frameworks as classifiers. For example, Jirayucharoensak et al. proposed a deep learning model to identify feature correlations [55]. With continued advances in deep learning, these methods have also demonstrated strong performance in automatic feature extraction and have been applied to learn representations directly from biosignals. Martinez et al. developed physiological modeling approaches that extract features using a convolutional neural network (CNN) and an auto-encoder [56]. Alhagry et al. proposed a Long Short-Term Memory (LSTM) model that learns EEG features and performs emotion classification based on arousal and valence values [57]. Yang et al. introduced a parallel architecture combining a Recurrent Neural Network (RNN) and a Convolutional Neural Network (CNN) on EEG signals to obtain meaningful features [58]. Overall, deep learning has enabled emotion recognition within short time windows, including one-minute segments.

When it comes to PPG for example, many studies have used deep learning to extract features automatically [45] [59].

Deep learning facilitates representation learning, which can fully or partially eliminate the need for manual feature engineering. This approach allows for discovering an effective set of features for a task in significantly less time compared to manual methods [6].

2.1.3 Emotion Assessment

Many studies have been conducted on estimating emotion using machine learning applied to physiological signals. In this subsection, I focus on PPG signals, as they are the most directly aligned with the wearable sensing scope of this thesis.

Recently, there has been an increase in research focused on estimating emotion using Machine learning on PPG signals alone. For instance, Lee M et al. [45], explored the use of PPG signals for short-term emotion recognition. They proposed a method based on a single pulse PPG signal, employing a one-dimensional convolutional neural network (1D CNN) to extract features and classify emotions, with validation using the DEAP dataset [14]. Other studies have used the WESAD dataset [12], including [60], in which the authors proposed a novel Hybrid Convolutional Neural Network (H-CNN) architecture for stress detection using wrist-based PPG sensor data. This H-CNN architecture combines hand-crafted features from the PPG signal with features automatically learned by the CNN to enhance stress classification accuracy. In 2022, Jin Y et al. [13] introduced a new dataset for emotional analysis based on PPG signals extracted from a fingertip sensor, and evaluated the dataset with various machine learning models.

Beyond CNNs, LSTM networks and hybrid CNN-LSTM architectures have also been explored for emotion recognition. For instance, Etienne et al. [61] developed a CNN-LSTM model for emotion recognition from speech data, while [62] demonstrated the application of LSTMs in physiological signal processing. Pre-trained CNNs combined with residual LSTMs have been applied to EEG signals for emotion detection [63]. Regarding PPG signals, CNN-LSTM architectures have been effectively used, as shown in [64], which employed a hybrid model combining 1D-CNN and LSTM for emotion recognition tasks using remote PPG signals.

Temporal Convolutional Networks (TCNs) have emerged as another promising architecture with unique characteristics, including maintaining the same output size as the input and ensuring no future information leaks into the network's computations [65]. Harb [66] highlighted the feasibility of using TCNs for emotion recognition and proposed a multi-modal TCN model integrating features

from audio, visual, and textual modalities. TCNs have also been applied successfully to emotion recognition tasks involving EEG and ECG signals [67] [68] [69].

For PPG signals, while TCNs have been used in heart rate monitoring [70] and to classify learner engagement levels in E-learning contexts [71], their application for emotion recognition remains unexplored. To the best of my knowledge, no prior work has utilized TCNs in combination with PPG for emotion recognition.

2.2 Emotional Model

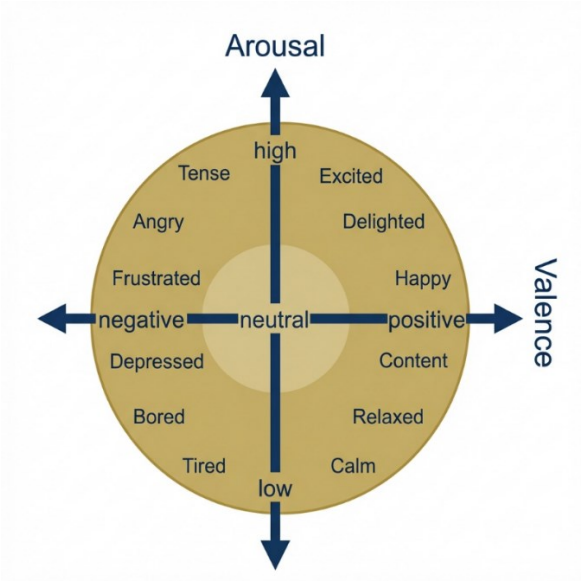


Figure 2.1. Dimensional Emotional Model

Emotions are commonly represented using two dominant approaches [72]. The categorical model assigns discrete labels to emotions, such as stress or joy. Despite its simplicity, this model has limitations, such as lacking exact translations and conceptual equivalence across languages [73]. Additionally, the categorical approach may fail to capture the full spectrum of emotions due to its reliance on a limited set of categories. Alternatively, emotions can be represented using an n-dimensional model [74]. A widely used dimensional approach includes the valence and arousal dimensions [75]. Valence describes the degree of pleasantness of an affective state, ranging from unpleasant to pleasant, whereas arousal describes the level of activation or intensity, ranging from low activation (calm) to high activation (excited).

In the valence-arousal model, discrete emotions can be positioned within four quadrants, which can be used to associate regions of the space with representative affective states as seen in Figure 2.1.

For example, high valence combined with high arousal is commonly associated with states such as excitement or happiness, while low valence combined with high arousal is often associated with states such as anxiety or anger.

From a discrete emotions perspective, each emotion, such as anger, sadness, or contempt, corresponds to a unique profile across experience, physiology, and behavior. According to [72], emotions are not simply varying degrees of valence and arousal but rather distinct categories with their own unique characteristics.

This representation is especially common in affective computing because it supports continuous modeling of affect intensity and accommodates mixed or ambiguous affective states that are not easily captured by a small set of discrete labels

2.3 Datasets for Affective Computing

2.3.1 Popular Datasets and their Limitations

Affective computing research has long emphasized the value of physiological signals for understanding affective responses to multimedia stimuli [76]. Numerous publicly available datasets have been proposed to support this goal, often incorporating modalities such as EEG, Electromyography (EMG), ECG, PPG, Electrodermal Activity (EDA), three-axis Acceleration (ACC) and skin Temperature (TEMP).

However, most widely used datasets were collected in controlled lab conditions and rely primarily on non-immersive visual stimuli and do not consider the growing relevance of immersive VR and multisensory environments. My work builds on this trajectory by introducing a multimodal dataset based on wearable sensing within immersive VR environments, enhanced with olfactory stimulation.

Table 2.1 summarizes the main affective datasets discussed in this section and highlights differences in sensing modality, wearability, stimuli type, and the use of AR/VR. Among popular datasets, DEAP [14] is a publicly available multimodal database. Despite its broad modality coverage and self-reports, the sensors used are not considered wearables.

Table 2.1. Existing Related Datasets

Year	Name	Signal Source	Use Wearables Sensors	Number of Subjects	Type of Stimuli	Use of AR/VR	Affects
2012	DEAP	EEG, EMG, EOG, GSR, PPG	No	32	Video	No	Arousal, Valence, Liking, Dominance
2018	WESAD	ECG, EMG, RSP, PPG, EDA, ACC, TEMP	Yes (ECG belt and PPG smartwatch)	15	Video, Assessment	No	Neutral, stress, amusement
	DEAR-MULSEMEDIA	EEG, GSR, PPG	No	18	Video, Fan, Heater, Olfaction Dispenser, and Haptic vest	No	Arousal, Valence
2022	PPGE	PPG	No	18	Video	No	Joy, sadness, anger, Relaxation
2022	OVPD	EEG	No	10	Video and Olfaction dispenser	No	Arousal, Valence
2022	BOOKAR	EEG	No	22	Reading, Augmented-Reality (AR) reading	Yes	Arousal, Valence, Dominance
2023	VREMO	EEG	No	32	VR scene	Yes	Cybersickness
2025	**WARM-VR**	**ECG, PPG, EDA, ACC, TEMP**	**Yes (ECG belt and PPG smartwatch)**	**31**	**Reading, Assessment, Olfaction dispenser, VR scene**	**Yes**	**Arousal, Valence, Relaxation**

It also lacks immersion and uses screen-based stimuli only. In contrast, WESAD [12] employs wearable sensors through a chest belt and a wristband to record ECG, PPG, EDA, TEMP, EMG, Respiration (RESP), and ACC signals to measure affect-related physiological responses. While WESAD is highly relevant to my focus on wearables, it is limited to visual and sound stimuli and does not consider immersive environments.

Several datasets have explored the integration of olfactory stimuli into affective elicitation. The OVPD dataset [77] for example, combines video and odor cues to enhance EEG-based affect classification, showing that multimodal sensory input can improve signal discriminability. Similarly, the DEAR-MULSEMEDIA dataset [78] includes EEG, Galvanic Skin Response (GSR), and PPG signals collected under visual, auditory, olfactory, haptic, and thermal stimuli. This dataset is notable for its multisensory scope, yet it lacks a VR-based or fully immersive delivery context. The influence of PPG, also referred to as PPG, on affect analysis has also been explored in isolation. For example, PPGE dataset introduced by Jin et al. [13] focuses exclusively on PPG-based affect classification using narrative videos and provides benchmark results for a deep learning model. Cognitive AR-based experiments have also been used to investigate affective engagement. Dasdemir introduced BOOKAR [79], a dataset using EEG data collected while participants read texts Augmented Reality (AR). Although this dataset moves toward immersive interaction, it does not include full VR immersion, wearables beyond EEG, or multisensory integration.

Immersive VR offers a compelling way to simulate real-world environments and scenarios, allowing users to feel a strong sense of presence and interact with the environment. By replacing the physical environment with a virtual one, VR creates the illusion of “being there” through sensory engagement. This sense of immersion is largely driven by the stimulation of various senses. By engaging multiple senses, VR can closely replicate real-life interactions. Although often associated with gaming, recent studies have highlighted VR’s potential in therapeutic contexts, including its ability to support relaxation and alleviate anxiety [80].

In 2023, Dasdemir introduced VREMO [81] dataset, which includes EEG signals in response to VR scenes to classify both cybersickness and emotional states within the valence–arousal space. This work shows the potential benefit of physiological measurements in immersive contexts. However, it relies solely on EEG and lacks common daily wearables like ECG belts and PPG wristbands.

I describe below the three datasets that are most closely aligned with my work.

Database for Emotion Analysis using Physiological signals (DEAP) Dataset

The DEAP [14] dataset is widely used in emotion recognition research. It consists of two parts: online ratings and participant ratings. The online self-assessment includes ratings of 120 1-minute music videos by 14-16 volunteers, focusing on arousal, valence, dominance, and liking of the videos. It is designed to support the analysis of human affective states. It was created to facilitate research in affect recognition by providing physiological and multimedia data collected while participants experienced emotional stimuli. The focus of the DEAP dataset was the EEG signals, but it also included other modalities such as PPG, which was collected using a pulse sensor attached to a fingertip.

The participant ratings involve 32 participants (16 male and 16 female) who watched a subset of 40 videos while recording various physiological signals. These participants also rate the videos on a scale of 1 to 9 for arousal, valence, dominance, and liking.

PPGE Dataset

This dataset [13] consists of 72 PPG recordings collected from 18 participants (13 men and 5 women) aged between 23 and 31. The dataset was collected while participants watched short video clips designed to evoke specific emotions. These emotions were: joy, sadness, anger, and relaxation.

The study emphasizes that the video clips were selected from YouTube, had durations of less than 5 minutes, and were sufficiently impactful to elicit emotional responses in participants. The dataset was collected using a pulse sensor attached to a fingertip, with a sampling rate of 100 Hz.

The paper highlights the limitations of the DEAP dataset in terms of the short duration of the video stimuli, stating that 1 minute may not be enough to evoke an emotional response. It also emphasizes the importance of narrative video clips in the PPGE dataset, suggesting that they may be more effective in eliciting emotions.

Wearable Stress and Affect Detection (WESAD) dataset

WESAD [12] is a multimodal dataset for wearable stress and affect detection. WESAD features physiological and motion data recorded from both a wrist-worn and a chest-worn device from 15 subjects during a lab study. The dataset includes the following sensor modalities:

- Wrist-worn device (Empatica E4): PPG, EDA, TEMP, and ACC.

- Chest-worn device (RespiBAN): ECG, EDA, EMG, RSP, TEMP, and ACC.

The dataset bridges the gap between prior laboratory studies on stress and emotion by including three affective states: neutral, stress, and amusement. The subjects were exposed to different affective stimuli (stress and amusement), and a baseline and two meditation periods were recorded.

- Baseline condition: This condition is aimed at inducing a neutral affective state. Participants were instructed to sit at a table and read neutral reading material for 20 minutes.
- Amusement condition: Participants watched a set of eleven funny video clips. Each clip was followed by a short neutral sequence of five seconds.
- Stress condition: Participants were exposed to the Trier Social Stress Test (TSST), which consists of a public speaking and a mental arithmetic task. These tasks are known to elicit stress reliably.

The TSST consisted of a five-minute speech on personal traits, followed by counting backward from 2023 to zero in steps of 17. Subjects were told they were being evaluated by human resource specialists, which added pressure to the situation.

The dataset also includes self-reported values on the perceived affective state of the subjects, which were obtained using several established questionnaires. These self-reports can be used to train personalized classifiers.

The advantage of WESAD over the DEAP dataset is its Real-world applicability due to the use of wearable sensors. But DEAP's use of valence and arousal modality allows the representation of blended emotions and is not limited to specific emotions. Additionally, DEAP has more than twice as many participants.

2.3.2 The Need for a New Dataset

Despite these advances, existing affect recognition datasets remain limited in several important aspects noted above. Most lack full immersion due to the absence of VR integration, even when multiple sensory stimuli are included. Additionally, physiological data are often collected using stationary laboratory equipment rather than wearable devices, which undermines real-world applications.

As shown in Table 2.1, only a small subset of datasets combine wearable sensing with immersive/VR stimuli. To address this limitation, I introduce WARM-VR (Wearable Affect Recognition from

Multisensory stimuli in Virtual Reality), a novel publicly available multimodal dataset designed to support affect recognition in immersive, multisensory environments using wearable sensing instrumentation.

2.4 LLMs in Mental Health

2.4.1 Existing solutions and their limitations

Mental illness is a significant health issue that impacts emotions, reasoning, and social interactions, often with far-reaching consequences for individuals and communities [82]. Effective mental health support calls for approaches that emphasize prevention and early detection.

Traditional diagnostics often rely on self-reported data collected through questionnaires that identify specific emotional or social patterns, thereby helping to detect individuals in need of support [83] [84]. Advances in predictive analytics and medical technology, as highlighted by recent studies, are reshaping this field, paving the way for more responsive and proactive healthcare solutions [84] [85] [10].

The rapid rise of mental health applications has underscored the need for support mechanisms that adapt to individual needs [10]. With a growing demand for accessible mental health tools, artificial intelligence (AI) has emerged as a promising avenue for providing customized guidance and resources to diverse users [82] [10]. These AI-driven approaches offer significant potential for addressing unique mental health needs at scale, ensuring that users have access to support in ways that traditional models often cannot [10] [82] [83] [86].

In recent years, advancements in Natural Language Processing (NLP) have opened up new possibilities for mental health applications, with LLMs emerging as powerful tools for enhancing user interactions across various domains, including healthcare [83]. LLM-based question-answering (QA) systems have shown promise in providing relevant and timely responses [15]. However, despite their potential, general QA systems face notable limitations, particularly around field-specific language and rapidly evolving knowledge bases [87]. Without real-time data access and specialized terminology, these systems often struggle to retrieve and contextualize information accurately [85]. This challenge is especially noticeable in complex fields like mental health [84] [88].

2.4.2 Retrieval-Augmented Generation (RAG) systems

2.4.2.1 General Advances In Rag Systems

I begin by reviewing general advances in RAG systems, covering their core components: embedding models, retrieval mechanisms, and generation processes.

1) Embedding Challenges: Data Gaps and Specificity

A persistent challenge is managing data gaps, which can lead to hallucinations or incomplete reasoning when crucial contextual information is missing. One study highlights this issue, noting that when key attributes such as demographic details are absent, models may inaccurately infer missing context, thereby reducing factual reliability [85]. For instance, retrieval-augmented systems may generate plausible but incorrect outputs, due to over-reliance on general training data and filling gaps with assumed information that may not align with the user's actual context [15]. Several studies highlight the complexity of developing embeddings that generalize across diverse cases while maintaining context-specific precision [89]. Additionally, embedding models like Bidirectional Encoder Representations from Transformers (BERT) and Sentence Transformers face challenges with complex document layout, such as multiple columns or embedded images, which complicate the process of text extraction and disrupt retrieval accuracy [90]. These findings underscore the need for advanced preprocessing techniques to ensure that extracted chunks remain coherent and relevant for similarity searches.

2) Retrieval Systems

Retrieval systems in RAG applications must effectively address data relevance and retrieval accuracy to ensure that only relevant information influences the generation process. They filter large volumes of data and prioritize information that is both contextually relevant and responsive to user specific needs. A primary technical challenge is token length limitations, which restrict the amount of text that models can process at once. To overcome this, retrieval systems segment lengthy clinical notes into manageable chunks, preserving critical information without overwhelming the model [85]. Other works emphasize real-time data matching [90], retrieval-aware fine-tuning [15], and query rewriting [89] as strategies to improve retrieval relevance and reduce hallucinations.

3) Generation: Prompts and Context Engineering

In RAG systems, the generation stage focuses on ensuring the accuracy and appropriateness of content produced by LLMs, with prompt sensitivity being a major challenge [85]. Retrieval-aware fine-tuning also emerges as a key method to ground outputs in retrieved content, reducing hallucinations and improving contextual relevance [15] [89].

2.4.2.2 Rag Systems In Healthcare

Building on these general insights mentioned previously, several studies have specifically applied RAG systems to healthcare contexts, where data complexity, accuracy, and scope control are especially important.

In health applications, embedding data from diverse sources such as Reddit or electronic medical records introduces significant challenges [91]. Studies highlight the complexities of creating embeddings that need to generalize across diverse cases while maintaining patient-specific accuracy [92], and emphasize the importance of semantic representation [89]. Embedding informal, unstructured health data requires preserving subtle relational cues to support accurate personalization.

For example, Yu et al. [92] introduced Health-LLM, a retrieval-augmented system for disease prediction. Their framework integrates patient-specific data with external medical knowledge bases to generate accurate, tailored predictions, demonstrating how retrieval grounding can enhance clinical decision support. Similarly, AlKhalaf et al. [85] explored the use of RAG for summarizing and extracting key information from electronic health records (EHRs). Their study showed that generative AI, when combined with retrieval, could produce concise, clinically relevant summaries that improve interpretability for healthcare professionals. This highlights the role of RAG in addressing the complexity of large, unstructured datasets such as EHRs.

Another review [91] proposed a two-layer modular RAG framework to support medical QA from social media data, combining query-focused summarization modules with LLMs to answer clinicians' questions about emerging drug use based on Reddit posts. Similarly, retrieval-aware finetuning has been shown to improve grounding in clinical contexts [15].

2.4.2.3 Rag Systems In Mental Health

Within the healthcare domain, mental health is particularly demanding due to its reliance on nuanced, context-sensitive, and personalized language. Recent research has begun to explore how RAG systems can specifically address these challenges

1) Text Mining and Preprocessing

Abbe et al. [87] reviewed text mining applications in psychiatry, showing how tokenization, lemmatization, and semantic structuring are essential to extract insights from unstructured records, patient narratives, and biomedical literature. The review identified four application areas: psychopathology, patient perspective, medical records, and medical literature. These findings reinforce the importance of preprocessing pipelines in preparing psychiatric datasets for retrieval augmented frameworks, ensuring that subtle patient cues and clinical nuances are preserved.

2) Domain-Specific Challenges in Mental Health RAG

Two recent studies propose RAG frameworks tailored for mental health information retrieval. First study [16] introduces a RAG-based solution to provide accurate mental health information that can be used by policymakers in designing interventions. It integrates Cochrane Reviews as a curated knowledge base, ensuring responses are grounded in high-quality, evidence-based literature. Their system utilizes Facebook AI Similarity Search (FAISS) indexing with BAAI General Embedding (BGE) and a quantized Mistral LLM, demonstrating the feasibility of scalable RAG for enhancing mental health literacy and supporting digital mental health interventions. The study did not use a quantitative measure of RAG accuracy; instead, the authors themself did a qualitative analysis of the system's performance. While promising, they acknowledge limitations in scope (restricted to Cochrane mental health reviews) and the need for validation against expert mental health professionals.

In the second study, the authors introduce SentimentCareBot [93], a chatbot designed to improve mental healthcare accessibility by integrating sentiment analysis with a RAG-based LLM. The study evaluates various RAG configurations (Naive RAG, Multi-query RAG, and Hypothetical Document Embeddings – HyDE), with and without a sentiment-based re-ranking component, using a public mental health counseling dataset. The key finding is that sentiment-aware reranking, particularly when applied to Multi-query RAG with the Mistral language model, improves answer quality according to automatic evaluation metrics, highlighting the potential of sentiment signals to enhance mental health chatbots.

Despite these advances, existing mental-health RAG systems primarily focus on information retrieval quality and sentiment-aware ranking; they do not benchmark multiple LLMs within a fixed

therapeutic knowledge base or examine how factual grounding relates to therapist-like conversational behavior.

2.4.2.4 Llms As Virtual Therapists

Recent studies have explored the use of LLMs as virtual therapists in psychotherapy applications [94]. These systems harness LLMs' ability to maintain context and perform multiturn reasoning, enabling more flexible and human-like interactions than traditional scripted systems. Research in this area focuses on developing models that can adapt conversational strategies in response to real-time patient feedback. Comparative studies by Iftikhar et al. (2024) [95] and Zhang et al. (2025) [96] have explored the differences between LLM-led and human-led cognitive behavioral therapy (CBT) sessions. Their results emphasize significant gaps in empathy and cultural understanding, with LLM-based conversations frequently lacking the depth found in interactions with human therapists. A recent review [94] of 69 studies in this field reported that the majority of existing LLM-based therapy systems, about 77%, utilize prompt-based techniques, and 74% rely on commercial LLMs. However, a key limitation identified is that most current LLM-based therapy applications depend on static models, which are unable to dynamically adjust responses to users' evolving needs. The authors of this review suggest RAG frameworks could provide the necessary technical foundation for real-time, adaptive strategies in therapeutic conversations. Specifically, RAG can interpret patient cues during ongoing dialogue, adapt interventions based on subtle emotional and cognitive changes, and go beyond simply reproducing responses from pre-existing datasets.

Building on these insights, I use RAG as a main component in my framework for mental health support.

2.5 Digital Twins in Health and Well-being

2.5.1 Definition and Applications

Digital twins are virtual representations of physical entities, originally developed to enhance manufacturing processes. They enable continuous data transmission between the physical and virtual worlds, allowing for the monitoring, understanding, and optimization of the functions of both living and nonliving entities. Digital twins provide continuous feedback to improve quality of life and well-being and are increasingly applied across fields such as health, wellness, security, transport, and communications. The concept has evolved to encompass a wide range of applications beyond

manufacturing, with the potential to have significant impacts across multiple industries [97]. This technology enables the monitoring, understanding, and optimization of the functions and behaviors of the real twin, providing continuous insights to improve quality of life and well-being [98].

When it comes to healthcare, digital twins are not only dashboards for tracking signals. In most health and well-being settings, the digital twin is expected to include an intelligence layer that can analyze data, support prediction, and generate recommendations within a feedback loop. In other words, the value of the digital twin comes from connecting continuous data collection with decision-making and personalized feedback over time, rather than only visualizing measurements.

As mentioned in Chapter 1, WDTs are usually created by integrating continuous wearable sensors, personal records, and an AI-powered feedback system to facilitate ongoing monitoring and personalized recommendations over time. Recent work on WDTs, often frames them as a pipeline built from four main parts: (1) heterogeneous data sources (e.g., wearable sensors and personal records), (2) an AI inferencing engine that performs analysis and prediction, (3) a user-facing interaction layer for delivering feedback, and (4) a closed feedback loop where the twin is continuously updated based on new data and user responses [11].

WDT technology serves a broad range of applications aimed at enhancing accessibility, personalization, and effectiveness of health services. These applications can be grouped into three main categories:

- **Clinical care and precision medicine:** Can support telemedicine and smart healthcare services, and telemedicine, improving the efficiency and effectiveness of medical care [99]. They can also enable personalized medicine by tailoring interventions to individual trajectories and supporting precision health decision-making [98], [10].
- **Continuous monitoring and preventive care**: Enable continuous tracking of health parameters and support real-time data collection and visualization [98]. Building on this continuous monitoring, they can provide biofeedback and personalized recommendations that promote healthier behaviors and help prevent disease or deterioration [99]. In mental health services, it enhances accessibility by providing continuous monitoring and feedback through wearables and Internet of Things devices, offering personalized recommendations for mental health improvement based on specific symptoms, and representing an individual's mental state to support precision care and insight into mental well-being [10].

- **Lifestyle:** In sports and fitness, for example, it can assist athletes with automated training programs and personalized feedback, optimizing training without a coach and monitoring physical activities to provide insights into physical health and activity levels [99]. These diverse applications leverage the data-driven nature of digital twins and their integration with artificial intelligence and machine learning to optimize health outcomes and enhance overall well-being.

2.5.2 Digital Twins in Mental Health

The integration of AI into mental health care can potentially transform the psychotherapy field due to its ability to enhance accessibility and continuous patient monitoring [100] [101]. One of AI's key strengths is its ability to analyze patient data, such as behavioral patterns, responses to treatment, and medical history [102], increasingly collected through consumer electronics such as wearables, and remote monitoring tools [103]. This is achieved through tools that utilize NLP, speech affect analysis, and other affective computing techniques. Leveraging these tools enables therapy sessions to be more personalized and tailored to each patient's unique needs and circumstances [104]. Some of these applications offer AI-powered virtual therapists [105] [106], which may can potentially achieve a safe and private space for people to find support and help reduce barriers that hinder individuals from seeking professional help [107] [108]. However, most such systems remain limited to user-initiated therapeutic interactions within the framework of conversational health agents [108] [109] [110]. These AI systems primarily respond to verbal or written communication but do not proactively engage in delivering mental health support, making them reactive rather than proactive. Moreover, the existing systems neglect real-time and continuous data analysis, which poses a significant drawback, particularly when immediate intervention is crucial. While chat-based therapy applications remain reactive, some AI-driven systems take a more proactive approach by incorporating physiological measurements for stress monitoring. For example, commercial applications such as EliteHRV [111] and CuesHub [112] focus on enhancing personal well-being by monitoring and analyzing data from wearables, proactively sending stress alerts, and suggesting relaxation techniques when high stress is detected. In a recent study exploring the integration of wearable devices with LLM for stress management, Neupane et al. [113] integrated CuesHub with LLM chatbots, offering proactive prompts, but their focus remained on stress reduction rather than structured therapy. Although these applications offer some aspects of proactive interventions, their primary focus is on stress reduction rather than delivering structured therapy sessions.

2.5.3 Ubiquitous Biofeedback

The Ubiquitous Biofeedback Multimedia Systems concept, first introduced by Al Osman et al. in 2014 [114], provides a foundation for proactive intervention through physiological measurements. A key part of these systems guides users through personalized relaxation practices by assessing their physiological measurements and adjusting interventions based on past effectiveness. More recently, the concept of digital twin has emerged as a way to utilize EI to enhance the quality of life and well-being [9]. In the context of consumer health, digital twins are particularly effective when coupled with wearable devices, which generate continuous personal data. They provide the enabling technology that allows for continuous monitoring, deeper understanding, and timely feedback to improve well-being, and hold significant potential for application in the mental health field [99].

An example of an emotion-aware digital twin system that uses ubiquitous biofeedback is inHarmony [115], which combines wearable sensing with live biofeedback to depict users' emotional states and deliver feedback and coping recommendations based on user preferences and past effectiveness. This line of work, together with advances in affective computing, conversational mental-health agents, and digital twins, shows strong progress toward technology-supported well-being. However, the literature is still fragmented. Many systems advance only one component (e.g., sensing and biofeedback, or dialogue systems, or digital twin concepts) without integrating robust wearable emotion estimation, grounded language generation, and an updateable personalization loop into a single end-to-end solution.

2.6 Gap Analysis: Toward Emotion-Aware Digital Twins for Personalized, RAG-Based Mental Health Support

Despite rapid progress in affective computing, wearable sensing, LLM-based mental health support, and digital twins, these areas are largely studied in isolation. To the best of my knowledge, there is still no comprehensive solution that integrates emotion sensing from wearable devices, reliable language generation, and digital twin personalization into a single system. This gap motivates the direction of this thesis and, in particular, the design of UbiMyTherapist, an emotion-aware well-being digital twin introduced in Chapter 3.

2.6.1 Gap 1: Limitations in existing affective datasets

Many datasets used for emotion recognition are collected in controlled laboratory environments with limited variability in stimuli, context, and user behavior. While useful for benchmarking, they do not fully represent everyday situations where stress and emotion naturally occur. In addition, immersive VR-based affect datasets combined with wearable sensors remain rare, even though such environments can more effectively simulate real-life emotional experiences. This limits the development of models that generalize well in practical settings.

2.6.2 Gap 2: Wearable-based emotion recognition models still struggle with robustness, generalization, and label uncertainty

PPG-based emotion recognition has shown promising results, especially with machine learning and deep learning methods. However, PPG signals are sensitive to motion artifacts, sensor placement, and individual differences. As a result, models often perform well within the same dataset but struggle when tested on new subjects or different conditions. There is a need for models that handle noise, imbalance, and subject variability more effectively, particularly when intended to support daily mental health applications.

2.6.3 Gap 3: LLM-based mental health assistants face reliability and personalization limitations that are not fully resolved by current approaches

LLMs can simulate conversations and provide general guidance, but they may hallucinate information, give non-professional advice, or respond without understanding the user's history or emotional state. RAG helps reduce hallucination by grounding responses in verified knowledge, but there are still no widely accepted protocols for mental-health scenarios. In particular, most works do not include emotional context from wearables or track user changes over time.

2.6.4 Gap 4: Digital twin research in health emphasizes physical states, while emotion-aware digital twins remain underdeveloped

Digital twins have been successfully applied in cardiovascular monitoring, fitness tracking, and disease management. However, emotion-aware digital twins for mental well-being are still uncommon. Existing work typically describes high-level architectures but does not address how

affect can be reliably sensed, how emotional states are stored and updated, or how feedback can be personalized for each individual. There is also limited evaluation on user experience or effectiveness in real use cases. In addition, emotion-aware digital twins require integrating heterogeneous data sources into an updateable state representation; otherwise, the "twin" becomes a collection of disconnected signals rather than a meaningful model of the user over time.

2.6.5 Gap 5: Lack of an end-to-end approach that combines all components

The four gaps above reinforce one another. Without valid datasets, the robustness claims of emotion recognition models remain fragile. Without robust emotion recognition, digital twin state representation will be unreliable. Without grounded and evaluable RAG mechanisms, LLM guidance remains difficult to trust.

2.6.6 Thesis approach to bridging the gap

This thesis addresses the above gaps through an end-to-end research strategy that links five components into a coherent WDT pipeline. The overall direction is guided by the UbiMyTherapist framework introduced in Chapter 3, and the subsequent chapters instantiate and evaluate the key components required to realize it.

1. **UbiMyTherapist (Chapter 3):** a framework under the umbrella of WDT, that integrates (i) continuous emotional state, (ii) user-specific information history and psychological knowledge bases, and (iii) RAG-enabled generation to support both reactive conversational assistance and proactive biofeedback-style interventions. This framework defines the target system and clarifies how the remaining components fit together.
2. **WARM-VR dataset (Chapter 4):** a data collection protocol that combines immersive stimuli with wearable physiological sensing and structured self-report to support affect recognition research in more realistic conditions.
3. **Wearable emotion recognition models (Chapter 5):** a structured evaluation of deep learning architectures for emotion recognition from wearable signals (with attention to noise, subject variability, and class imbalance) to identify modeling strategies that better support generalization.
4. **RAG-LLM evaluation for mental health support (Chapter 6):** a balanced approach between generating context-grounded answers while keeping the language therapist-like, to be considered effective for a mental-health assistant.

5. **Prototype and user case study (Chapter 7)**: Building a prototype of UbiMyTherapist and conducting a user case study to evaluate the feasibility and user experience, as a proof of concept.

By explicitly connecting dataset design, emotion recognition model development, context-grounded language generation, and a digital twin personalization loop, the thesis moves beyond isolated contributions and targets a unified solution to the central problem: enabling accessible, context-aware, and personalized support for mental well-being at any time and in any place.

For the remainder of this thesis, the term digital twin refers specifically to a WDT as defined in this chapter, unless stated otherwise.

Chapter 3. UbiMyTherapist Digital-twin Framework

Building upon the two concepts discussed in Chapter 2: Digital Twin and Ubiquitous feedback, my proposed Digital-Twin Ubiquitous AI Personal Therapist Assistant (UbiMyTherapist) extends beyond traditional conversational chatbots by integrating agents and RAG-LLMs with emotion detection models that can operate on bio-signals, speech intonation, or text sentiment. Unlike other AI-powered systems, my framework does not rely solely on user interactions. It continuously monitors the user's emotional state through wearables, such as smartwatches, smartphones, and earbuds, and delivers timely and appropriate support, even without user prompts. This Digital Twin-driven, multimodal, and proactive approach is designed to ensure individuals receive dynamic and personalized assistance based on both their near-real-time emotional conditions and their historical context.

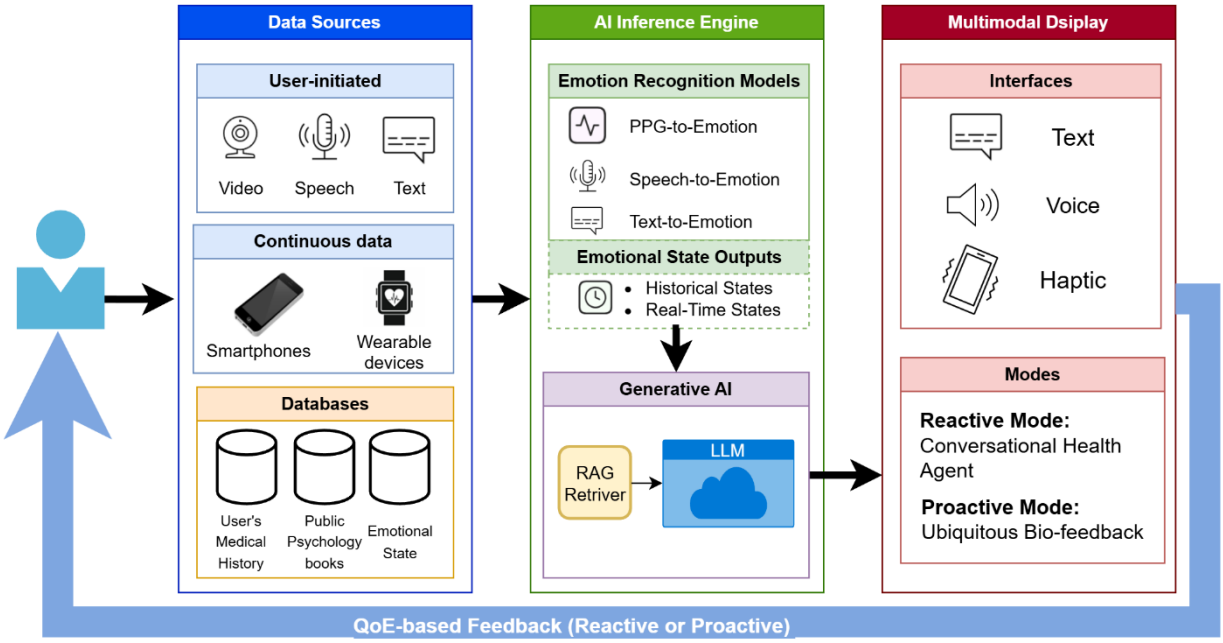


Figure 3.1. UbiMyTherapist Framework Architecture, composed of three layers: Data Sources, AI Inference Engine, and Multimodal Display

3.1 System design

As illustrated in Figure 3.1, the framework is composed of three layers: (1) Data Sources, (2) AI Inference Engine, and (3) Multimodal Display. The Data Sources layer collects information from both user-initiated interactions and continuous measurements captured by wearables, while also

managing knowledge stored in dedicated databases. The AI Inference Engine layer processes these inputs through different emotion recognition models to estimate near-real-time and historical emotional states. It also features a Generative AI module that integrates information retrieved by the RAG from databases, detected emotional states, and the user's prompt, if available, to produce personalized therapeutic responses using an LLM. Finally, the Multimodal Display layer determines the system interaction with the user, supporting text, voice, and haptic interfaces to deliver responses in two complementary modes: Reactive Mode, which acts as a conversational health agent, and Proactive Mode, which is based on ubiquitous biofeedback, enabling interaction without the user explicitly seeking help.

3.1.1 Data sources

The system gathers data from three primary sources: (1) user-initiated inputs, (2) continuous data from wearables, and (3) stored databases. User-initiated inputs can include video, speech, and text, enabling direct interaction with the system in a conversational setting. On the other hand, continuous data is collected passively from wearables, such as smartwatches, smartphones, and other devices. These devices can capture PPG, ECG, EEG, and Skin Conductance Response (SCR) through built-in sensors. Their portability and unobtrusiveness allow seamless monitoring of the user's physiological state without disrupting daily activities. The third source, the databases, plays a crucial role in supporting the digital twin of the user by contextualizing and personalizing therapeutic responses:

3.1.1.1 Public Psychology Database

Contains psychology books and literature. Instead of relying solely on the LLM's pre-trained parameters, our RAG module retrieves information from this source. This method improves therapeutic responses by ensuring they are fact-checked and contextually accurate.

3.1.1.2 User's Medical History Database

Stores personal and medical information, such as age, gender, mental health diagnoses, and regular medications.

3.1.1.3 User's Emotional States Database

Maintains a bidirectional interaction with the AI Inference Engine. Outputs from Emotion Recognition Models are continuously stored here, creating a historical record of the user's affective state. The

RAG retriever later queries this database to improve therapeutic responses, supporting the digital twin concept by incorporating emotional data history.

3.1.2 AI Inference Engine

3.1.2.1 Emotion Recognition Models

This framework includes three models: Biosignals-to-Emotion, Speech-to-Emotion, and Text-to-Emotion, which can be used separately or combined through a decision strategy that weighs predictions by reliability and context for accurate emotional state estimation. The biosignals model can process physiological signals such as PPG, ECG, EEG, or SCR to predict emotions using machine learning. The second model can analyze speech input, focusing on vocal traits like tone, pitch, rhythm, and intensity. The third, Text-to-Emotion model, uses NLP to identify emotions in the user's text. Their outputs support both historical and real-time analyses and are stored in the User's Emotional States Database, which tracks emotional history across adjustable time scales. The RAG retriever integrates this context with knowledge from other databases, enabling responses to adapt in both content and style (e.g., tone and empathy), mirroring human therapists.

3.1.2.2 Generative AI Module

This component features the RAG architecture, combining a retriever and a MultiModal Large Language Model (MM-LLM). The retriever queries databases for relevant information based on semantic similarity to the user's prompt, prioritizing meaning over exact wording. After retrieving content and the user's emotional state, the MM-LLM generates a personalized response aligned with it. The response is delivered either reactively or proactively, depending on the interaction mode.

3.1.3 Multimodal Display

The Multimodal Display layer in Figure 3.1 controls how UbiMyTherapist provides feedback via text, voice, and haptic interfaces, adapting to the user's emotional state to enhance engagement. For example, text responses offer detailed advice and coping tips within an App, voice responses deliver conversational guidance, such as breathing exercises, through earbuds, and haptic feedback, which subtly alerts users to stress, encouraging them to take deep breaths. By offering flexible interaction methods, the framework ensures responses remain contextually appropriate across both reactive (conversational) and proactive (biofeedback-triggered) modes.

3.2 RAG integration

RAG is incorporated into UbiMyTherapist to improve the reliability and personalization of LLM-based mental well-being support.

In this framework, the retriever is part of the AI inference Engine, more specifically, the Generative AI module. The module queries three main sources: (1) the Public Psychology Database, which contains psychology books and literature used to ground responses in evidence-based content, (2) the User's Medical History Database which contains historic information specifically related to the user from demographic data to medication history (3) the User's Emotional States Database, which stores near real-time and historical emotion predictions produced by the emotion recognition models.

The retrieval process is based on semantic similarity to the user's prompt, prioritizing meaning over exact wording, and it retrieves both relevant knowledge and the user's emotional context. The multimodal LLM then generates a response aligned with retrieved knowledge and the user's current emotional state, enabling the system to adapt the response content and style (e.g., tone and empathy) in a therapist-like manner.

More details about the reason of choosing RAG is discussed in chapter 6.

3.3 Reactive vs Proactive Modes

3.3.1 Reactive Mode (Conversational Health Agent)

In this mode, the user initiates interaction via text or voice, and the framework responds in a structured, personalized, and therapist-like manner.

3.3.2 Proactive Mode (Ubiquitous Biofeedback)

In this mode, the system can autonomously detect emotional distress from near real-time physiological signals collected by wearables. When a negative state is identified, it proactively recommends interventions, such as relaxation techniques, motivational prompts, or context-tailored behavioral adjustments. For example, while driving, the system may deliver short voice prompts for breathing exercises, whereas at home it could propose guided meditation or extended therapy sessions. By grounding decisions in the user's emotional state, the proactive mode

implements the digital-twin concept, ensuring support is timely, non-intrusive, and contextually appropriate, even when the user does not explicitly seek help.

3.4 Comparison with Existing Work

As shown in Table 3.1, our proposed framework differs from existing AI-powered mental health systems. The majority of these systems primarily function as chat-based health agents, relying on user-initiated interactions. While some systems, such as Ellie [105] and EmoAda [106], extend beyond traditional text input to include voice and video, they remain reactive, responding only to user-initiated interactions, and lack the continuous monitoring capabilities that wearable devices provide through physiological measurements.

OpenCHA [116], on the other hand, is a multimodal conversational health agent system designed to accept diverse input types, including physiological signals, images, text and voice. However, it is limited by requiring users to manually upload or input these data types rather than collecting them automatically in real-time. Additionally, OpenCHA does not specialize in mental health support but rather functions as a general-purpose conversational health agent. Some systems, such as Woebot [108], Wysa [109], and Youper [110], offer a proactive approach through daily check-ins, but scheduled interactions alone are insufficient for immediate intervention when needed. In contrast, mobile apps like EliteHRV [111] and CuesHub [112] can proactively send stress alerts when wearables detect stress through PPG and ECG measurements. However, these apps are designed to enhance general mental well-being and do not provide a structured therapeutic plan similar to a human therapist. It is worth noting that Ellie [105] and EmoAda [106] have built-in emotion detection from facial expressions and voice analysis, but their monitoring is limited to active sessions.

What differentiates UbiMyTherapist from existing work is its ability to provide users with access to a personal AI-powered therapist assistant that is always ready to assist, even when the user is not actively engaged with the system. My proposed framework bridges the gap between AI-based mental well-being assistants and continuous, near real-time mental health monitoring, integrating an AI therapist assistant with ubiquitous biofeedback from wearable devices to provide timely and personalized support.

Table 3.1. Comparison Of UbiMyTherapist With Some Of The Existing Systems And Applications

System	Type	Input(s)	Emotion Detection Modality (ies)	Physiological Data	Continuous Monitoring	Adaptive Therapy Tone	Interaction Style
Woebot [108]	AI CBT Therapy Chatbot	Text	None	No	No	Yes, based on user mood input	Reactive (user-initiated), Proactive (daily check-ins)
Wysa [109]	AI Mental Health Chatbot	Text	Text	No	No	Yes, based on text analysis	Reactive (user-initiated), Proactive (daily check-ins)
Youper [110]	AI CBT Therapy Chatbot	Text	None	No	No	No	Reactive (user-initiated), Proactive (daily check-ins)
Ellie [105]	Virtual Therapist Avatar	Text, voice, video	Face, body, and voice	No	No	Yes, based on detected emotion	Reactive
OpenCHA [116]	AI Multi-Modal General Health chatbot	Text, voice, video	None	Manual user input	No	No	Reactive
EmoAda [106],	AI Multi-Modal Mental Health Chatbot	Text, voice, video	Face and voice	No	No	Yes, based on detected emotion	Reactive
EliteHRV [111]	Mental Well-Being App	Text	Physiological	From wearables	Yes	No	Proactive (wearable-based)
CuesHub [112]	Mental Well-Being App	Text	Physiological	From wearables	Yes	No	Proactive (wearable-based)
UbiMyT-herapist	**AI Personal Therapist**	**Text, voice**	**Voice and physiological**	**From wearables**	**Yes**	**Yes, based on detected emotion**	**Reactive (user-initiated), Proactive (wearable-based distress detection)**

Chapter 4. WARM-VR Dataset

A core requirement for implementing UbiMyTherapist, introduced in Chapter 3, is the ability to infer users' emotional states reliably from wearable physiological signals using emotion recognition models. Based on the gaps identified in Chapter 2, achieving this requires a dataset that captures affect responses in settings closer to real life, while still using a reproducible protocol and reliable ground truth.

Therefore, Chapter 4 presents the WARM-VR dataset, describing its motivation, data acquisition protocol, sensing equipment, labeling strategy, and initial analyses that demonstrate its value for studying affective responses and supporting the development of wearable emotion recognition models.

4.1 Motivation

HCI has become integral to daily life, particularly with the rapid advancement of machine learning technologies [117]. To further improve the quality of these interactions, the field of affective computing has emerged as an interdisciplinary domain focused on designing systems that recognize, interpret, and respond to human emotions [118]. Specifically, Affective Computing in Multimedia focuses on recognizing the emotional responses that a particular stimulus is likely to elicit in users [119].

In recent years, the rise of wearable devices has provided a valuable new data source for mental well-being and affect recognition [120] [121] [122]. These devices enable continuous, non-intrusive monitoring of physiological signals such as PPG and ECG, which are commonly used to infer affective states [123]. Unlike facial expressions or vocal tone, physiological signals offer objective and reliable indicators of affect, as they are less susceptible to social masking [124] [125] [126] [127]. These advances have inspired the development of datasets aimed at modeling affective states from physiological signals. However, most are constrained by non-immersive environments.

VR is an effective tool for simulating real-life environments, enabling users to interact with digital surroundings and experience a strong sense of presence despite not being physically there. It has gained increasing recognition for its potential to promote relaxation [128] and reduce anxiety [129] through the integration of multiple sensory modalities.

This integration has been shown to enhance the depth of presence and realism in VR systems [128]. While visual and auditory stimuli have traditionally dominated VR experiences, olfactory stimuli are increasingly being explored for their potential to modulate mood and improve emotional well-being [130]. The use of scent in virtual settings, similar to aromatherapy, involves the controlled dispersion of essential oils known for their calming properties. For instance, lavender and orange scents have been shown to reduce anxiety and elevate mood [130].

Despite these advances, existing affect recognition datasets remain limited in several important aspects. Most lack full immersion due to the absence of VR integration, even when multiple sensory stimuli are included. Additionally, physiological data are often collected using stationary lab equipment rather than wearable devices, which undermines real world applications. In contrast, WARM-VR is, to the best of my knowledge, the first dataset specifically designed for affect recognition from wearable devices within a fully immersive virtual environment that also incorporates olfactory stimulation.

The main contribution of this chapter is presenting a new publicly available multimodal dataset (Table 4.1), conducted in an immersive VR environment. Two wearable devices are used: a wrist-based sensor and a chest-based belt, capturing high-resolution physiological signals (PPG, ECG, EDA, TEMP) and motion signal (ACC). The dataset also includes self-reported data of participants' affective states, collected through questionnaires. These responses can serve as valuable labels for training personalized affect recognition models.

Table 4.1. WARM-VR Features

Signal Source	Wearables	Nb of Subjects	Type of Stimuli	Annotations
ECG, PPG, EDA, ACC, TEMP	ECG: Polar H10 belt PPG: Empatica E4 watch	31	Reading, Math test, Olfaction, VR scene	SAM (valence/arousal), RRS (relaxation), and post-experiment questions

4.2 Data Acquisition and Protocol

The data collection process is shown in Figure 4.1.

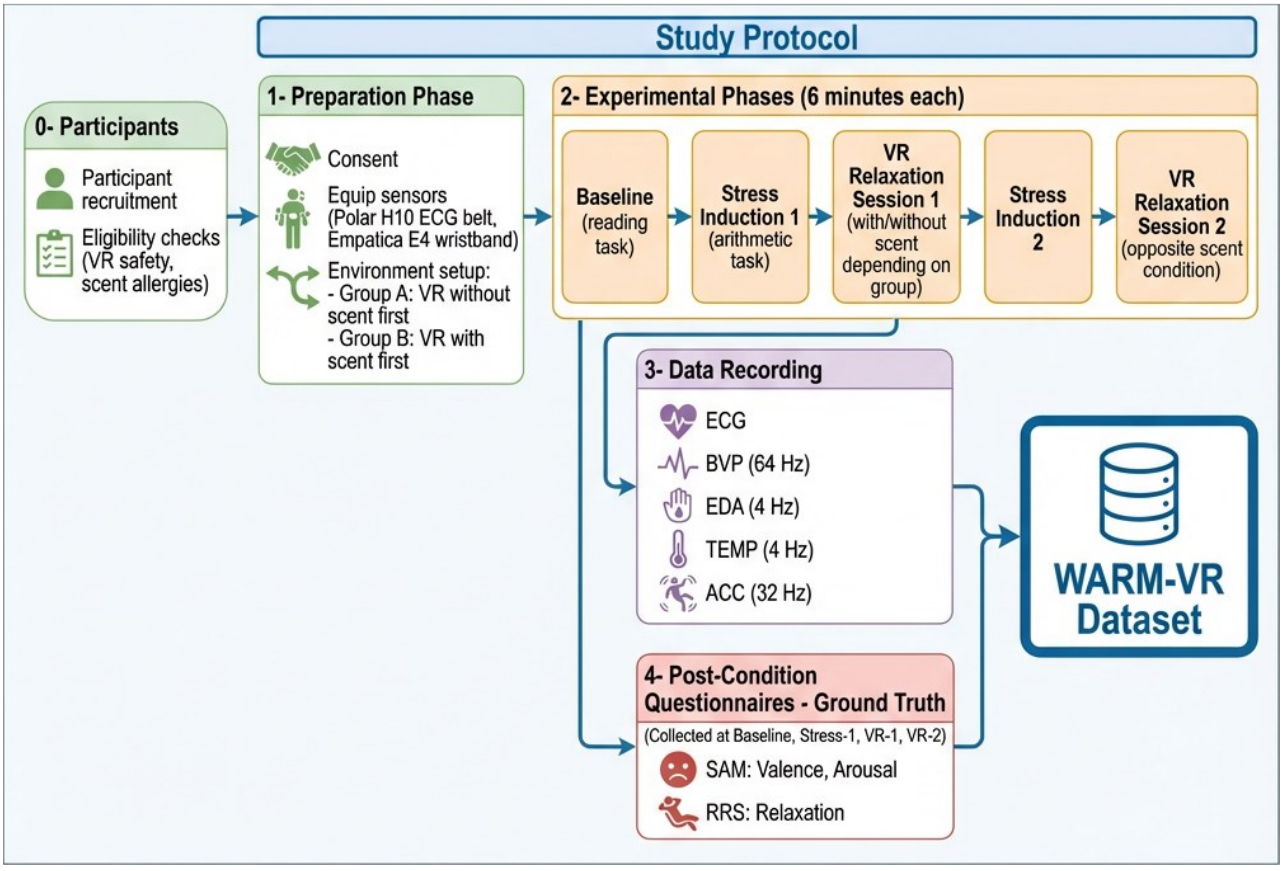


Figure 4.1. Overview of the Dataset Creation Process

4.2.1 Participants

The study involved 31 participants (13 female, 17 male, 1 preferred not to specify) between the ages of 19 and 37 (Mean age = 26.6). Data from the Empatica E4 watch of three participants were excluded due to an error in sensor placement, which affected the data. A significant focus of my dataset was ensuring diversity among the participants, an aspect that has been largely overlooked in popular existing datasets where the majority of participants were from a single background [14] [12] [13]. To achieve this, the participants were from various ethnic backgrounds Figure 4.2. This inclusion of participants from multiple ethnic backgrounds enhances the dataset's demographic diversity, which is important for developing more robust and generalizable affect recognition models. While the sample size is limited and some groups remain underrepresented, it offers broader representation than many existing datasets in the field.

To ensure safety during the VR simulation, individuals with a history of seizures, epilepsy, severe motion sickness, or other neurological conditions were excluded [131]. Participants with known sensitivities or allergies to the scents used in the experiment were also not eligible.

Recruitment was primarily carried out through posters placed around the University of Ottawa campus. Additional participants were recruited using a snowball sampling method, where initial participants were invited to refer friends and acquaintances to the study.

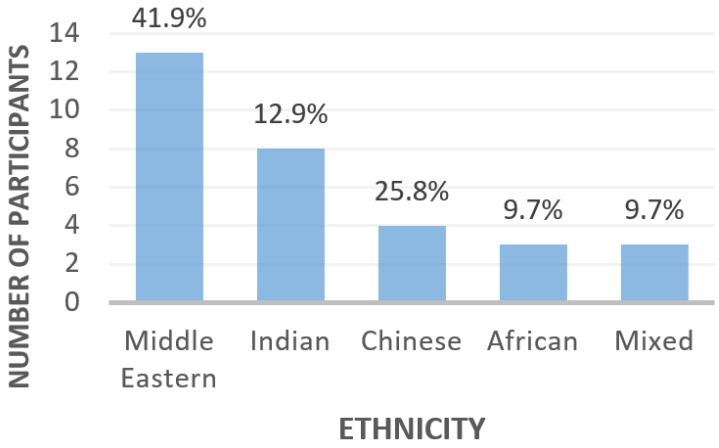


Figure 4.2. Ethnicity Distribution of Participants

4.2.2 Equipment Used

To ensure accurate and high-resolution physiological data collection in a fully immersive environment, a combination of wearable sensors and sensory stimulation tools was employed. The selected equipment enabled synchronized recording of cardiac activity, delivery of olfactory stimuli, and presentation of a visually and auditorily rich virtual environment. Below is an overview of the key devices used in the study:

1. **ECG Belt:** Cardiac electrical activity was measured using the Polar H10 chest strap [132], which participants wore around their ribcage. The device uses two electrode pads to capture ECG signals and records R-R intervals with a sampling rate of 1000 Hz. The R-R intervals data were transmitted in real time via Bluetooth.
2. **Wristband:** Participants also wore the Empatica E4 wristband [133] on their non-dominant hand. This wearable device continuously recorded multiple physiological signals: PPG at 64 Hz, Electrodermal Activity (EDA) at 4 Hz, TEMP at 4 Hz, and ACC at 32 Hz. The E4 is widely

used in affective computing and clinical research [134] due to its high signal fidelity and unobtrusive design [135].

3. **VR Headset:** To create an immersive virtual environment, I employed the Meta Quest 2 headset [136]. Its high-resolution display and integrated motion tracking enabled participants to engage with the VR scenario naturally while remaining seated. The headset's wireless design helped eliminate external distractions.
4. **Olfactory Diffuser:** Olfactory stimulation was delivered via an ultrasonic diffuser that emitted a mist of essential oils chosen to evoke a beach-like scent. The diffuser was activated prior to the participant's entry, ensuring a consistent ambient aroma throughout the session. This passive exposure enhanced immersion without requiring any participant interaction.

Group A		Baseline		Stress Induction		VR-No Olfactory		Stress Induction	VR-With Olfactory
Group B		Baseline		Stress Induction		VR-With Olfactory		Stress Induction	VR-No Olfactory

Figure 4.3. Overview of the Two Study Protocol Groups. Each Phase Lasted 6 Minutes. Blue Boxes Indicate The Time Points when participants completed questionnaires

4.2.3 Study Protocol

The objective of the study was to induce and measure three different affective states: Baseline, Stress, and Relaxation. As shown in Figure. 4, participants followed a structured sequence of phases in one of two protocol groups (Group A or Group B), depending on the order in which they were exposed to olfactory stimuli. Below are the different phases of the study protocol:

1. **Preparation Phase:** The study was conducted in two controlled office spaces: one scent-free for Group A and the other infused with a beach scent for Group B, allowing for controlled exposure to olfactory stimuli. Upon arrival, each participant provided informed consent and completed a pre-experiment questionnaire. They were then fitted with the two wearable devices: a chest-worn ECG belt (Polar H10) and an Empatica E4 wristband, to be used in all the next phases to record the physiological data. Participants were instructed to refrain from alcohol for 12 hours, and from caffeine, food, smoking, and exercise for 3 hours to ensure consistent physiological baselines.

2. **Baseline Phase:** Participants were seated comfortably and asked to read The Silk Roads by Peter Frankopan for six minutes to establish a resting baseline. Reading was selected over personal device use to minimize external stressors and ensure consistency.
3. **Stress Induction Phase:** Participants completed a six-minute arithmetic stress task adapted from [137], involving rapid-fire multiplication problems (numbers from 0 to 12) on a screen with a time limit for each question. Incorrect answers led to score penalties displayed in real time. To increase cognitive pressure, participants were told that their scores were compared to other participants.
4. **VR Relaxation Phase – First Session:** After the first stress phase, participants were fitted with the Meta Quest VR headset and immersed in a virtual beach environment as seen in Figure 4.4. Depending on their group (as shown in Figure 4.3, this first session either included olfactory stimuli (Group B) or not (Group A). The VR session lasted six minutes, during which participants were encouraged to explore the virtual scene naturally while remaining seated.
5. **Second Stress phase and Second VR session:** Participants then completed a second arithmetic stress task identical to the first to re-induce stress, followed by a second VR session in the opposite olfactory condition. Group A, for example, experienced scent-free VR first and scented VR second, while Group B experienced the reverse. This crossover design ensured that the influence of olfactory stimuli could be evaluated independently of the order.
6. **Post-Condition Questionnaires:** Participants completed the Self-Assessment Manikin (SAM) and Relaxation Rating Scale (RRS) questionnaires after the baseline phase, first stress induction phase, and after each VR relaxation phase (as indicated in Figure 4.3). To reduce the study's time and not burden the participants, questionnaires were not repeated after the second stress induction phase. Answers from the questionnaires were used as subjective labels for participants' affective states.

(a) (b)

Figure 4.4. (a) Virtual beach environment with an avatar inside the scene (b) A participant seated in a chair wearing a Meta Quest VR headset, with a diffuser dispersing beach scent to enhance immersion

4.2.4 Ground Truth

To evaluate participants' affective states throughout the experiment, a combination of self-report tools was used to capture both affective responses. These subjective measures served as the ground truth for labeling affective states associated with each experimental phase:

1. **SAM:** Affective states were measured using the SAM, a non-verbal pictorial tool that assesses affect along three dimensions: valence (Unpleasant–Pleasant), arousal (Calm–Excited), and dominance (Control). In this study, I excluded dominance as it is not commonly used in the literature. In the valence-arousal space, affect can be mapped into four quadrants (e.g., high valence and high arousal may indicate excitement or happiness). Participants used visual aids (see Figure 4.5) to rate their affective state on both dimensions.
2. **RRS**: Participants also self-reported their level of relaxation using the RRS. Participants rated their perceived relaxation on a scale from 0 (not relaxed at all) to 10 (extremely relaxed).
3. **VR Relaxation Sessions:** To gain insight into participants' subjective experience with the VR environment, two additional post-experiment questions were included: (1) *"Would you use this method as a relaxation technique?"* and (2) *"Did you feel more immersed with the scent compared to without the scent?"* These questions aimed to assess both the perceived effectiveness of the olfactory-enhanced VR setup and the depth of user immersion, providing valuable context for future real-world applications.

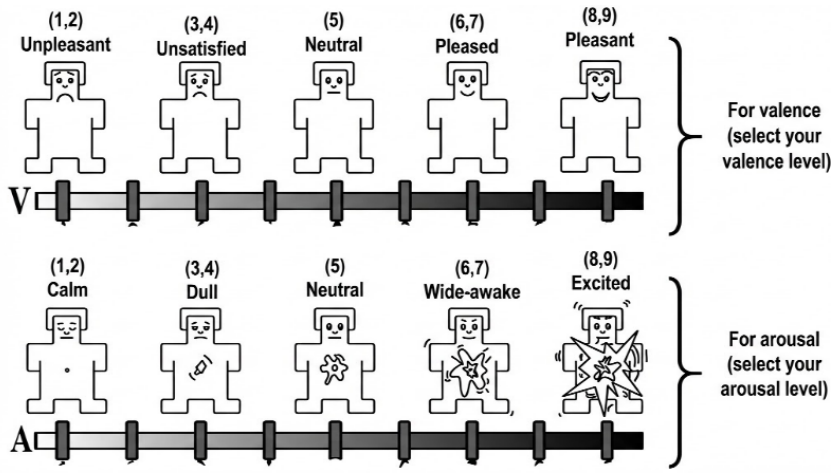


Figure 4.5. Image Shown To Participants To Help Them Answer The SAM Questionnaire

4.3 Use Cases and Methods

The primary goal of this dataset is to explore how immersive environments, specifically those incorporating VR and olfactory stimuli, can support relaxation and reduce stress. Additionally, the dataset was intentionally designed with broader applicability in mind by collecting a diverse set of physiological signals and self-reported measures.

4.3.1 Use Cases

4.3.1.1 Investigating the Effect of Olfactory Stimuli

Because each participant experienced both scented and unscented VR conditions in a counterbalanced design, this dataset allows researchers to isolate and analyze the specific impact of olfactory input on relaxation and stress. The inclusion of paired conditions supports within-subject comparisons and enables analysis of the physiological and subjective differences associated with scent-enhanced immersive experiences.

4.3.1.2 HRV Analysis

The ECG data collected using the Polar H10 chest strap include precise R-R interval measurements, which are ideal for traditional HRV analysis. HRV is a well-established method for assessing Autonomic Nervous System (ANS) activity, particularly in stress research. Frequency-domain features, such as the high-frequency (HF) component (0.15–0.4 Hz), have been shown to correlate strongly with relaxation state [138]. Researchers can extract these features to evaluate physiological responses across different experimental conditions.

4.3.1.3 Affect Measurement via Machine Learning:

The combination of physiological data from both the ECG belt and the E4 wristband, along with subjective relaxation scores from the RRS questionnaire (0–10 scale), offers a robust foundation for training machine learning and deep learning models aimed at relaxation estimation. Additionally, valence and arousal ratings collected via the SAM questionnaire (see Section 3.4) allow for broader affective modeling within the dimensional affective space. These multimodal labels support applications in biofeedback, wellness monitoring, and personalized affect-aware systems.

4.3.1.4 Statistical Analysis of Subjective Responses:

The dataset includes several validated psychological questionnaires, including SAM, and RRS, administered at multiple time points. These tools can be used for traditional statistical analysis to explore the effects of different experimental conditions on mood. This makes the dataset useful not only for signal processing and machine learning but also for hypothesis-driven research in psychology and HCI.

4.3.2 Methods

In this chapter, I demonstrate the applicability of the WARM-VR dataset through three out of four use cases mentioned in the previous subsection, for which I provided experimental results. The Affect Measurement via Machine Learning case is kept for Chapter 5.

4.3.2.1 Investigating the Effect of Olfactory Stimuli:

In my published work [23], I demonstrated using the WARM-VR dataset that olfactory stimuli can subconsciously enhance relaxation in VR. To evaluate this effect, I analyzed the HRV derived from the ECG RR-intervals collected from the Polar H10. Particular attention was given to the HF band (0.15–0.4 Hz), which reflects parasympathetic nervous system activity linked to relaxation [139]. I then computed the Average HF values across all participants and compared them between experimental phases: the stress test and relaxation sessions with and without olfactory stimulation.

4.3.2.2 Statistical Analysis of Subjective Responses:

To validate the study protocol, I evaluated the self-reported questionnaires (Tables 4.2 and 4.3). This helped us verify that the experimental conditions effectively manipulated the subjects' affective states as intended.

4.4 Results

4.4.1 Investigating the Effects of Olfactory Stimuli

4.4.1.1 HRV Analysis

HRV refers to the fluctuations in the time intervals between successive heartbeats. HRV metrics are used to evaluate cardiac health and to provide insights into the ANS status [114]. During relaxation, the parasympathetic nervous system is predominantly activated. Monitoring the balance between the sympathetic and parasympathetic branches allows for inferences regarding the degree of relaxation experienced by an individual.

Assessing relaxation levels using HRV is primarily conducted in the frequency domain. Of particular interest is the high-frequency (HF) band, which ranges from 0.15 to 0.4 Hz. There is substantial evidence that an increase in the HF component of HRV is associated with increased relaxation levels [138]. The HF component is widely regarded as indicative of parasympathetic (vagal) activity, which plays a key role in promoting relaxation and reducing stress. Furthermore, research has demonstrated that individuals with higher HF baseline tend to exhibit better emotional regulation, lower anxiety, and a better capacity to cope with stress, all of which are indicators of a more relaxed physiological state [138].

For the HRV analysis, I utilized raw ECG signals collected from the Polar H10 belt. The subjects' RR intervals were accessed and exported using the Android application “Elite HRV.” Afterward, the data was processed using the Python package “hrv-analysis” [140].

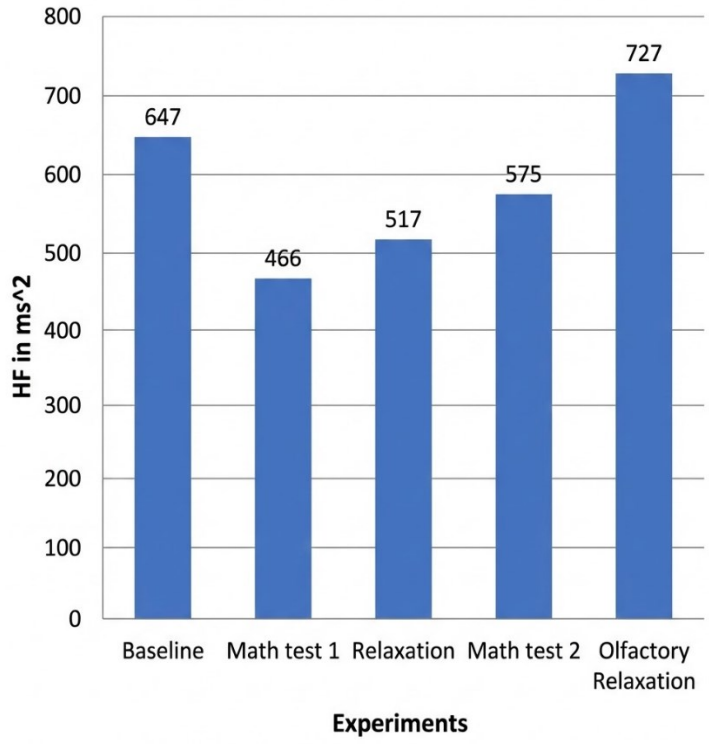


Figure 4.6. Mean High-Frequency HRV Values Across All Participants

Prior to data analysis, RR-interval artifacts, such as outliers and ectopic beats, were identified and removed using the library's built-in artifact-correction methods, which apply filtering techniques to clean the signal. Next, I performed frequency domain analysis by calling the “get frequency domain features” function. This function computes HRV parameters using the Welch method for spectral estimation, applying Fast Fourier Transform (FFT) to obtain power across standard frequency bands. In my case, I specifically extracted the HF component, which reflects parasympathetic activity related to relaxation. The high-frequency parameter was then extracted for each of the five parts of each recording session. The average HF values of all the subjects are shown in Figure 4.6. The percentage increase in HF between the Math stress test and the relaxation experiment was calculated using Equation 1.

$$Percentage\ Increase = (HF_{relaxation} - HF_{Stress\ Test}) \times 100 \quad \textit{(Equation 1)}$$

To compare the HF increase under two conditions—relaxation without olfactory stimulation and relaxation with olfactory stimulation—I calculated the average HF increase for all participants.

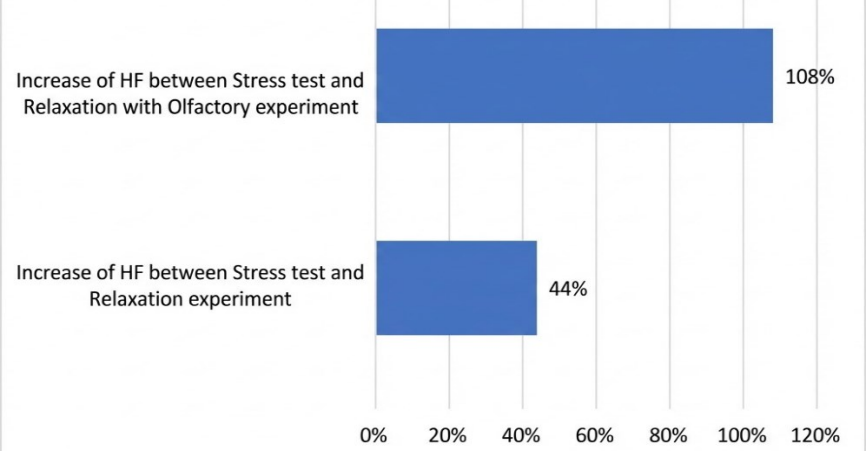


Figure 4.7. HF Percentage Increase Between Stress Test And The Relaxation Experiment

To evaluate the impact of olfactory stimulation on relaxation, I measured the increase of the high-frequency component under two distinct conditions: relaxation without olfactory stimulation and relaxation with olfactory stimulation.

Relaxation without Olfactory Stimulation: In this experiment, participants underwent a relaxation session without olfactory stimuli following a math stress test. The analysis revealed a 44% increase in the HF component when compared to the HF values recorded during the math test. This significant increase suggests that even without olfactory input, VR-based relaxation techniques can effectively enhance parasympathetic activation, promoting a state of relaxation.

Relaxation with Olfactory Stimulation: On the other hand, when olfactory stimulation was introduced during the relaxation session, the HF component saw an even more substantial increase. As shown in Figure 4.7, there was a 109% rise in HF values from the math stress test to the relaxation with scent condition. This substantial increase underscores the potential additive effect of pleasant scents in enhancing the relaxation response beyond VR-based relaxation techniques alone.

4.4.1.2 Statistical Comparison and Significance of the HRV Analysis

To assess the statistical significance of these observations, a paired-samples t-test was conducted. The paired t-test was chosen because it compares two conditions within the same participants, reducing individual variability and providing a more precise measure of the effect of olfactory stimulation. This approach ensures that any observed differences in HF values are directly attributable to changes in conditions rather than to personal factors such as baseline stress levels, age, or health status. This way, I isolate the effect of olfactory stimulation from other personal

factors. The test compared the mean increases in HF values between the two relaxation conditions. The resulting p-value of 0.002 is well below the standard threshold of 0.05, indicating that the observed differences are statistically significant. Therefore, I conclude with confidence that relaxation sessions augmented by olfactory stimulation resulted in greater parasympathetic activation than sessions without such stimulation. This finding highlights the potential of integrating olfactory elements into relaxation practices to enhance their effectiveness.

A subjective rating scale (e.g, Likert scale 1 to 10) was used in measuring perceived stress reduction and relaxation in the olfactory and non-olfactory conditions. A paired t-test was used because the study involved the same participants experiencing both conditions (olfactory and non-olfactory), making the data paired and dependent. This test is ideal for within-subject designs as it accounts for individual differences and focuses on the mean difference between conditions. The mean relaxation score is slightly higher for the olfactory condition (7.68) compared to the non-olfactory condition (7.29). This suggests that participants generally perceived the olfactory-enhanced VR environment as slightly more relaxing than the scent-free environment. However, the difference is quite small (only 0.39 points on a 1-10 scale). The standard deviation is slightly higher in the olfactory condition (2.25) than in the non-olfactory condition (1.94), indicating greater variability in participants' relaxation ratings in the olfactory condition. This also suggests that not everyone experienced the same degree of relaxation from the scent addition. In other words, the perceived relaxation experiences of participants with scent varied more than those without scent.

The increased variability (SD) in the olfactory condition may indicate that individual differences (e.g., personal preferences, scent sensitivity) influence the effectiveness of the olfactory display for different users. The p-value (0.371) is greater than 0.05; the difference in relaxation ratings between the olfactory and non-olfactory conditions is not statistically significant. This means that while the olfactory condition had a higher mean relaxation score, the difference is likely due to random variation rather than a true effect of the olfactory display.

Table 4.1. Mean And SD Scores Of The Valence (SAM Questionnaire), Arousal (SAM Questionnaire), And Relaxation (RRS Questionnaire) During The Three Phases

Group A	**SAM – Valence**	**SAM – Arousal**	**RRS**
Baseline phase	6.13 ± 2.06	4.26 ± 2.66	6.80 ± 2.56
Stress induction phase	5.66 ± 2.69	7.13 ± 1.85	4.33 ± 3.02
VR Relaxation phase	7.13 ± 1.58	4.40 ± 2.72	7.26 ± 1.80
Group B	**SAM – Valence**	**SAM – Arousal**	**RRS**
Baseline phase	6.61 ± 1.77	4.76 ± 2.83	7.69 ± 1.13
Stress induction phase	5.15 ± 2.71	7.76 ± 1.12	2.84 ± 2.03
VR Relaxation with Olfactory phase	7.69 ± 1.13	4.23 ± 3.14	7.30 ± 2.46

4.4.2 Statistical Analysis of Subjective Reponses

1. **SAM and RRS analysis:**

   As shown in Table 4.2, subjective questionnaire analysis confirmed that the stress induction phase was effective in elevating stress before the VR relaxation phases. relaxation's mean scores from the RRS decreased (Group A: 6.80 to 4.33; Group B: 7.69 to 2.84) between the baseline and stress phases, then increased (Group A: 4.33 to 7.26; Group B: 2.84 to 7.30) during the VR relaxation phases. To evaluate the statistical significance of these changes, a one-way repeated measures ANOVA followed by post-hoc paired t-tests was conducted separately for Group A and Group B. The results across the three phases revealed significant effects on arousal (Group A: p=0.0017; Group B: p=0.0009) and relaxation (Group A: p=0.0038; Group B: p=0.00001) in both groups, and on valence in Group B (p=0.0098). Post-hoc paired t-tests showed that, consistent with my expectations, arousal significantly increased from Baseline to Stress induction phases (Group A: p=0.0008; Group B: p=0.0059) and decreased from Stress induction to VR relaxation phases (Group A: p=0.0045; Group B: p=0.0041) in both groups. Additionally, relaxation significantly decreased from Baseline to Stress induction phases (Group A: p=0.0283; Group B: p=0.0017) and increased from Stress induction to VR relaxation phases (Group A: p=0.00001; Group B: p=0.00005). For Group B, valence did not change significantly from Baseline to Stress induction phases, but it did from Stress induction to VR relaxation phases (p=0.0049). These results align with my expectations that during stress induction, I expected arousal to increase and both relaxation and valence

to decrease; conversely, during VR sessions, I expected valence and relaxation to increase and arousal to decrease.

2. **Post-experiment questions analysis:** Most of the participants (74.2%) confirmed that they would use the VR sessions in my study as a relaxation technique. This high percentage of positive responses supports the feasibility of deploying fully immersive VR-based systems for mental well-being applications. Table 4.3 also shows the Mean and SD values of question 2, where I asked the participants to rate how immersive the VR beach environment with olfactory stimuli was compared to one without. Participants rated the olfactory-enhanced session as more immersive, with a Mean score of 3.62 (SD = 0.70), approaching the "very immersive" score of 4. These results indicate that olfactory input contributes meaningfully to the sense of presence in VR, aligning to enhance realism and affective impact in immersive systems.

Table 4.2. Post-Experiment Questionnaire Answers

Question	Answer
Q1. Will you use this as a relaxation technique?	Yes (73.4%), No (13.3%), and Maybe (13.3%)
Q2. How immersive was the beach environment with scent compared to without scent (Use Likert scale from 1 to 5, with 1 being not at all immersive and 5 being the most immersive)	Mean and SD: 3.63 ± 0.70

Chapter 4 introduced WARM-VR and demonstrated that the collected physiological and subjective responses can capture meaningful affect-related variations under immersive conditions. However, for an emotion-aware digital twin, raw signals and labels must be converted into reliable emotion estimates that generalize across individuals and real-world variability. This motivates the next stage of the thesis: developing and evaluating wearable-based emotion recognition models.

Therefore, Chapter 5 investigates multiple deep learning architectures for PPG-based emotion recognition using WARM-VR, with a focus on addressing signal noise, class imbalance, and subject variability to support robust emotion inference.

Chapter 5. Emotion Recognition Models

The introduction of wearable technology has enabled the non-invasive, continuous acquisition of physiological signals, such as EEG, ECG, and PPG [27]. These signals provide objective and reliable indicators of emotional states, as they are less susceptible to social masking compared to physical signals like facial expressions or voice tones [127] [141]. The widespread availability of wearable technology has also driven the development of advanced machine learning algorithms capable of processing the rich data generated by these devices, enabling progress in fields such as mental health monitoring and HCI [25].

Among physiological signals, PPG offers distinct advantages for emotion recognition, particularly when integrated into smartwatches. Unlike EEG or ECG sensors, PPG sensors are highly portable and unobtrusive, allowing for seamless, continuous monitoring without disrupting users' daily activities [6]. These features, combined with the increasing affordability and accessibility of wearable devices, have popularized emotion recognition technologies, making them more practical for a broader audience [7].

This chapter provides a controlled benchmark of multiple CNN-based deep learning architectures for contact-PPG-based emotion recognition, evaluated alongside Transformer- and Mamba-based architectures that have not previously been explored in this context. It also introduces a hybrid CNN–LSTM–TCN architecture to better model both local morphological patterns and longer temporal dependencies in wearable signals. All these models are evaluated under the same preprocessing, labeling, and subject-independent cross-validation protocol.

5.1 Generalization Challenge and Architectural Strategies for Robust PPG-Based Affect Recognition

Despite the progress in emotion detection using PPG, much of the research focuses on achieving high test accuracy while overlooking the crucial aspect of model generalization. Models that perform well in controlled settings often degrade when evaluated on unseen subjects, limiting their practical deployment in real-world scenarios [142]. Prior work has explored CNN-based approaches for PPG emotion recognition, but cross-subject robustness remains an unresolved issue.

This motivates exploring architectural strategies that go beyond a pure CNN baseline by combining models that capture different characteristics of physiological time series. CNNs are effective feature extractors for PPG waveforms, as they can learn local morphological patterns from short temporal neighbors. However, affective responses also exhibit temporal structure that may benefit from explicit sequence modeling [143]. Accordingly, this chapter evaluates multiple model families under a consistent subject-independent protocol: (i) CNN baseline and CNN–RNN hybrids (CNN–LSTM and CNN–GRU), (ii) a proposed hybrid architecture that combines CNN feature extraction with TCN, and (iii) modern long-range sequence models (Transformer and Mamba).

In this context, TCNs represent a promising option for sequence modeling because they support causal temporal processing and can capture longer temporal dependencies through dilated convolutions. While TCNs have been explored for emotion recognition across other modalities, their use for PPG-based emotion recognition remains largely unexplored. To the best of my knowledge, no prior work has combined TCNs with PPG for this task. Based on these considerations, this thesis introduces a hybrid CNN-LSTM-TCN approach. At a high level, raw PPG segments are first processed by a CNN feature extractor, then the learned features are modeled using parallel LSTM and TCN branches, and the resulting representations are fused to support affect classification.

Because generalization is the primary objective, evaluation emphasizes subject-independent validation and metrics that capture robustness beyond accuracy alone. In particular, Area Under the Curve (AUC) summarizes class separability across decision thresholds, while class-wise and average F1 scores provide a more informative view of performance under class imbalance by reflecting how well the model detects both low- and high-affect classes at the chosen operating point.

5.2 Methodology:

Fig. 5.1 shows the overall workflow of the machine learning validation pipeline used in this study.

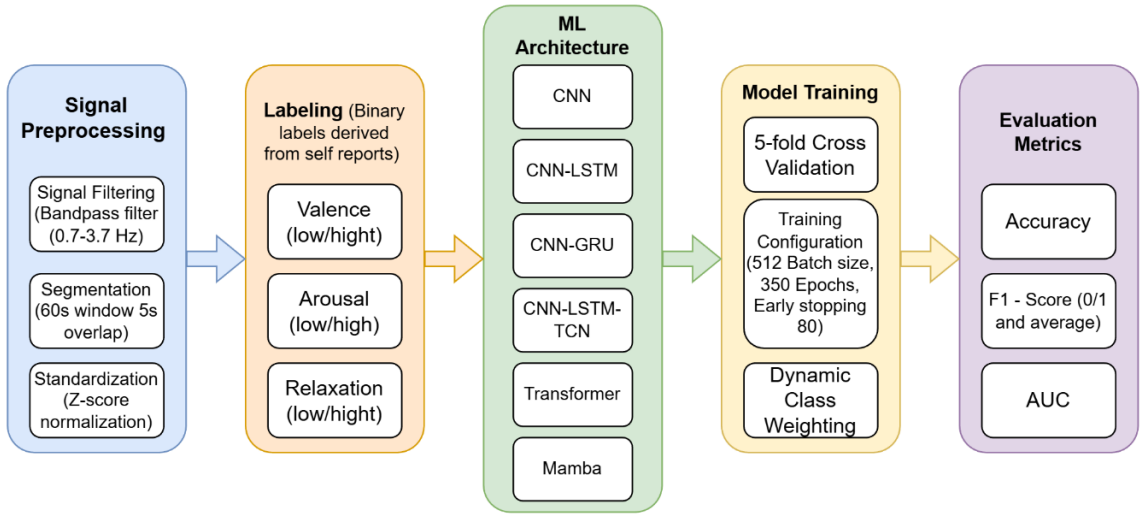


Figure 5.1. Workflow Of The Machine Learning Validation Pipeline For Affect Recognition Using WARM-VR Dataset

5.2.1 Signal Preprocessing

I applied identical filtering, segmentation, and normalization procedures to all the architectures to ensure consistency and reproducibility.

1. **Signal Filtering:** To enhance signal quality, a third-order Butterworth bandpass filter with cutoff frequencies of 0.7–3.7 Hz is applied, corresponding to heart rates of approximately 40–220 bpm. This range helps suppress motion artifacts and low/high-frequency noise. The filtering approach follows the methodology used in [144].
2. **Segmentation:** Following filtering, the PPG signal was divided into fixed-length windows to capture temporal variations relevant to emotional responses. A sliding-window segmentation with 60-second windows and a 5-second overlap is employed, consistent with previous studies [20] [60]. Each segment was assigned an arousal, valence, and relaxation label based on the annotation protocol described in Section IV-A.
3. **Standardization:** Before model training, each segment was standardized using Z-score normalization.

5.2.2 Labeling

Affect representations capture more nuanced physiological variations and avoid limitations of categorical labels, which may not translate consistently across contexts or languages [73]. Accordingly, I used the WARM-VR dataset introduced in Chapter 4, as the main dataset, which provides valence, arousal, and relaxation annotations.

This dataset is particularly relevant for wearable affective computing because compared to popular datasets, it includes more participants than PPGE [13] or WESAD [12] datasets and uses wrist-worn PPG rather than fingertip sensors, as in DEAP [14].

5.2.3 Machine Learning Architectures

To assess the real-world applicability of emotion recognition models, I analyzed the most commonly used architectures from the literature. Over the past few years, CNN-based models have been the most widely used architectures for PPG-based affect recognition [45] [60] [59] [13].

More recently, long-range sequence models, such as Transformers, have become state-of-the-art (SOTA) in many fields, particularly in NLP [145]. In parallel, state-space models, such as Mamba, have emerged as alternatives to Transformers for efficient long-sequence modeling [146]. Despite their success in other domains, it remains unclear whether Transformers and Mamba actually provide added value over CNN-based architectures for PPG-based affect recognition, especially under real-world constraints on wearable data.

Additionally, I introduced a new novel hybrid architecture that combines the strengths of CNNs, LSTMs, and TCNs.

5.2.3.1 CNN-based models

5.2.3.1.1 CNN baseline

For the baseline, I employed a 1D CNN due to its demonstrated effectiveness in affect recognition from PPG signals, as shown in previous works such as Lee M. et al [59] [45] and Rashid N. et al [60].

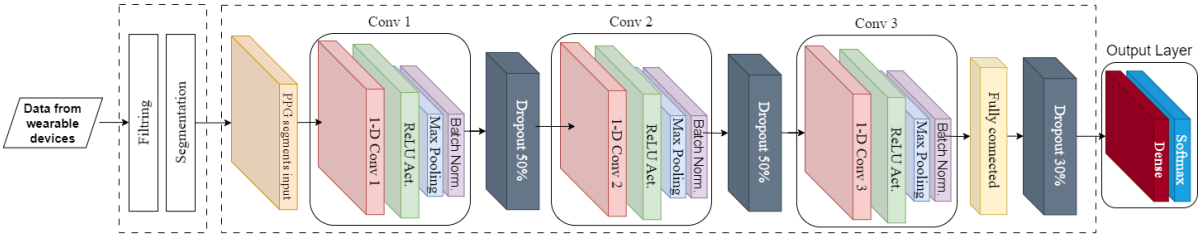


Figure 5.2. PPG Based CNN Architecture

The CNN model (Figure 5.2) begins with a first convolutional layer that utilizes 8 filters, each with a kernel size of 64, and a stride of 4. The convolution operation is applied with 'same' padding, ensuring the output maintains the same length as the input. The Rectified Linear Unit (ReLU) activation function is employed to introduce non-linearity. Following the convolution, a max pooling layer with a pool size of 4 is applied, reducing the spatial dimensionality and allowing for the extraction of dominant features. Batch normalization is then performed to stabilize the learning process and improve convergence.

A dropout layer with a 50% dropout rate is subsequently included to prevent overfitting by randomly deactivating neurons during training. The second convolutional layer follows, with 16 filters, a kernel size of 32, and a stride of 2. Similar to the first layer, ReLU activation and 'same' padding are used. Max pooling and batch normalization are repeated to further condense and normalize the feature maps. Another 50% dropout layer is added to enhance generalization.

The third convolutional layer employs 8 filters with a kernel size of 16 and a stride of 1, again using ReLU activation and 'same' padding. This is followed by a final max pooling layer with a pool size of 4, which further reduces the feature map size.

The model then includes a flatten layer, which transforms the 3D feature maps into a 1D vector, preparing it for the fully connected layers. A dropout layer with a 30% dropout rate is applied before the final output layer to prevent overfitting further.

The output layer is a dense layer with two neurons, using a SoftMax activation function to produce probabilities for the binary classification of valence or arousal states. This design allows the model to learn complex temporal features from the input signals, ultimately facilitating the accurate prediction of emotional states based on physiological data.

5.2.3.1.2 CNN-RNN variants

In addition to the baseline CNN, I implemented several variants that combine the CNN feature extractor with a recurrent layer to model temporal dependencies across the learned feature sequence. Specifically, the first two convolutional blocks of the CNN baseline, along with a 50% dropout, serve as a shared front-end, and the resulting feature sequence is passed to an LSTM or Gated Recurrent Unit (GRU) layer.

To examine the effect of recurrent capacity and directionality, I evaluated models with 12 and 32 recurrent units and included both unidirectional and bidirectional configurations.

The evaluated models are shown in Table 5.1.

Table 5.1. CNN-Based Architectures And Variants

Model name	Convolutional Blocks	RNN type	Units	Bi-directional	Dropout
CNN-GRU-12	Conv 1 and Conv 2 from CNN baseline	GRU	12	No	50% before the Layer
CNN-GRU-32	Conv 1 and Conv 2 from CNN baseline	GRU	32	No	50% before the Layer
CNN-Bi-GRU-12	Conv 1 and Conv 2 from CNN baseline	GRU	12	Yes	50% before the Layer
CNN-LSTM-12	Conv 1 and Conv 2 from CNN baseline	LSTM	12	No	50% before the Layer
CNN-LSTM-32	Conv 1 and Conv 2 from CNN baseline	LSTM	32	No	50% before the Layer
CNN-Bi-LSTM-12	Conv 1 and Conv 2 from CNN baseline	LSTM	12	Yes	50% before the Layer

5.2.3.2 CNN-LSTM-TCN

My proposed model architecture combines CNN, LSTM, and TCN to leverage the strengths of each network type for enhanced feature extraction and temporal modeling in physiological signal analysis. This integrated design aims to improve the robustness of emotion detection by harnessing the complementary capabilities of the three components. The architecture is illustrated in Figure 5.3.

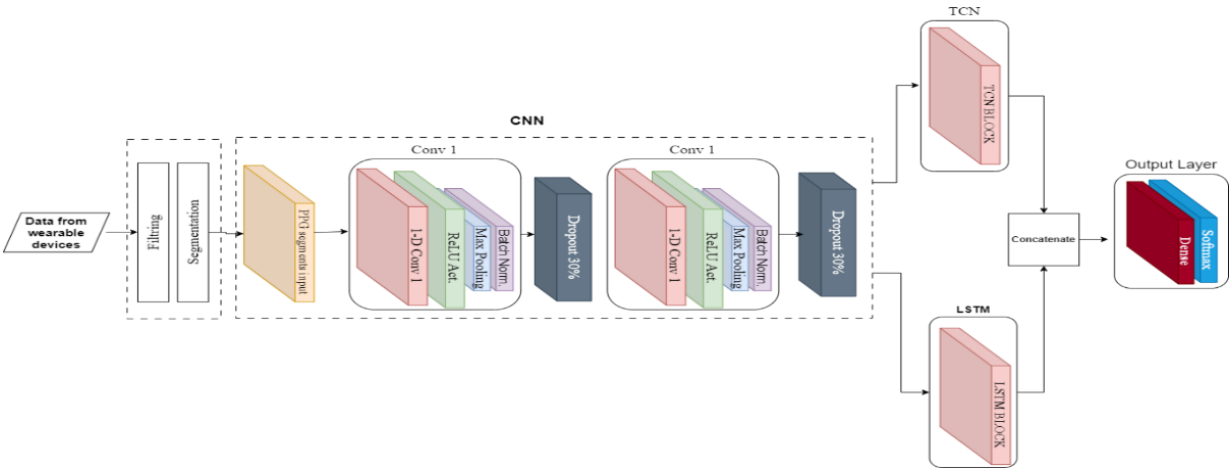


Figure 5.3. Overview Of The Proposed Model

The model begins with a shared CNN branch that processes the input signal through a series of convolutional layers. The first CNN layer applies 8 filters with a kernel size of 64 and a stride of 4, using ReLU activation and 'same' padding. This is followed by max pooling and batch normalization to reduce dimensionality and stabilize learning, respectively. To prevent overfitting, a dropout layer with a 30% rate is included. A second convolutional layer, comprising 16 filters with a kernel size of 32 and a stride of 2, further refines the feature maps. This layer is also followed by max pooling, batch normalization, and dropout to ensure robust feature extraction.

The processed features from the CNN layers are then fed into two parallel branches: a TCN branch and an LSTM branch.

TCN Branch: This branch employs a Temporal Convolutional Network (TCN) with 8 filters, a kernel size of 32, and dilation rates set to [1, 2, 4, 8]. Causal padding is used to maintain the temporal order of the input signal. Skip connections are incorporated to improve gradient flow, and a dropout rate of 30% is applied to mitigate overfitting. I implemented the TCN using the KerasTCN library [147].

LSTM Branch: This branch includes an LSTM layer with 12 units to capture sequential dependencies and long-range temporal dynamics within the data. The LSTM focuses on modeling the inherent temporal patterns in the signal.

The outputs from the TCN and LSTM branches are concatenated to create a combined feature representation that integrates spatial and temporal information. This fused representation is passed to a fully connected output layer consisting of two neurons with a SoftMax activation function, enabling the model to perform binary classification for valence or arousal states.

By combining the spatial feature extraction capability of CNNs, the sequential modeling strengths of LSTMs, and the hierarchical temporal processing power of TCNs, I hypothesize that the proposed architecture may achieve better robustness and accurate emotion detection from PPG signals, compared to using CNN and LSTM alone.

5.2.3.3 Transformer

The Transformer architecture adapts an encoder-style design to 1D PPG segments by first converting the raw signal into a sequence of patch embeddings. Given an input segment, a 1D convolutional patch-embedding layer (kernel size 32, stride 32) projects a non-overlapping patch into a latent vector of dimension *dmodel*. I evaluated two configurations of model dimensionality: a lightweight version "Transformer_v1" with *dmodel* = 32 and 4 attention heads, and a higher-capacity version "Transformer-v2" with dmodel = 64 and 2 heads. To preserve temporal structure, a learned positional embedding is added to each token. The embedded sequence is then passed through a Transformer encoder block (depth = 1), consisting of multi-head self-attention, residual connections, and layer normalization, followed by a feed-forward network with GELU activation and an expansion ratio of 2, wrapped in normalization and residual connections. A 30% dropout is applied to both attention and feed-forward outputs for regularization. A global average pooling operation aggregates the token representations, and a final softmax layer performs the binary classification.

5.2.3.4 Mamba

I also evaluated state-space sequence models based on the recent Mamba architecture. Similar to the Transformer, the Mamba networks first apply a 1D convolutional patch embedding (kernel size = 32) to convert the raw PPG waveform into a sequence of tokens. The resulting feature map is transposed to a (tokens, dmodel) representation and passed through a single Mamba block (depth = 1), operating in the token space with model width (dmodel), state size (dstate = 64), local convolution width (dconv = 4), and expansion factor 2. These blocks implement a selective state-space model that jointly captures local and long-range temporal structure while maintaining linear complexity with respect to the sequence length. After the Mamba block, I perform global average pooling over the tokens, apply dropout, and use a final softmax layer for binary classification. I investigate two configurations. The first, "Mamba", uses dmodel = 64, non-overlapping patches (stride = 32), and 30% dropout, yielding a relatively higher-capacity model with fewer tokens. The second, "Mamba-v2", reduces the model width to dmodel = 48 but introduces 50% overlapping

patches (stride = 16) and a slightly higher dropout rate (40%), approximately doubling the number of tokens and providing finer temporal resolution. These two variants enable us to examine the trade-off among model width, token density, and regularization in PPG-based affect recognition.

Due to the limited dataset size, increasing the model depth beyond a single block led to rapid overfitting for both the Transformer and Mamba architectures. I therefore adopt depth = 1 for all long-range models, yielding compact variants better suited to short PPG windows and small-sample training conditions.

5.3 Model Training

To prevent hyperparameter tuning from biasing the final evaluation on WARM-VR, I followed a two-stage training procedure. In the first stage, a trial-and-error hyperparameter tuning was performed on learning rate, dropout, and architectural width parameters, such as dmodel, using the PPGE dataset (18 participants), which served exclusively as a development set.

The second stage was evaluating the models on WARM-VR (28 participants), where I adopted a subject-independent 5-fold cross-validation protocol to ensure that no subject's data appeared in both training and testing splits. This setup reflects real-world deployment scenarios in which models must generalize to unseen users. All reported results in this paper are based solely on the WARM-VR dataset, using the selected hyperparameters without further modification.

5.3.1.1 Training Configuration:

Following the training protocol described in [20], all models were trained with a batch size of 512, up to 350 epochs, and an early stopping patience of 80 epochs based on validation accuracy. This configuration was found to provide stable convergence across all architectures.

5.3.1.2 Class Weighting

To address class imbalance across affective states, I used dynamic class weighting during training. The class weights w_c (Equation 1) computed from the training fold at each split and applied to the loss function (Equation 2), ensuring that minority classes contributed proportionally to the gradient updates. This approach improved the model's sensitivity to underrepresented affect states.

$$w_c = \frac{N}{K\, n_c} \qquad Equation\ 1$$

$\boldsymbol{N}$ is the number of training samples
$\boldsymbol{K}$ is the number of classes
nc is the number of training samples belonging to class c

$$\mathcal{L}(y, \hat{y}) = -\sum_{c=1}^{K} w_c y_c \log(\hat{y}_c) \qquad Equation\ 2$$

$\boldsymbol{y_c}$ true label for class c
$\hat{y}_c$ predicted probability for class c

5.3.1.3 Optimization and Loss Function

For binary classification tasks (low vs. high arousal), I used categorical cross-entropy as described above, with class weights set to be consistent across all models. The feed-forward networks within the Transformer blocks used GELU activation, whereas convolutional layers in CNN, CNN–RNN and CNN-LSTM-TCN models employed ReLU activation. The Mamba models were implemented using the official PyTorch Mamba library, ensuring correct usage of selective state space modules and compatibility with the latest Mamba kernel implementations

5.4 Evaluation Metrics

I evaluated model performance using validation accuracy, the class-wise F1 scores (F1-0 and F1-1), the average F1 score, and AUC. Accuracy provides a general measure of correctness but may not adequately reflect performance differences between classes, particularly in affect-related tasks where class distributions may be uneven. This is because it cannot measure misclassification costs, which is a major drawback in real-world applications where these costs are often unequal [38]. To capture class-specific behavior, I report F1-0 and F1-1, which correspond to the F1 scores for the low and high affect states, respectively (e.g., low vs. high arousal or low vs. high valence). The average F1 score is computed as the unweighted average of F1-0 and F1-1, giving equal weight to both classes and providing a more reliable measure than Accuracy under potential class imbalance.

5.5 Results

Table 5.2, 5.3 and 5.4 summarize the classification performance of all evaluated architectures across the three affective states: Arousal, Valence, and Relaxation. Performance is reported using validation accuracy, F1-0, F1-1, average F1, and AUC computed under a subject-independent 5-fold cross-validation protocol on the WARM-VR dataset.

Table 5.2. Average Of Subject Independent 5-Fold Cross-Validation Results For Arousal Binary Classification Affect States (Low Vs. High) Using PPG Data Of WARM-VR Dataset (28 participants)

Model	Accuracy	F1-0	F1-1	Average F1	AUC
CNN	**0.7**	0.47	0.68	0.58	**0.68**
CNN-GRU-12	0.63	0.32	0.68	0.50	0.62
CNN-GRU-32	0.67	0.4	**0.7**	0.55	0.67
CNN-Bi-GRU-12	0.64	0.38	0.69	0.54	0.59
CNN-LSTM-12	0.61	0.48	0.55	0.52	0.57
CNN-LSTM-32	0.66	0.45	0.67	0.56	0.59
CNN-Bi-LSTM-12	0.66	0.40	0.68	0.54	0.63
CNN-LSTM-TCN	0.64	0.52	0.56	0.54	0.55
Transformer_v1	0.62	**0.54**	0.63	**0.59**	0.65
Transformer_v2	0.63	0.50	0.65	0.58	**0.68**
Mamba	0.64	0.50	0.63	0.57	0.54
Mamba_v2	0.68	0.53	0.61	0.57	0.57

5.5.1 Arousal Classification

The CNN baseline achieved the highest validation accuracy (0.70), but the Transformer_v1 model (d_model=32, heads=4) achieved the highest average-F1 (0.59), primarily by improving the low-arousal class (F1-0 = 0.54 compared to 0.47 for the CNN), at the cost of lower accuracy (0.63). Mamba models showed slightly lower average-F1 scores (0.57) and validation accuracy of 0.64 and 0.68, respectively. The proposed CNN–LSTM–TCN model did not improve arousal classification under this setting, achieving an average-F1 of 0.54 (accuracy 0.64, AUC 0.55). This suggests that, for

arousal, the additional temporal modeling capacity introduced by the CNN–LSTM–TCN architecture does not translate into improved generalization under the current dataset size.

Table 5.3. Average Of Subject Independent 5-Fold Cross-Validation Results For Valence Binary Classification Affect States (Low Vs. High) Using Ppg Data Of WARM-VR Dataset (28 participants)

Model	Accuracy	F1-0	F1-1	Average F1	AUC
CNN	0.69	**0.50**	0.76	**0.63**	**0.69**
CNN-GRU-12	0.64	0.44	0.73	0.59	0.63
CNN-GRU-32	0.64	0.46	0.71	0.59	0.62
CNN-Bi-GRU-12	0.68	0.49	0.76	**0.63**	**0.69**
CNN-LSTM-12	0.67	0.35	0.78	0.57	0.57
CNN-LSTM-32	0.67	0.36	0.75	0.56	0.57
CNN-Bi-LSTM-12	**0.70**	0.38	**0.80**	0.59	0.59
CNN-LSTM-TCN	0.66	0.38	0.74	0.56	0.66
Transformer_v1	0.68	0.40	0.77	0.59	0.62
Transformer_v2	0.68	0.40	0.78	0.59	0.63
Mamba	0.63	0.30	0.74	0.52	0.52
Mamba_v2	0.67	0.29	0.77	0.53	0.54

5.5.2 Valence Classification

For valence, the CNN baseline remained the strongest overall model, achieving high validation accuracy (0.69) and a top average F1 score (0.63) and AUC (0.69). A closely related CNN-based variant, CNN-Bi-GRU-12, matched the CNN baseline on average F1 (0.63) and AUC (0.69). Transformer variants performed moderately well, with close accuracy of 0.68 but a lower average-F1 value of 0.59. The proposed CNN–LSTM–TCN model achieved an average-F1 of 0.56, and AUC 0.66, which did not compete with the CNN baseline results. Mamba and Mamba_v2 had the lowest

average F1 scores (0.52 and 0.53), indicating weaker discrimination capability of the valence dimension.

Table 5.4. Average Of Subject Independent 5-Fold Cross-Validation Results For Relaxation Binary Classification Affect States (Low Vs. High) Using Ppg Data Of WARM-VR Dataset (28 participants)

Model	Accuracy	F1-0	F1-1	Average F1	AUC
CNN	**0.71**	0.46	**0.79**	0.63	0.60
CNN-GRU-12	0.65	0.56	0.68	0.62	0.64
CNN-GRU-32	0.62	0.54	0.66	0.60	0.63
CNN-Bi-GRU-12	0.66	**0.59**	0.69	0.64	0.69
CNN-LSTM-12	0.62	0.48	0.69	0.59	0.61
CNN-LSTM-32	0.68	0.36	0.78	0.57	0.57
CNN-Bi-LSTM-12	0.68	0.48	0.75	0.62	0.63
CNN-LSTM-TCN	**0.71**	0.56	0.78	**0.67**	**0.70**
Transformer_v1	0.68	0.52	0.76	0.64	0.68
Transformer_v2	0.67	0.51	0.74	0.63	0.67
Mamba	0.67	0.49	0.76	0.63	0.65
Mamba_v2	0.69	0.48	0.78	0.63	0.6

5.5.3 Relaxation Classification

Relaxation is the task where the proposed hybrid model demonstrates its main advantage. While the CNN baseline achieved the highest accuracy (0.71) with an average-F1 of 0.63, the CNN–LSTM–TCN model achieved the highest average-F1 score (0.67) and the highest AUC (0.70) among the evaluated models, while maintaining the same accuracy (0.71). This improvement was driven primarily by a substantial gain in the low-relaxation class (F1-0 = 0.56 compared to 0.46 for the CNN baseline),

while preserving strong performance on the high-relaxation class (F1-1 = 0.78 compared to 0.79). Transformer_v1 and CNN-Bi-GRU-12 remained competitive (average-F1 = 0.64), but neither matched the hybrid model's average-F1 and AUC. These results indicate that the hybrid architecture's combination of local CNN feature learning with complementary LSTM and TCN can be particularly beneficial for relaxation detection in contact-PPG signals.

5.5.4 Discussion

Across the three affective dimensions, the CNN baseline remains a strong and reliable choice, achieving high accuracy for all tasks and maintaining competitive average-F1 performance. However, the results also show that the best-performing architecture depends on the affective dimension being modeled, and that improvements are not uniform across arousal, valence, and relaxation.

First, the CNN–RNN variants (CNN–LSTM and CNN–GRU configurations) do not provide consistent gains over the baseline. In several cases, adding recurrent layers increases model complexity without improving subject-independent performance, suggesting that the temporal structure captured by simple recurrent layers is either insufficiently informative in these 60-second PPG segments or difficult to optimize robustly under the dataset's size and noise characteristics.

Second, Transformer models produced more balanced F1-scores between the minority and majority classes. For both arousal and relaxation, the Transformer improved the F1-score of the low-affect class without substantially sacrificing the high-affect class. For example, in relaxation classification, the Transformer_v1 increased F1-0 from 0.46 to 0.52 compared to CNN, while maintaining an F1-1 close to the CNN's value. A similar pattern appears in arousal. The Transformer_v2 with dmodel=64 and heads=2 had a similar pattern but lower values. This suggests that Transformer architectures may be less biased toward dominant classes and could be advantageous in scenarios where minority-class sensitivity is a primary objective. Although this did not translate into a higher accuracy, the improved balance in class-wise performance may be meaningful for certain application contexts.

Third, the proposed CNN–LSTM–TCN hybrid model provides a meaningful contribution, but its benefit is task-dependent. It does not improve arousal or valence classification under the current experimental conditions but it substantially improves relaxation detection, achieving the best average-F1 and AUC and notably improving the low-relaxation class performance. This indicates

that relaxation-related physiological patterns in contact-PPG may contain temporal structure that is better captured by the combined LSTM and TCN modeling strategy than by CNN-only or CNN–RNN baselines.

Finally, the Mamba models demonstrate stable but not superior performance. While competitive in some cases, they do not outperform the CNN baseline or the best Transformer results in any affective dimension. This is likely because their long-range modeling capacity is underutilized in small, noisy, short-context PPG datasets.

These findings indicate that neither Transformer nor Mamba architectures currently provide clear advantages over simpler CNN models for contact PPG-based affect recognition, particularly when dataset size is limited. This reinforces the importance of model simplicity and robustness when designing wearable affect recognition systems, with CNNs remaining the most reliable and computationally efficient choice at present. At the same time, the hybrid CNN–LSTM–TCN results for relaxation highlight that carefully designed hybrid architectures can yield meaningful gains.

It is important to note that the evaluation is constrained by the small dataset size, short window lengths, and the absence of self-supervised pretraining, which may prevent long-range models from realizing their full potential. Until larger PPG datasets emerge and until pretrained models become more feasible, it will be important to revisit these comparisons to determine whether long-range sequence models can offer advantages under different data conditions.

Chapter 5 addressed the sensing and prediction component of the pipeline by developing models that estimate emotional state from wearable physiological signals. However, recognizing emotion alone does not provide the user with meaningful support unless the system can translate that state into safe, helpful, and personalized guidance. LLMs offer an effective interface for delivering such guidance, but their reliability and tendency to hallucinate raise concerns, especially in mental health contexts. For this reason, Chapter 6 examines RAG as a strategy to ground responses in curated knowledge and evaluates how different LLM configurations balance factual accuracy with therapist-like communication quality.

Chapter 6. RAG-LLM Systems for Mental Health

As described in Chapter 3, implementing UbiMyTherapist requires addressing two core requirements: reliable wearable-based emotion recognition and reliable conversational support that can adapt to the user over time. After establishing the emotion recognition component in Chapter 4 and 5, this chapter focuses on the second requirement: generating safe, grounded, and personalized responses for mental well-being support.

With the increasing demand for mental health support, RAG systems offer a promising solution by combining retrieval and generative capabilities to provide accurate and contextually relevant responses. This chapter examines the application of RAG systems in mental health. I describe the implementation of a RAG system for a mental health therapist assistant chatbot, evaluating how different LLMs perform within the RAG framework and assessing their ability to generate accurate and therapist-like responses in a mental health context. I used similarity metrics (ROUGE, BLEURT and BERTScore) to measure response accuracy, and human experts to judge how well responses align with actual therapist techniques and conversation style.

6.1 Problem: hallucination and lack of personalization in LLMs

As discussed in Chapter 2, LLMs, when used in isolation, exhibit limitations that restrict their reliability in real-world mental health support. Two challenges are particularly important: hallucination and insufficient personalization.

LLMs are known to occasionally generate information that appears plausible but is factually incorrect or unsupported, a phenomenon commonly referred to as hallucination. Within mental health applications, this risk is particularly concerning due to the sensitivity of user's questions and the potential consequences of inaccurate guidance. Providing misleading or incorrect responses may negatively affect users' well-being, highlighting the importance of minimizing hallucination when designing virtual therapeutic agents.

In addition, personalization is a core component of effective mental health support. A therapist-like system must consider the user's background and medical history to provide guidance aligned with their specific needs. Generic or context-agnostic responses may reduce the perceived empathy and effectiveness of the support provided. Therefore, integrating mechanisms that allow the model to

recall user-specific information and adapt to individual profiles is essential for building trustworthy and useful mental health assistants.

6.2 RAG architecture as a solution

To address these limitations, RAG systems represent an evolution in AI technology, merging retrieval and generative capabilities to create context-aware responses that are both accurate and relevant [10]. RAG systems incorporate external knowledge sources, allowing them to dynamically pull from up-to-date, specialized information during response generation, an approach that overcomes the static knowledge constraints of standard LLMs [16]. This hybrid model is particularly valuable in healthcare, where precise, adaptable responses are crucial [84]. By combining a retrieval component that searches for relevant information with a generation component that creates responses, RAG systems can attain a depth of understanding and adaptability that standalone LLMs cannot [148] [16].

In mental health applications, where user needs are complex and evolving, the value of RAG systems is especially apparent. RAG systems can enrich therapeutic interactions by tailoring responses based on individual user profiles while grounding responses in both evidence-based mental health resources and user-specific data [16]. Such systems are poised to support timely interventions, enhance user engagement, and provide mental health guidance that is not only contextually relevant but also dynamically informed by the latest knowledge. However, to realize this potential, these systems must navigate substantial challenges, ensuring response accuracy across diverse user needs.

6.3 Evaluation of different LLMs

6.3.1 Methodology

To perform our evaluation, I designed a RAG framework to act as a mental health therapist assistant chatbot, with a focus on delivering accurate, evidence-based, and therapeutic-aligned responses grounded in CBT principles. I chose CBT as the focus, so I am able to evaluate the RAG better.

6.3.1.1 Overview of the RAG Framework

6.3.1.1.1 Architecture description

The architecture integrates components for embedding, retrieval, generation and citation management. At its core, the system consists of several components, including

- **Central chatbot orchestrator**: manages user queries and coordinates information flow.
- **Embedding Model:** generates dimensional vector representations of text.
- **Retrieval Index:** implements vector-based, graph-based, or hybrid.
- **Citation extraction and management tools:** ensure transparency and traceability of generated responses.

6.3.1.1.2 Processing pipeline

The query processing pipeline is the core mechanism that enables the RAG system to retrieve and generate responses dynamically. It ensures that user queries are efficiently matched with the most relevant document chunks from the knowledge base before passing the retrieved context to the language model for response generation. This process involves multiple steps, such as document retrieval, context processing, and response synthesis. All these steps help improve the accuracy and coherence of the final output. Below is a breakdown of the key stages in the pipeline:

- **Database segmentation:** Documents in the database are segmented into optimized, overlapping text chunks to preserve semantic cohesion. Each chunk is embedded into a high-dimensional vector space and indexed.
- **User Input:** User submits a question to the chatbot.
- **Document Retrieval:** Relevant document chunks are retrieved using the designated retrieval strategy.
- **Context Processing:** Retrieved information is passed as context to the LLM.
- **Response Generation:** The system synthesizes an answer based on both retrieved content and model knowledge.

6.3.1.2 Database

The dataset consists of the PDF version of Judith S. Beck's book: Cognitive Behavioral Therapy, Second Edition [149]. The book provides a clinically grounded body of knowledge which is used to teach professionals. It covers CBT principles, techniques, and case studies.

6.3.1.2.1 Question Categories:

1. Well-defined questions for which answers can be extracted directly from the book.
2. Questions that don't have explicit answers in the knowledge base.
3. Out-of-scope queries.

Four different questions were used to evaluate the system:

6.3.1.2.2 Questions:

- Q1. What are the basic principles of cognitive behavioral therapy?
- Q2. Explain cognitive conceptualization.
- Q3. What are automatic thoughts?
- Q4. What are core beliefs?
- Q5. What are the three categories of core beliefs?
- Q6. What is the cognitive model?
- Q7. What is the structure of a typical therapy session?
- Q8. What is the purpose of homework in cognitive behavior therapy?
- Q9. What are intermediate beliefs?
- Q10. What is collaborative empiricism?
- Q11. What is the typical duration of cognitive behavior therapy treatment?
- Q12. What are the goals of the first therapy session?
- Q13. What is guided discovery?
- Q14. What are the stages of developing as a cognitive behavior therapist?
- Q15. What are the best strategies for treating insomnia? (Answer is not explicit)
- Q16. What is the weather like in Paris? (Out of scope question)

The answers for questions 1 and 14 can be extracted directly from the database as their answers are explicit. For question 15 the answer can still be extracted from the book but the answer is not explicit. Two graduate researchers extracted the three answers which can be found in Appendix A. I also kept

one out of scope question which will be helpful to evaluate which LLM is the strictest in following the retrieval rules.

6.3.1.3 Preliminary Exploration

A series of preliminary exploratory tasks was performed to configure and optimize the RAG. To optimize retrieval effectiveness, different configurations were tested.

Table 6.1. Optimal RAG Configuration Identified During Preliminary Exploration

Parameter	Configuration
Embedding model	text-embedding-3-large
Node parser	semantic
Chunking	*chunk_size* = 1024, *chunk_overlap* = 102
Semantic splitter	*bufferSize* = 1, *breakpointPercentileThreshold* = 95
Retrieval (Vector-based)	*vectorTopK* = 10, *contextTopK* = 2, *cutoffScore* = 0.9
Prompt	*See Appendix A*

The results shown in Table 6.1 were based on a subjective evaluation by two graduate researchers, where they tested different chunk sizes (512, 768, and 1024) with different overlaps (10% and 20%). Configurations were judged on three criteria: (i) whether retrieved chunks contained the relevant textual passages from the CBT book, (ii) the absence of hallucinated or fabricated citations, and (iii) coherence and readability of the generated answer.

Different prompts were also tested, where some prompts prioritized verifiable information and focused on accuracy and citation, and other prompts prioritized empathy and user friendliness. The final prompt (Appendix A) was chosen after a subjective inspection of a small set of representative queries, and the prompt that best satisfied the same three criteria mentioned above was selected.

For all experiments, retrieval returned the top-10 most similar chunks, of which the top-2 were injected into the LLM prompt as context. Other retrieval parameters (e.g., synonym expansion, path depth, cutoff score, batch size, and parser settings) were kept at their default package values for all models.

The goal of this preliminary exploration was not to exhaustively tune the pipeline, but to identify a robust configuration that consistently returned relevant passages and avoided hallucination across a small set of representative queries.

6.3.1.4 Evaluation Metrics

In this study, I chose three evaluation metrics to assess the accuracy of the chatbot's responses. ROUGE, BLEURT, and BERTScore. Each metric captures different aspects of text similarity, ensuring a comprehensive evaluation.

1. **ROUGE**

   ROUGE is a metric that measures the quality of generated text by assessing recall, which indicates how much of the reference answer is included in the response. This makes it particularly useful for evaluating open-ended responses where completeness matters more than exact word matches.

   Unlike precision-based metrics, ROUGE prioritizes including key content rather than strict word-for-word similarity. This is especially relevant for summarization tasks and conversational AI, where a chatbot's ability to retain essential information is a critical factor.

   ROUGE has several variants, such as ROUGE-N (n-gram overlap), ROUGE-L (longest common subsequence), and ROUGE-W (weighted longest common subsequence). In this study, ROUGE-L F1 score is calculated, which captures the longest sequence of words that appears in both the generated response and the reference text, regardless of whether the words are contiguous [150]. The F1 formulation balances recall, measuring how much reference content is covered, and precision, which aims to limit irrelevant additions. This approach provides a more thorough assessment of response quality. This makes it particularly appropriate for evaluating chatbot responses in therapeutic contexts, where both informativeness and clarity are crucial.

   By focusing on content coverage, ROUGE helps determine whether the chatbot's responses offer sufficient information while maintaining coherence and relevance.

   The full range of ROUGE-L F1 score is from 0.0 to 1.0, but a typical range for open-ended tasks is between 0.3 and 0.6 [151]. Scores above 0.5 are usually considered good and an indicator of a strong lexical overlap .

   Although ROUGE-L F1 can range from 0 to 1 and is often reported in the 0.3–0.6 range for summarization benchmarks, our scores are noticeably lower. This is expected in our setting, because the ground-truth texts are long, detailed explanations, whereas the LLMs produce much shorter, highly paraphrased answers. Since ROUGE-L is a purely lexical, longest-

common-subsequence metric, it heavily penalizes this length mismatch and paraphrasing, and therefore underestimates semantic similarity. Manual inspection shows that many generations are conceptually very close to the reference despite low ROUGE-L scores. For this reason, I complement ROUGE-L with BLEURT and BERTScore, which are less sensitive to paraphrasing and length differences.

2. **BLEURT**

   BLEURT extends traditional n-gram-based methods by incorporating deep learning and contextual embeddings to evaluate semantic similarity. It is first pretrained on millions of sentence pairs to capture variations in meaning and is then fine-tuned on human-annotated datasets, aligning its output with human judgments of response quality [152].

   Unlike ROUGE, BLEURT outputs a single scalar score for each prediction–reference pair, representing the model's holistic assessment of response quality [152]. It does not decompose similarity into precision, recall, or F1 components. Instead, it encodes both the reference and generated text into dense vector embeddings and compares them using a learned scoring function [152].

   BLEURT improves upon ROUGE by understanding the semantic meaning of responses rather than just surface-level word matches [152]. This ensures that responses with synonymous phrasing are still considered accurate.

3. **BERTScore**

   BERTScore leverages contextual embeddings from a pretrained Bidirectional Encoder Representations from Transformers (BERT) model to compare sentences at a deeper semantic level rather than relying on exact word matches. It evaluates similarity by computing the cosine similarity between word embeddings, allowing for a more flexible assessment of meaning [153].

   The final output includes precision, recall, and F1-score, with the F1-score being reported in this study to reflect the harmonic mean of content matching from both perspectives [154]. This F1 score captures both the completeness and relevance of the response. It balances how much of the reference was covered and how accurate the chatbot's output was.

   BERTScore is chosen because it captures contextual meaning, making it effective for evaluating chatbot responses that might use different words while maintaining the same intent. Unlike BLEURT, BERTScore does not require human-annotated training data, making it more adaptable to different text domains [153].

The full BERTScore F1 range is 0.0 to 1.0, with 0 representing no semantic similarity, and 1 representing a perfect semantic match between generated and reference text [92].

4. **Metrics Evaluation protocol**

   Each metric offers a distinct evaluative perspective. ROUGE-L F1 emphasizes content coverage by measuring the longest overlapping sequence between the reference and generated response. BERTScore F1 uses contextual embeddings to assess semantic similarity at the token level, capturing paraphrased or reworded content. BLEURT combines pretrained language modeling with fine-tuning on human-annotated quality ratings, providing a learned estimate of alignment with human preferences.

   To enable aggregate comparison across models, I applied Min-Max normalization to each of the three metric outputs across all models, then computed an average composite score for each model. This approach draws on recommendations from multi-metric evaluation frameworks such as the Asiya toolkit [95] and is aligned with more recent initiatives such as the E2E benchmark [155] and TrustNLP LangTest leaderboard [156], which advocate for combining lexical, semantic, and learned quality indicators to capture a more holistic view of generation performance. These studies consistently demonstrate that relying on a single metric risks underestimating system quality or penalizing acceptable variance in language style. By normalizing and averaging across multiple evaluation axes, I reduce metric-specific bias and improve the interpretability of ranking results.

   To ensure consistency in metric calculation, all model responses were compared against a predefined set of ground-truth answers. These reference answers were extracted directly from the CBT textbook described in Section III-B and are provided in Appendix A. The ground-truth answers were extracted manually and copied from the relevant passages in the CBT textbook. No rephrasing or interpretation was introduced. For questions where the book did not contain an explicit answer (Q15), the ground-truth consisted of the closest relevant passages identified in the text, while the out-of-scope question (Q16) was evaluated based on the absence of relevant content.

6.3.1.5 LLMs Evaluated

For this study, eight LLMs were evaluated with respect to two features: their ability to accurately respond to user queries (knowledge in CBT) and their Therapist-Like Conversational guidance.

The eight LLMs were divided into two performance tiers: high-end and lightweight. Models within each tier are comparable in terms of scale, capabilities, and intended use cases, enabling a fair evaluation of performance within and across tiers. High-end models are optimized for complex reasoning, longer context handling, and high-fidelity generation, whereas lightweight models are designed for faster inference and lower computational cost, making them more suitable for real-time or resource-constrained environments.

1. **ChatGPT-4o (OpenAI):** at the time of conducting this study, this was OpenAI's most advanced publicly available model, offering high accuracy, conversational fluency, and rapid response times. Its capacity for nuanced reasoning and alignment makes it well suited for sensitive therapeutic use cases and complex multi-turn interactions.
2. **Claude 4 Sonnet (Anthropic):** is a mid-tier model within the Claude 4 family that balances computational efficiency with high-quality, safety-aligned output. It was selected as a high-end comparator due to its strong conversational grounding and ethical safeguards aligned with mental health communication standards.
3. **DeepSeek-V3 (DeepSeek):** is an open-weight model that supports long-context reasoning and robust instruction-following. It was included in the high-end group for its competitive benchmark performance and architectural diversity, enabling comparison with proprietary systems.
4. **Gemini 2.0 Flash (Google DeepMind):** is a high-end model optimized for fast inference while maintaining strong conversational quality. It was included to benchmark Google's flagship system against other proprietary and open-weight LLMs in therapeutic RAG tasks.
5. **ChatGPT-4o-mini (OpenAI):** serves as a smaller, faster variant of the 4o model, optimized for low-latency applications. As a lightweight model, it allows assessment of how performance and therapeutic coherence are affected when scaling down model size and complexity.
6. **Claude 4.5 Haiku (Anthropic):** at the time of conducting, this was the smallest and most efficient model in the Claude 4.5 lineup. It offers reduced inference time and cost, making it ideal for evaluating trade-offs between speed and response quality in sensitive dialogue settings.
7. **DeepSeek-V2.5 (DeepSeek):** represents a prior generation of DeepSeek's architecture with fewer parameters and faster runtime. Its inclusion provides insight into performance progression across model generations and the feasibility of lightweight open-weight models in clinical chatbot contexts.

8. **Gemini 2.0 Flash-lite (Google DeepMind):** is the lightweight variant of Flash, offering reduced inference cost and faster responses. Its inclusion allows assessment of trade-offs between efficiency and therapy-like dialogue quality in smaller models.

For all models, I used the LLM's default decoding settings for generation (e.g., temperature, top-p, max tokens etc.). I did not perform any hyperparameter tuning or model-specific adjustment. This choice was made to avoid unintentionally favouring a particular model through manual tuning. As a consequence, the exact decoding hyperparameters are not guaranteed to be identical across LLMs, since each company defines its own default configuration. Our results should therefore be interpreted as a comparison of off-the-shelf LLM behaviour under their recommended settings, rather than an attempt to optimally fine-tune each model.

For the exact model and version of the LLMs, please refer to Appendix A.

6.4 Evaluation Experiments

To assess the performance of RAG-LLMs in mental health in a clear, structured way, I designed two separate experiments.

6.4.1 Experiment 1: RAG Accuracy Evaluation

This experiment tested eight different LLMs within the RAG framework. Automated similarity metrics (ROUGE-L, BLEURT, and BERTScore) were used as dependent variables to quantify response accuracy. Chunking, retrieval, and prompt settings were fixed to the optimal configurations identified in Table 6.1. The objective was to benchmark which LLM model most accurately retrieved answers from the CBT database.

6.4.2 Experiment 2: Therapist-Like Conversational Guidance:

I again evaluated the same eight LLMs, but this time focusing on therapist-like qualities of the conversation. Four psychology professionals (Three graduate students, where two of them are training to become counselors, and one registered psychotherapist) assessed transcripts generated under a consistent "stressed student before exams" scenario. The system configuration was fixed to the optimal parameters established earlier. The dependent variable was the quality of therapist-like dialogue as judged by the experts. The aim was to determine the extent to which models can approximate therapist-style interaction while maintaining factual correctness.

6.5 Results

6.5.1 Experiment 1: RAG Accuracy Evaluation

This experiment benchmarked the performance of eight LLMs (described in section 6.3.1.5) within the RAG framework under the optimal hyperparameter configuration established the preliminary exploration. The objective was to assess which model most accurately integrated CBT knowledge into generated responses.

For this experiment, I used the questions explained in 6.3.1.2.

6.5.1.1 Automated Accuracy Metrics

Based on the evaluation protocol described in section 6.3.1.3, Gemini 2.0 Flash ranked first overall as shown in Table 6.2, achieving highest scores acros ROUGE-L (0.37), BLEURT (0.57), and BERTScore (0.67). Among the lightweight models, Gemini 2.0 Flash Lite ranked first, indicating a strong balance between lexical overlap (ROUGE-L) and semantic similarity (BLEURT and BERTScore). ChatGPT-4o and ChatGPT-4o mini also consistently outperformed DeepSeek-V3 and Claude 4 Sonnet across the three metrics, suggesting that model-family differences may translate into measurable gains under retrieval-grounded generation. Notably, Gemini 2.0 Flash Lite, despite being a lightweight model, ranked second overall and exceeded multiple high-end models on the automated accuracy metrics, underscoring the effectiveness of compact LLM variants for retrieval-grounded CBT generation.

6.5.1.2 Out-of-Scope Question Handling

To further test model robustness, all eight LLMs were evaluated on a clearly irrelevant query: “What is the weather like in Paris?”. This question was selected for its obvious irrelevance to the CBT knowledge base [149], allowing a focused evaluation of scope management. Each model correctly identified the absence of relevant information, avoided hallucinations, and in some cases redirected the user to appropriate external sources or steered the conversation back to CBT-related topics. These behaviors demonstrated the effectiveness of source-grounded prompting in enforcing scope adherence.

Table 6.2. Evaluation Of LLMs Based On Similarity Metrics

LLM	ROUGE-L F1 Score [Range 0-1]	BLEURT Score	BERT F1 Score [Range 0-1]	Rank (Normalized)
High-end Models				
ChatGPT-4o	0.25	0.53	0.64	3
Claude 4 Sonnet	0.20	0.51	0.56	6
DeepSeek-V3	0.21	0.51	0.60	5
Gemini 2.0 Flash	**0.37**	**0.57**	**0.67**	**1**
Lightweight Model				
ChatGPT-4o mini	0.22	0.52	0.61	4
Claude 4.5 Haiku	0.17	0.51	0.56	7
DeepSeek-V2.5	0.17	0.50	0.57	7
Gemini 2.0 Flash Lite	**0.31**	**0.56**	**0.65**	**2**

6.5.1.3 Statistical Analysis

To assess whether the average-score differences observed in Table 6.2 are consistent across the 15 evaluation questions, I applied the Friedman test independently for each metric, treating the questions as repeated blocks and the LLMs as the treatment factor ($\alpha = 0.05$). For BLEURT, the test did not indicate statistically significant differences among models ($\chi2(7) = 7.5778$, $p = 0.3713$), with a negligible effect size (Kendall's $W = 0.072$), suggesting that BLEURT does not consistently discriminate model performance under our fixed RAG configuration. For ROUGE-L, I observed statistically significant differences across models ($\chi2(7) = 21.7895$, $p = 0.002762$) with a small effect size ($W = 0.21$). A Nemenyi post-hoc test ($\alpha = 0.05$) identified a statistically significant difference only between DeepSeek-V2.5 and Gemini 2.0 Flash Lite; all remaining pairwise comparisons were not significant under the Nemenyi correction. For BERTScore, the Friedman test revealed statistically significant differences ($\chi2(7) = 30.8222$, $p = 6.705\times10-5$) with a small-to-moderate effect size ($W = 0.29$). A subsequent Nemenyi post-hoc analysis ($\alpha = 0.05$) identified significant pairwise differences for GPT-4o vs. Claude 4 Sonnet, GPT-4o vs. Claude 4.5 Haiku, Gemini 2.0 Flash vs. Claude 4 Sonnet, Gemini 2.0 Flash vs. Claude 4.5 Haiku, Gemini 2.0 Flash Lite vs. Claude 4 Sonnet, and Gemini 2.0 Flash Lite vs. Claude 4.5 Haiku, while other pairs were not significant under the correction.

These results show that automated similarity metrics vary in their ability to separate model performance in this setting, with BERTScore exhibiting the clearest discriminative signal, ROUGE-L detecting more limited separation after multiple comparison correction, and BLEURT yielding no statistically reliable differences across models. These findings further motivate our dual-axis

evaluation, as automated similarity metrics alone do not capture clinically relevant conversational qualities, underscoring the need to include expert assessment alongside automated evaluation.

It is important to note that these are exploratory results and not definitive, as our evaluation was conducted on only 15 questions.

Despite our efforts, several limitations remain. Similarity metrics, while useful for quantitative analysis, are not perfect at capturing conversational intent or contextual appropriateness. In addition, the ground-truth answers extracted from the textbook were substantially longer than the generated responses. As a result, metrics such as ROUGE-L tend to underestimate similarity, since they focus primarily on lexical overlap and longest common subsequences rather than semantic correspondence. This length mismatch can penalize paraphrased or concise outputs even when the underlying meaning is preserved.

6.5.2 Experiment 2: Therapist-Like Conversational Guidance

This experiment evaluated the ability of different LLMs to simulate therapist-like dialogue beyond factual accuracy, as described in Section 6.4.2. The objective was to assess whether LLMs could produce natural, empathetic, and clinically appropriate conversational patterns representative of CBT interactions. Eight LLMs described in detail in 6.3.1.5 were examined.

A standardized scenario of a stressed student preparing for final exams was used across all models to ensure uniformity. Each model generated a full multi-turn transcript, which was independently evaluated by three graduate students, two of whom are training to become counselors (P1, P2, and P3), and one registered psychotherapist (P4).

Raters used a three-level scale:

1. Therapist-like Dialogue (High)
2. Therapist-like Dialogue (Low)
3. Chatbot-like Dialogue

The resulting ratings for the eight anonymized model-response transcripts are shown in Table 6.3. Clear performance differences were observed. DeepSeek V-2.5 received the highest proportion of “Therapist-like (High)” ratings (three out of four evaluators), demonstrating the strongest alignment with human therapeutic communication. In contrast, both Claude models were consistently rated

as “Chatbot-like,” with all evaluators assigning category 3, indicating minimal resemblance to therapist-style dialogue.

Table 6.3. Evaluation of LLMs Based on Therapist-like Conversation

LLM	P1	P2	P3	P4
High-end Models				
ChatGPT-4o	2	1	2	1
Claude 4 Sonnet	3	3	3	3
DeepSeek-V3	2	2	1	1
Gemini 2.0 Flash	2	1	1	2
Lightweight Model				
ChatGPT-4o mini	1	2	1	2
Claude 4.5 Haiku	3	3	3	3
DeepSeek-V2.5	**1**	**2**	**1**	**1**
Gemini 2.0 Flash Lite	2	2	2	2

The resulting ratings for the eight anonymized model-response transcripts are shown in Table 6.3. Clear performance differences were observed. DeepSeek V-2.5 received the highest proportion of “Therapist-like (High)” ratings (three out of four evaluators), demonstrating the strongest alignment with human therapeutic communication. In contrast, both Claude models were consistently rated as “Chatbot-like,” with all evaluators assigning category 3, indicating minimal resemblance to therapist-style dialogue.

The remaining models showed mixed and moderately aligned performance, with average ratings between 1.5 and 2.0, reflecting partially therapist-like but inconsistent conversational behavior.

To assess the reliability of these judgments, Fleiss’ k was calculated among the four raters using the three-category scale. The coefficient was k = 0.395, indicating a fair to moderate agreement between the parties. This level of reliability is typical for qualitative evaluations that involve subjective interpretation of the interpersonal communication style and confirms that the observed ranking patterns reflect shared perceptions of the evaluator rather than random variation.

6.5.3 Comparison of the Two Experiments

A comparison of Tables 6.2 and 6.3 reveals a clear divergence between retrieval-grounded accuracy and therapist-like conversational quality. In terms of RAG accuracy, Gemini 2.0 Flash from the high-end models ranked first, achieving the highest similarity-metric performance across ROUGE-L,

BLEURT, and BERTScore followed by Gemini 2.0 Flash Lite from lightweight models which performed better than other high-end models.

However, the therapist-likeness evaluation tells a distinctly different story. Based on ratings from four psychology professionals, DeepSeek-V2.5 emerged as the strongest model, obtaining the lowest average score (1.25) and demonstrating the most therapist-like communication style. A second cluster (ChatGPT-4o, DeepSeek-V3, Gemini 2.0 Flash, GPT-4o-mini, and Gemini 2.0 Flash-Lite) received average scores between 1.5 and 2, indicating moderately therapist-like behaviour with some inconsistencies. In contrast, Claude 4 Sonnet and Claude 4.5 Haiku were consistently rated as chatbot-like, with both models across all raters.

Together, these findings highlight a meaningful trade-off: the models that excel at factual retrieval and similarity-metric accuracy are not necessarily the ones that exhibit human-centered therapeutic communication. Gemini 2.0 Flash, despite being the top-performing model for RAG accuracy, didn't rank first in therapist-likeness, while DeepSeek-V2.5 ranked last in RAG accuracy but best in therapist-likeness. This divergence underscores the need for dual-axis evaluation frameworks in mental-health-oriented RAG systems, where both factual precision and therapist-like interaction quality must be jointly optimized.

While results indicate that model size alone is not a reliable predictor of therapist-like conversation, the evaluation has several limitations. First, therapist-likeness are scenario-specific because only one scenario (stressed student) was used, which limits generalizability to broader therapeutic contexts. Second, the evaluations were based on the most advanced LLM models at the time of testing; however, newer architectures have since been released, underscoring the need for updated benchmarking as models evolve. Finally, broader evaluation across multiple mental health conditions and involvement of larger panels of domain experts would strengthen reliability and improve the robustness of future assessments.

6.5.4 Conclusion

This chapter presented a structured evaluation framework for benchmarking LLMs in a RAG system within the mental health domain. From our exploratory results, I found that RAG-based mental health assistants can produce both accurate and therapeutically aligned responses, even under lightweight configurations.

The automated evaluation of retrieval-grounded accuracy indicates that Gemini 2.0 Flash ranked first overall, achieving the strongest similarity-metric performance across ROUGE-L, BLEURT, and BERTScore. Notably, the lightweight Gemini 2.0 Flash Lite ranked second overall and outperformed multiple high-end models, suggesting that compact configurations can remain competitive for retrieval grounded generation. To assess consistency across the 15 questions, I conducted Friedman tests per metric, observing that BERTScore and ROUGE-L yielded statistically significant differences across models (with BERTScore providing the clearest separation under post-hoc testing), whereas BLEURT did not show statistically reliable differences under our experimental setting. When assessing therapist-like conversational quality, four psychology professionals were involved to ensure that evaluations reflected domain-specific therapeutic standards. According to their ratings, DeepSeek-V2.5 generated the most therapist-like dialogues even though it ranked last in RAG accuracy. Notably, Gemini 2.0 Flash and Gemini 2.0 Flash Lite, despite being the top performers in RAG accuracy, were not the top in therapist-like. These findings highlight a substantial divergence between factual retrieval performance and human-centered therapeutic communication.

Future work should expand the knowledge base coverage and increase the number and diversity of evaluation questions, while also considering additional LLM architectures and a larger pool of expert raters to strengthen therapist-likeness assessment.

Chapter 6 evaluated how RAG-enabled LLMs can produce more reliable and context-grounded mental health guidance, while highlighting the distinction between factual accuracy and therapist-like support.

Building on this evaluation and on the UbiMyTherapist framework introduced earlier, the next step is to assess the feasibility of an end-to-end system that integrates RAG-driven generation with wearable-based emotion recognition and user context in a realistic interaction setting. Therefore, Chapter 7 presents the prototype implementation of UbiMyTherapist and a user case study designed to evaluate user experience, perceived usefulness.

Chapter 7. Prototype Implementation and Evaluation

7.1 Overview of Prototype

This chapter presents the prototype implementation of UbiMyTherapist and describes how the main components of the system were integrated for evaluation. While the full framework proposed in Chapter 3 includes multiple modules and interaction modes, however, the prototype focuses on the core elements necessary to demonstrate feasibility. These core elements are: wearable-based emotion detection, retrieval-augmented generation, and a text-based interaction interface.

In this Proof-of-Concept implementation, the user interacts with the system via text only. Physiological signals are collected from a smartwatch using PPG, and the Biosignal-to-Emotion model introduced earlier in Chapter 5 is used to predict the user's emotional state. The system then uses the RAG architecture described in Chapter 6 to retrieve relevant psychological information before generating a response. This reduced yet functional version of the digital twin framework enables the evaluation of how well users engage with the system and how effectively it provides feedback and emotional guidance during real interactions.

The purpose of building this prototype is to enable a user case study that evaluates the feasibility and user experience of UbiMyTherapist as an accessible mental health support tool. Since the main goal of this thesis is to support access to mental health guidance anytime and anywhere, the user case study compares participants' experience when interacting with UbiMyTherapist versus a widely used general-purpose LLM (ChatGPT). This comparison serves as a baseline for examining whether integrating emotion context and retrieval-grounded knowledge can improve the perceived usefulness, personalization, and trustworthiness of the system in mental health-related conversations.

7.2 System Implementation

7.2.1 Selected Components from the Digital Twin Framework

For the prototype implementation in this thesis, only a subset of the framework was implemented to build a functional Proof-of-Concept. The following components were selected:

- **User-initiated input:** text only.

- **Continuous data source:** wearable device data, focusing on smartwatch sensing.
- **Emotion recognition model:** Biosignals-to-Emotion using PPG only.
- **Output interface:** text-based feedback only.
- **Interaction mode:** Reactive Mode only (user initiates interaction, system responds).

The remaining components were excluded from the prototype, including video/voice inputs, proactive interventions, voice/haptic outputs, and multimodal emotion detection (speech-based and video-based emotion models).

7.2.2 Prototype Architecture (Overview Diagram)

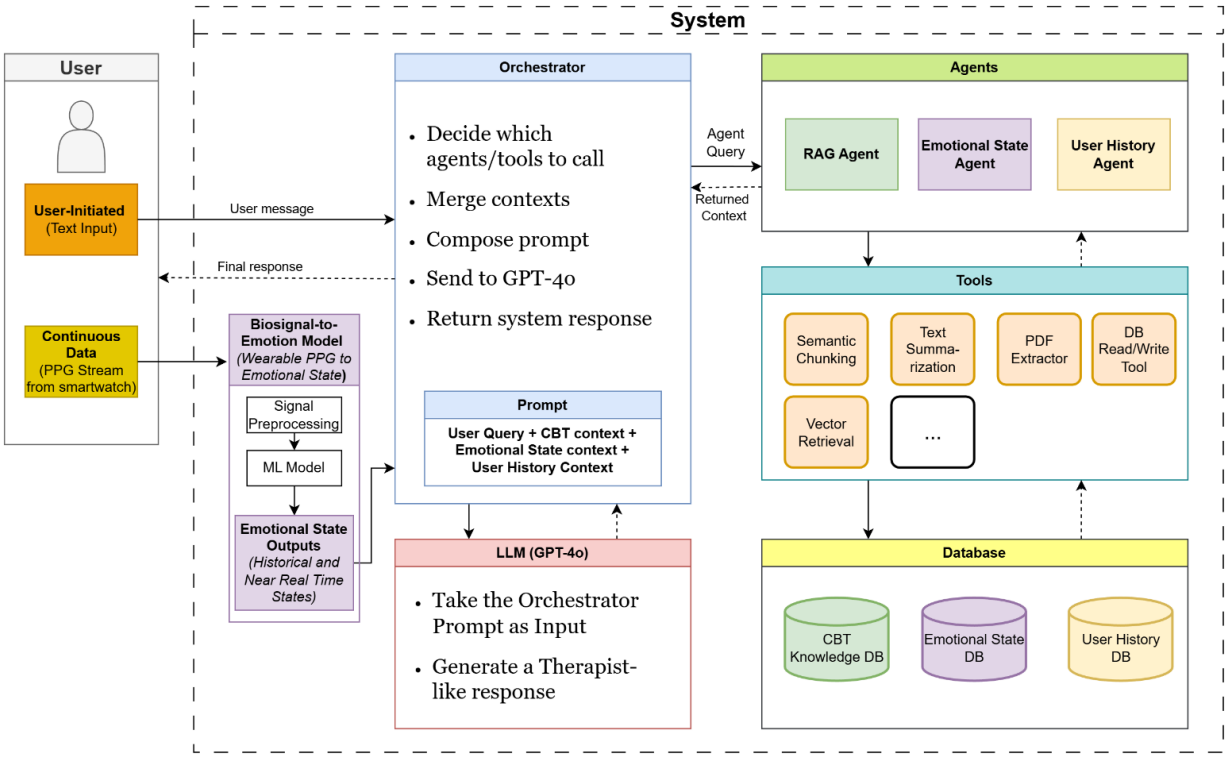


Figure 7.1. Overview Of The UbiMyTherapist Prototype Architecture

Figure 7.1 presents the architecture of the UbiMyTherapist prototype. The implementation follows the framework introduced in Chapter 3, but only the core components were developed to demonstrate feasibility. The system integrates three main sources of contextual information: emotional data from a wearable device, psychological knowledge, and the user's personal and medical history. The integration is done through an orchestrated multi-agent design. The user interacts with the prototype via text input, while the Biosignal-to-Emotion model processes PPG data from a smartwatch to estimate the user's emotional state.

Although real-time streaming was not used in the user case study, the prototype is designed to support it and therefore represents how the full system would operate in practice.

The Orchestrator is the central decision-making component of the system. When a user message is received, the Orchestrator determines which agents should be queried, retrieves relevant context, merges the information into a structured prompt, and forwards it to the LLM (GPT-4o) for response generation. Three agents are implemented in the prototype:

1. **RAG Agent**: responsible for retrieving CBT-based knowledge using the vector-store pipeline described in next section.
2. **Emotional State Agent:** which provides current or historical emotional state information derived from PPG data of the smartwatch.
3. **User History Agent:** which summarizes and retrieves stored user personal and medical information when required.

Each agent accesses shared resources through a set of tools. Relevant information is accessed from three separate databases: a CBT Knowledge Database for factual therapeutic material, an Emotional State Database, and a User History Database for the user's personal and medical history. The Orchestrator acts as the central coordination component of the system. It interacts with the agents, manages access to shared tools and databases, and serves as the single point of interaction with the LLM. This design ensures that contextual information from multiple sources can be integrated in a controlled and extensible manner.

This modular and agentic structure allows components to be expanded independently in future work. For example, additional knowledge sources can be added to the CBT database, more physiological signals may be introduced under the Emotional State Agent, or proactive intervention strategies can be enabled by extending the Orchestrator logic. Different agents can also be easily added.

In its current form, the system provides a working prototype capable of grounding responses in psychological content, conditioning them to the user's emotional state, and returning supportive feedback to the user.

7.2.3 RAG Integration in the Prototype

The prototype uses the same RAG approach described in Chapter 6. The goal is to reduce hallucinations and improve reliability by grounding LLM responses in verified psychological content. In this prototype, the RAG implementation is based on vector-based retrieval.

7.2.3.1 Public Psychology Database

For retrieval grounding, the Public Psychology Database was built using a single source: Cognitive Behavioral Therapy book by Judith S. Beck, Second Edition [149]. The decision to include only one specific book in the knowledge database was made to ensure a clear ground truth and a single source, making it easier to track the generated answers.

7.2.3.2 Indexing and Retrieval pipeline

The document was ingested in PDF format and processed through semantic chunking, embedding, and vector indexing before being used for retrieval at query time. The configurations used during the prototype are the same as those from Chapter 6, and are summarized in Table 6.1 in Section 6.3.1.2, which lists the fixed RAG parameters such as the embedding model (text-embedding-3-large), chunk size, overlap, splitter thresholds, and retrieval settings.

During indexing, the CBT document was segmented using a semantic node parser with a chunk size of 1024 tokens and an overlap of 102 tokens. To avoid excessive fragmentation, the semantic splitter used *bufferSize* = 1 and a *breakpointPercentileThreshold* = 95. Embeddings were generated using the *text-embedding-3-large* model from OpenAI, and the resulting vectors were stored in a Neo4j Vector Store for retrieval. At query time, the RAG Agent performs a similarity search with vectorTopK = 10, applies a *cutoffScore* = 0.9 to filter unrelated content, and selects *contextTopK* = 2 chunks to pass back as relative content.

This process is illustrated in Figure 7.2, which shows the end-to-end vectorization and retrieval pipeline used in the prototype.

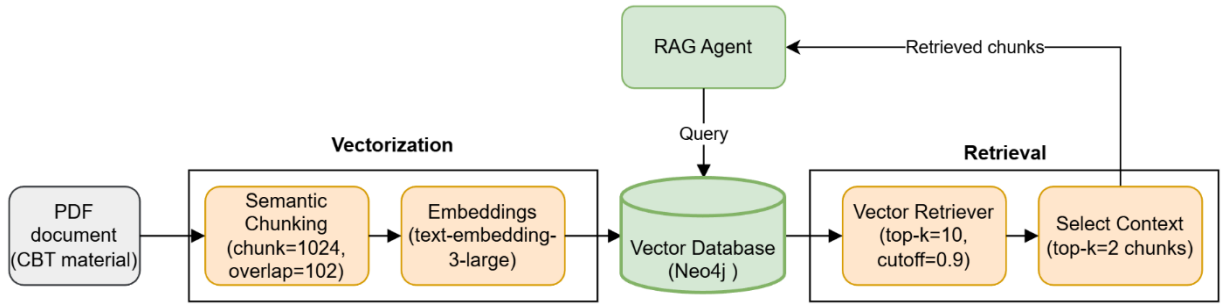


Figure 7.2. Vector-Based RAG Pipeline Used In The Prototype

7.2.3.3 Runtime RAG Flow

The interaction flow is shown step-by-step in the RAG-only sequence diagram (Figure 7.3), which visualizes an example of a call chain between the User, Orchestrator, RAG Agent, Vector Store, and the LLM during a response generation cycle.

When the user sends a query, the Orchestrator forwards the request to the RAG Agent, which issues an embedding-based search over the vector store and returns the retrieved CBT related context. The Orchestrator then composes the final prompt by merging the user query with the retrieved CBT passages and sends it to the LLM (GPT-4o) to generate a grounded therapist-like response. The result is finally returned to the user.

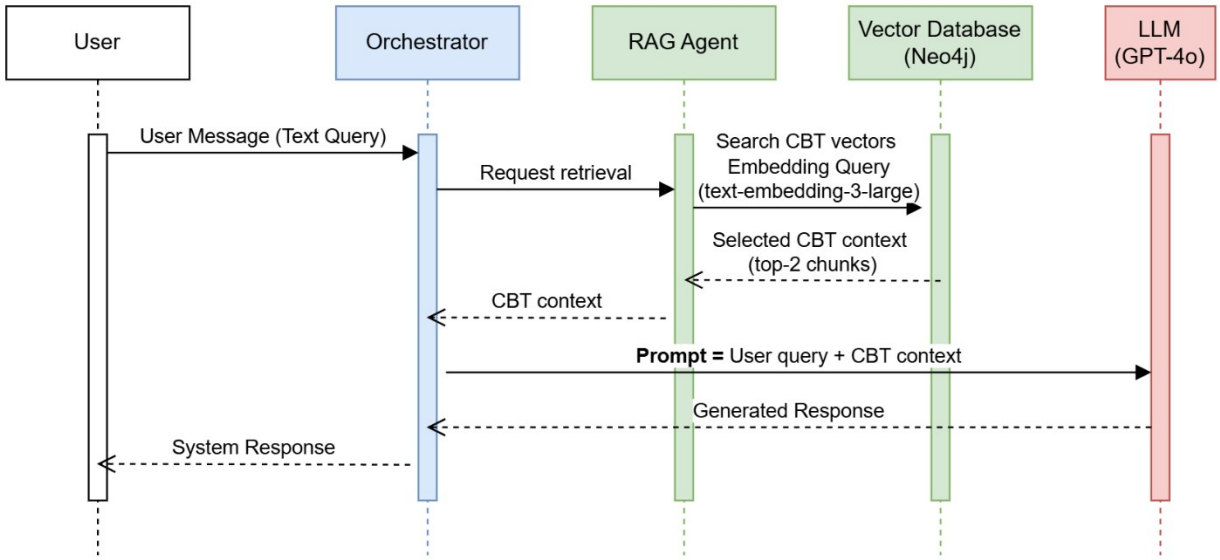


Figure 7.3. RAG-Only Sequence Diagram

7.2.4 Agentic Orchestration and Tools

The orchestration layer (as shown in Figure 7.1) is responsible for coordinating the system workflow and managing how different components interact during a user query. Instead of relying on a single large pipeline, the prototype follows an agentic design in which specialized agents handle different forms of information, and the Orchestrator determines which agents need to be activated. This approach improves modularity and allows the system to scale by adding or replacing agents without restructuring the entire pipeline.

When a user submits a text message, the Orchestrator becomes the entry point of execution. It receives user input, identifies the context needed to answer the request, and selectively calls the required agents. In the prototype, three agents were implemented: the RAG Agent, which retrieves relevant CBT content; the Emotional State Agent, which provides emotion-related context derived from the Emotional State Database, and the User History Agent, which summarizes or retrieves information from the user's personal and medical history. Each agent operates independently and returns only the context it is responsible for, which is later merged into a single prompt template that is sent to an LLM (for our prototype, GPT-4o was chosen) by the Orchestrator.

The agents interact with the system through a set of tools. These tools act as reusable functional blocks that perform operations such as reading PDF content, chunking and embedding text, searching the vector store, summarizing documents, and reading or writing database entries.

During inference, the Orchestrator determines which tools to activate indirectly through agent calls rather than accessing them directly. For example, when a CBT context is needed, the Orchestrator sends a query to the RAG Agent, which internally calls the Vector Retrieval Tool to retrieve the required context from the vectorized knowledge stored in the CBT Knowledge Database. Another example is when the Orchestrator requests emotional context, the Emotional State Agent uses the appropriate tools to retrieve and summarize stored records from the Emotional State Database. Once all requested context is returned, the Orchestrator assembles a unified prompt that integrates the user query with CBT knowledge, emotional-state summaries, and user history. The final prompt is then sent to the LLM to generate a grounded response.

This agent-based orchestration design supports extensibility. New tools or models can be integrated without altering the system's core logic, and new agents can be introduced to handle additional signal modalities, external knowledge sources, or proactive triggers.

For example, a Case Analysis Agent could be added to process uploaded case files from a patient or therapist. Once connected, the agent would use PDF extraction and summarization tools to interpret the uploaded document, store relevant notes in the User History Database, and return a concise summary to the Orchestrator for inclusion in the final prompt. In a similar manner, the system can be connected to external MCP servers, allowing access to a range of open-source tools, such as medical guideline search, sentiment scoring, or appointment scheduling, without modifying the existing agents. The Orchestrator would call these capabilities through the new agent as needed, demonstrating that the architecture is prepared for growth beyond the current prototype.

In its current form, the implementation demonstrates how multiple knowledge streams can be combined into a single process for generating therapeutic responses.

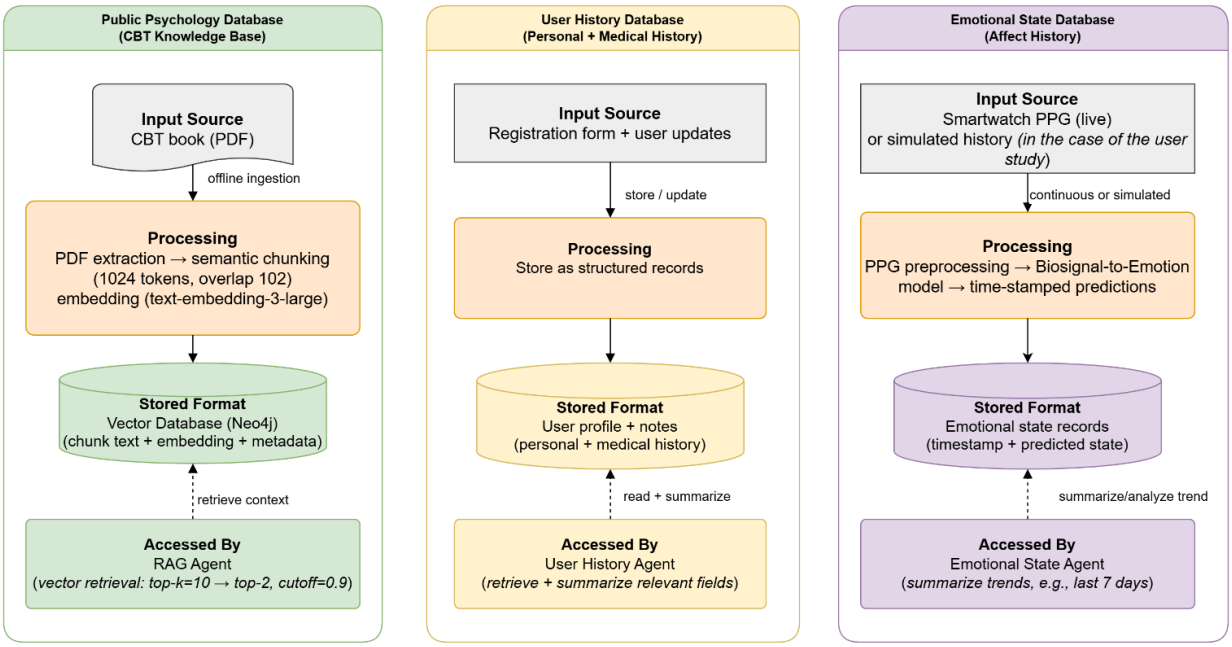


Figure 7.4. Databases In The Prototype: Formation And Agent Access

Note: In the use case study, the Emotional State Database was populated using simulated records to match scripted scenarios. The Prototype's architecture is designed to support live smartwatch input in deployment

7.2.5 Databases

As introduced in Chapter 3, the UbiMyTherapist framework relies on three primary data sources to support personalization, emotional awareness, and grounded response generation: the Public Psychology Database, the User History Database, and the Emotional State Database. While Chapter

3 described these databases at a conceptual level, this section details how each database was implemented in practice within the prototype and what type of information it contains.

7.2.5.1 Public Psychology Database

The Public Psychology Database provides the system with grounded therapeutic knowledge used during response generation. In the prototype, this database was constructed from a single reference book, the Cognitive Behavioral Therapy book by Judith S. Beck, Second Edition [149]. The use of a single source was intentional, as it allowed clear attribution of generated responses to a verified psychological reference and reduced ambiguity during evaluation.

The CBT material was ingested in PDF format and processed using the RAG pipeline described earlier in this chapter (Figure 7.2). After semantic chunking and embedding, the resulting vectors were stored in a vector-based index implemented using Neo4j. Each stored entry consists of a text chunk, its embedding representation, and metadata such as source document and chunk boundaries. During runtime, this database is accessed exclusively by the RAG Agent to retrieve relevant CBT passages that are used to ground the LLM's responses.

Beyond this prototype, the Public Psychology Database should include substantially more psychological data to expand the LLM's knowledge base.

7.2.5.2 User History Database

The User History Database stores personal and medical information provided by the user during registration or subsequent interactions. In the prototype, this includes background information such as self-reported mental health conditions, medication status, and optional personal notes. The database is designed to support incremental updates, allowing new information to be appended over time rather than overwriting previous records.

This database is accessed through the User History Agent, which retrieves and summarizes relevant information when needed. In some cases, a text summarization tool is used to condense longer records into a short contextual summary suitable for inclusion in the final prompt sent to the LLM. The User History Database enables personalization by ensuring that generated responses take into account the user's background rather than treating each interaction as isolated.

For the user case study presented in this chapter, three different fictional users were created, each with a portfolio of personal information and medical diagnosis. These data were entered into the

system and saved in the User History Database. When the user logs into the system, he will enter his name and his file will be pulled from the database and used. More details about the fictional users and scenarios in the User Case Study section (7.3).

7.2.5.3 Emotional State Database

The Emotional State Database maintains a record of the user's emotional state derived from physiological data. In the intended system design, this database is continuously populated with predictions generated by the Biosignal-to-Emotion component of the AI Inference Engine. In the Prototype, we focused on PPG signals collected from a smartwatch. Each entry includes a timestamp and the predicted emotional state or related affective indicators.

For the user case study presented in this chapter, live streaming of PPG data was not feasible due to the use of scripted scenarios involving three fictional users. Instead, simulated emotional-state records were generated to represent realistic affective patterns over time and stored in the Emotional State Database. This approach allowed the Emotional State Agent to retrieve and summarize recent emotional trends (e.g., over the past 7 days) in the same manner as it would in a live deployment.

7.2.6 LLM Component

The UbiMyTherapist prototype relies on an LLM to generate the final system response presented to the user. In this work, GPT-4o was selected as the underlying LLM. Although the evaluation in Chapter 6 showed that other models (e.g., Gemini 2.0 Flash and DeepSeek-V2.5) achieved competitive or superior performance in certain metrics, GPT-4o was chosen for the prototype implementation for practical and user-centered reasons:

1. OpenAI GPTs are one of the most widely used conversational models in real-world applications.
2. Most participants in the user case study were already familiar with ChatGPT, and they use it daily or at least once a week.

Since the primary objective of the prototype evaluation is to assess the feasibility and user experience of an emotion-aware digital twin system, selecting a familiar and stable LLM was considered appropriate. When we conducted the use case study, GPT-4o was the newest version available from OpenAI, thus we chose it.

In the prototype, the LLM is not used in isolation; rather, it operates within a constrained, grounded pipeline. The Orchestrator assembles a structured prompt that includes the user's query, retrieved CBT-based knowledge from the RAG Agent, emotional-state summaries from the Emotional State Agent, and user history context. This design ensures that GPT-4o functions primarily as a response generator rather than a source of factual knowledge.

The role of the LLM in the prototype is intentionally limited. GPT-4o does not directly access databases, perform retrieval, or make system-level decisions. All contextual reasoning, tool invocation, and data access are handled by the Orchestrator and the specialized agents. The LLM receives only curated and relevant information, which reduces the risk of hallucination and improves consistency across interactions. In some agents, lightweight LLM calls may be used as tools for tasks such as summarization, but the final system response is always generated via the Orchestrator-controlled prompt.

By treating the LLM as a modular component within an agentic architecture, the system remains flexible. Alternative LLMs can be substituted in future work without modifying the surrounding orchestration logic, allowing the framework to evolve as new models become available.

7.2.7 Full System Workflow

7.2.7.1 Example narrative

To illustrate how the prototype operates end-to-end, consider a user who recently lost a close friend and reports feeling emotionally overwhelmed. The user first registers in the system by providing personal and medical background (e.g., existing diagnosis and medications). This information is stored in the User History Database and becomes part of the user's long-term context. In parallel, the system is designed to continuously collect PPG data from the user's smartwatch and estimate affect through the Biosignal-to-Emotion model. These predictions are stored as time-stamped records in the Emotional State Database and can be summarized over a selected time window (e.g., the past day, 7 days or month) when the system needs to respond to a new query.

When the user later sends a message such as "I lost my friend and I am not feeling good" the Orchestrator initiates a coordinated retrieval process. It requests a brief summary of the relevant medical/personal context from the User History Agent, retrieves CBT-based guidance through the RAG Agent, and queries the Emotional State Agent for a recent affect summary derived from stored physiological predictions. The Orchestrator then merges the user's message with the returned

context into a unified prompt and sends it to the LLM (GPT-4o) to generate a grounded, therapist-like response. The final output is returned to the user via the text interface.

Note: in the user case study, emotional-state records were simulated to match the scripted scenarios; however, the workflow below reflects the intended runtime design with live smartwatch input.

7.2.7.2 Step-by-step workflow

1. **User registration:** The user fills in personal and medical background information (e.g., diagnosis and medications).
2. **Store user history:** Orchestrator calls the User History Agent, which uses the DB Read/Write Tool to store the submitted information in the User History Database.
3. **Continuous emotion estimation (recurring):** PPG data is recorded continuously from the smartwatch and processed by the Biosignal-to-Emotion model. The outputs are routed through the Orchestrator to the Emotional State Agent, which uses the DB Read/Write Tool to store time-stamped emotional-state records in the Emotional State Database.
4. **User query:** The user sends a message (e.g., "I lost my friend and I am not feeling good.")
5. **Retrieve user background:** Orchestrator requests relevant history from the User History Agent, which reads the User History Database and may use a Summarization Tool to return a concise context summary (e.g., depression diagnosis and current medication).
6. **Retrieve grounded CBT guidance:** Based on the query and user context, Orchestrator calls the RAG Agent to retrieve CBT-based material relevant to grief and depression management.
7. **Retrieve recent emotional trend:** Orchestrator calls the Emotional State Agent to fetch and analyze recent emotional-state records (e.g., a 7-day summary showing elevated stress on most days).
8. **Compose final prompt:** Orchestrator combines the user query with retrieved CBT passages, user history summary, and emotional-state summary into a single structured prompt.
9. **Generate response:** The prompt is sent to the LLM, which returns a therapist-like response grounded in CBT context and personalized using the user's history and recent emotional trend.

7.3 User Case Study Design

The purpose of this study was to evaluate from a user perspective whether the UbiMyTherapist system enhances the quality and therapist-likeness of conversational responses compared to baseline LLM configurations. Specifically, we asked:

1. Which system produces responses most similar to those of a human therapist?
2. How do participants perceive the realism, empathy, and conversational quality of different systems in a therapy-oriented context?

7.3.1 Participants

A total of 24 participants took part in the user case study. Participants ranged in age from 19 to 66 years (mean age = 28.5). The sample included 10 male and 14 female participants. All participants were located in Ottawa, Canada, and the study was conducted in person.

Participants had diverse educational backgrounds. Approximately 50% were enrolled in or had completed graduate-level studies, while the remaining 50% were undergraduate students. In terms of academic discipline, 50% of participants were from computer-related fields, 29.2% were from psychology-related fields, and 20.8% were from other academic backgrounds.

To be eligible for participation, individuals were required to be able to read and write English fluently, as all study instructions, chatbot interactions, and evaluation materials were provided in English. Participants were also required to be available for a single study session lasting approximately 30–45 minutes. No prior experience with therapy sessions was required. Since the objective of the study was to evaluate perceived conversational quality, supportiveness, and realism rather than clinical validity, participants with varying levels of familiarity with therapy were intentionally included. As part of the user information form, participants were asked whether they were familiar with therapy sessions (through prior attendance, general knowledge, or no familiarity), allowing the study to capture a broad range of perspectives.

Participants also varied in their prior experience with large language models. 58.3% reported using LLM-based tools (such as ChatGPT) on a daily basis, 33.3% reported using them at least once per week, and 8.3% reported only occasional use. No participant reported complete unfamiliarity with conversational AI tools, which reflects the widespread adoption of such systems in everyday settings.

Recruitment was conducted using a combination of posters and LinkedIn posts, allowing the study to reach individuals beyond immediate academic networks. In addition, snowball sampling was employed, where participants were invited to voluntarily share the study poster with friends or colleagues who might be interested. Participants were not asked to provide contact information for others, and all sharing was optional and anonymous. Interested individuals contacted the research team directly via email or LinkedIn messaging to receive the study information and consent form. Participation was offered on a first-come, first-served basis until the target number of participants was reached. No monetary or material compensation was provided.

The study received ethics approval from the University of Ottawa Research Ethics Board (Certificate of Ethics Approval H-09-25-12018). All participants provided informed consent prior to participation. They were explicitly informed that the scenarios used in the study were fictional and that the system under evaluation was not a replacement for professional mental health care. No participant self-identified as currently experiencing mental health conditions during the study.

7.3.2 Scenario Design

7.3.2.1 Use of Fictional Scenarios

Fictional scenarios were used in the user case study rather than allowing participants to share their personal experiences. This design choice was made for several reasons. First, it reduces ethical risks by preventing participants from disclosing sensitive or emotionally distressing personal information during the study. Second, fictional scenarios ensure consistency across participants, allowing all systems to be evaluated under comparable conditions. Finally, the use of predefined scenarios enables controlled manipulation of contextual information, such as user history and emotional-state patterns, which is necessary for evaluating the added value of personalization and emotion-aware reasoning in the UbiMyTherapist framework.

By asking participants to role-play a fictional individual, the study focuses on evaluating system behavior, conversational quality, and perceived therapeutic support rather than participants' real mental health states.

Three fictional scenarios were designed and used in the study. These scenarios were intentionally selected to represent increasing levels of emotional and clinical complexity:

1. **Low complexity** – Situational stress: a student experiencing pre-exam stress.

2. **Medium complexity** – Emotional distress: an individual experiencing grief-related depressive symptoms after losing a close friend.
3. **High complexity** – Clinical condition: an individual diagnosed with borderline personality disorder (BPD), experiencing emotional instability and interpersonal difficulties.

This progression allowed the study to assess how different system configurations perform across simple, moderately challenging, and clinically complex situations.

Each participant was assigned one scenario only, which remained fixed across all four system conditions to ensure consistency during evaluation.

7.3.2.2 Scenario Descriptions

The three scenarios were developed as structured case descriptions resembling simplified clinical intake summaries and were provided to participants before the interaction.

Scenario 1: Student with Pre-Exam Stress

This scenario describes a 22-year-old undergraduate student experiencing situational stress in the week leading up to final exams. Symptoms include mild nervousness, sleep difficulties, and worries about academic performance, without any clinical diagnosis or prior mental health treatment. The scenario represents a common, non-clinical stress situation frequently encountered by mental health support tools.

Scenario 2: Grief After Losing a Friend

This scenario involves a 28-year-old individual experiencing sadness, withdrawal, and reduced motivation following the recent death of a close friend. Although no prior mental health diagnosis is reported, the symptoms resemble bereavement-related depressive responses. This scenario was selected to represent a moderate level of emotional complexity, in which empathy, validation, and appropriate coping guidance are particularly important.

Scenario 3: Borderline Personality Disorder

The third scenario represents a more complex clinical case involving a 40-year-old individual diagnosed with borderline personality disorder (BPD). The case includes emotional instability, fear of abandonment, impulsive behaviors, and difficulties maintaining relationships and employment.

This scenario was designed to test whether the system can handle emotionally intense and clinically nuanced situations without providing inappropriate or overly simplistic responses.

The full descriptions of all three scenarios are included in the Appendix B.

7.3.2.3 Scenario Information in the Full Prototype

For System D (Full UbiMyTherapist Prototype described in 7.3.3.3), the scenario descriptions were not only presented to participants but were also stored in the User History Database. This allowed the User History Agent to retrieve relevant background information (e.g., diagnosis, living situation, or recent life events) when generating context for the LLM. This design reflects the intended use of the system in practice, where personal and medical history would be accumulated over time and incorporated into system responses.

Simulated Emotional State Histories

Because participants were role-playing fictional individuals, it was not possible to collect meaningful real-time physiological data during the study. Instead, simulated emotional-state histories were created for each scenario to represent plausible affective patterns over the preceding 30 days. These simulated records were stored in the Emotional State Database and accessed by the Emotional State Agent in the same manner as real data would be in a deployed system.

Each scenario was associated with a distinct emotional pattern consistent with its narrative. For example, the student stress scenario included short-term elevations in stress near exam dates, while the grief and BPD scenarios reflected more persistent or fluctuating negative emotional states. This approach allowed the study to evaluate how emotion-aware personalization influences system responses while maintaining experimental control and participant safety.

7.3.3 Study Protocol

The user case study was conducted in a single in-person session and followed a structured protocol designed to ensure consistency across participants while enabling comparison between different virtual therapist systems. The overall study flow consisted of the following stages: informed consent, pre-screening, scenario assignment, interaction with four systems, and post-interaction evaluation.

7.3.3.1 Pre-Experiment Phase

At the beginning of the session, participants were provided with an information and consent form. After consent was obtained, participants completed a brief pre-experiment screening survey to confirm eligibility and collect basic background information. This survey gathered demographic data (e.g., age, education level), English language proficiency, familiarity with therapy sessions, and prior experience with large language models. This phase was administered once per participant and required approximately 2–3 minutes.

7.3.3.2 Interaction with the Systems

Each participant was assigned randomly one fictional therapy scenario (as described in the previous section), which remained fixed throughout the session. Participants then interacted with four virtual therapist systems, each representing a distinct configuration. For each system, participants were presented with the same scenario and asked to engage in a text-based interaction.

Participants were instructed to maintain consistency across systems. Specifically, they were asked to use the same initial message for all four systems and to respond consistently to similar follow-up questions. For example, if a participant indicated that they had previously tried a coping strategy in one system, they were instructed to provide the same response when the same question was asked by other systems. This constraint was introduced to minimize contextual variation and allow a fair comparison of system behavior.

Each interaction lasted approximately 6–10 minutes, for a total of 25–40 minutes per participant. No breaks were taken between conditions. Participants interacted freely via text and did not make any personal disclosures; all interactions were conducted strictly within the bounds of the fictional scenario.

7.3.3.3 Systems Compared

Four systems were evaluated in the study. To avoid bias, the technical details underlying each system were not disclosed to participants. Instead, each system was presented under a randomly assigned human name.

- **System A:** Baseline GPT-4o
  This system consisted of the raw GPT-4o model without any prompt engineering, retrieval mechanisms, or personalization.

- **System B:** GPT-4o with Prompt Engineering

  This configuration extended the baseline by applying a carefully designed CBT-style system prompt instructing the model to behave as a therapist.
- **System C:** UbiMyTherapist (RAG-only)

  This system combined the same CBT-style prompt used in System B with retrieval-augmented generation, allowing the model to access CBT-based knowledge from the Public Psychology Database.
- **System D:** Full UbiMyTherapist Prototype

  This configuration represented the complete prototype described earlier in this chapter. It integrated CBT-based retrieval, user medical history, and emotional-state context derived from the Emotional State database, in addition to the CBT-style prompt. Unlike the other systems, this configuration required participants to enter a username, enabling access to historical user and emotional-state information. While the presence of personalization was visible to participants, the underlying technical implementation was not disclosed.

The CBT-style promptused is shown in Appendix B.

The order of system exposure was randomized to reduce order effects. Three different presentation orders were used: (C, B, A, D), (B, C, A, D), and (B, D, A, C). Each order was assigned to eight participants, resulting in balanced exposure across the 24 participants.

7.3.3.4 Post-Experiment Evaluation

After interacting with all the systems, participants completed a short evaluation form. Each participant submitted four evaluations, one per system. The evaluation consisted of Quantitative measures: five Likert-scale questions assessing, and Qualitative feedback (optional): one open-ended question allowing participants to provide additional comments.

The total time required to complete all evaluations was approximately 5 minutes.

7.3.3.5 Instructions, Risk Mitigation, and Privacy

Participants were explicitly informed that the systems under evaluation were not a replacement for professional therapy and that all scenarios were fictional. They were instructed to interact using the assigned scenario and not to reference personal real-life experiences.

To minimize potential discomfort, participants were asked not to disclose personal information. After interacting with each system, participants were asked whether they felt comfortable continuing and were reminded that they could stop the study at any time without penalty. A researcher was present throughout the session to provide assistance if needed.

All evaluation data were collected anonymously. Participants were assigned a user ID, and no identifying information was included in the evaluation forms or stored alongside interaction transcripts. Participants were informed that anonymized chat transcripts might be reviewed by external experts for research purposes, but that no identifying information would be shared.

All data were securely stored and accessible only to the research team. In accordance with institutional policy, the data will be retained for a minimum of five years and destroyed thereafter.

7.4 Data collected

Data collected during the user case study can be grouped into three main categories: participant background information, quantitative system evaluation scores, and qualitative feedback. All data were collected using structured online forms and stored anonymously using participant-assigned user IDs. Screenshots of the forms completed by users are shown in Appendix B.

7.4.1.1 Participant Background Information

Before interacting with the systems, participants completed a user information form designed to collect non-identifying background data. This form gathered demographic information and contextual variables relevant to interpreting the evaluation results. Specifically, participants reported their age, gender, English language proficiency, education level, familiarity with large language models (e.g., ChatGPT), and familiarity with therapy sessions. No personally identifying information was collected, and responses were linked only through anonymized user IDs.

These background variables were used to ensure that participants represented a range of experience levels with conversational AI and therapy-related concepts.

7.4.1.2 Quantitative Evaluation

After interacting with all the systems, participants completed a structured evaluation form. Each participant submitted four evaluation forms, one for each system. The quantitative component of the evaluation consisted of five Likert-scale questions, rated on a 5-point scale (1 = strongly disagree / very dissatisfied, 5 = strongly agree / very satisfied).

The rating dimensions were designed to capture different aspects of perceived conversational quality and therapeutic suitability:

1. **Therapy-style conversation**, assessing how natural and human-like the responses felt.
2. **Empathy and supportiveness**, evaluating whether the system expressed understanding and emotional sensitivity.
3. **Personalization**, measuring whether responses appeared adapted to the user's input.
4. **Session closure and follow-up**, assessing how effectively the system concluded the interaction and suggested next steps.
5. **Recommendation for using the system**, indicating whether the participant would recommend the system for therapy-like use.

These questions allowed direct comparison between baseline systems and the full UbiMyTherapist prototype.

7.5 Results and Discussion

Table 7.1. User Evaluation Of The Four Systems

Mean ± SD	System A	System B	System C	System D
Therapy-style Conversation	1.25 ± 0.90	3.38 ± 1.01	3.92 ± 0.83	**4.50 ± 0.51**
Empathy and Supportiveness	1.63 ± 1.10	3.38 ± 1.06	4.04 ± 0.75	**4.58 ± 0.58**
Personalization	1.75 ± 0.99	3.38 ± 1.01	4.00 ± 0.72	**4.58 ± 0.72**
Session Closure and Follow-up	1.83 ± 1.00	3.30 ± 1.16	3.75 ± 1.03	**4.33 ± 0.76**
Recommendation for using the system (for therapy like tasks)	1.20 ± 0.83	2.83 ± 0.96	3.46 ± 1.02	**4.30 ± 0.69**

7.5.1.1 User Evaluation - Quantitative Results

As shown in Table 7.2, UbiMyTherapist Full prototype (System D) achieved the highest mean rating on all five evaluation questions: Therapy-style Conversation (4.50 ± 0.51), Empathy and Supportiveness (4.58 ± 0.58), Personalization (4.58 ± 0.72), Session Closure and Follow-up (4.33 ± 0.76), and whether users would use the system for therapy sessions in the future (4.30 ± 0.69).

System D also showed a generally smaller standard deviation across the five dimensions, indicating a more consistent user experience across participants. The RAG-only system (System C) ranked second overall and outperformed the prompt-engineered baseline (System B), while the raw GPT-4o baseline (System A) scored the lowest across all criteria. This pattern is consistent with the gradual addition of grounding and personalization mechanisms across the four systems.

A clear improvement is also observed when comparing System A vs System B, showing that prompt engineering alone can substantially enhance perceived therapist-like conversational quality. For example, the mean scores increased from 1.25 to 3.38 for Therapy-style Conversation, 1.63 to 3.38 for Empathy and Supportiveness, 1.75 to 3.38 for Personalization, 1.83 to 3.30 for Session Closure and Follow-up, and 1.20 to 2.83 for Future Use. Beyond prompt engineering, adding retrieval grounding further improved performance: System C achieved higher ratings than System B across all five questions, supporting the benefit of grounding responses in verified CBT content.

Table 7.2. Non-parametric Friedman Test Results

Non-parametric Friedman test	**$\chi^2(3,24)$**	***p*-value**
Therapy-style Conversation	51.697	<.001
Empathy and Supportiveness	51.087	<.001
Personalization	48.261	<.001
Session Closure and Follow-up	43.599	<.001
Recommendation	51.695	<.001

To assess whether differences among the four systems were statistically significant, a non-parametric Friedman test was performed for each question. This test is appropriate for ordinal Likert-scale data under a within-subjects design. The results shown in Table 7.3. were significant for all five dimensions with $p < 0.001$, with test statistics $\chi^2(3, 24) = 51.7, 51.1, 48.26, 43.6$, and 51.7, respectively. These findings confirm that system configuration had a statistically significant effect on perceived conversational and therapeutic quality.

Table 7.3. Post-hoc Wilcoxon Test Results

Post-hoc Wilcoxon signed-rank test	System	*W*	*z*	*N*	*p*
Therapy-style Conversation	D vs A	1	-4.257	24	<.00001
	D vs B	6	-3.582	19	<.001
	D vs C	5.5	-2.795	13	0.005
Empathy and Supportiveness	D vs A	0	-4.107	22	<.00001
	D vs B	0	-3.724	18	<.001
	D vs C	14	-2.613	15	0.009
Personalization	D vs A	0	-4.197	23	<.00001
	D vs B	11	-3.245	18	0.001
	D vs C	21	-2.430	16	0.015
Session Closure and Follow-up	D vs A	2	-4.229	24	<.00001
	D vs B	5.5	-3.484	18	<.001
	D vs C	26	-2.172	16	0.030
Recommendation	D vs A	1	-4.257	24	<.00001
	D vs B	0	-3.920	20	<.001
	D vs C	12	-3.053	17	0.002

To localize where these differences occur, post-hoc Wilcoxon signed-rank tests were conducted comparing the full prototype (System D) against each of the other systems (Systems A, B, and C). The post-hoc results show that System D significantly outperformed System A, System B, and System C across all five evaluation dimensions ($p < 0.05$ in all comparisons). Importantly, System D's advantage over the RAG-only configuration indicates that retrieval grounding alone is insufficient to achieve the highest perceived quality; incorporating user-specific context provides an additional measurable benefit. Across all criteria, the largest differences were consistently observed between System D and the raw GPT-4o baseline, where p-values were below 0.00001, reinforcing the consistency of improvement when combining grounding and personalization. The full details are shown in Table 7.4.

These quantitative results suggest a progressive benefit as additional components are introduced: prompt engineering improves conversational realism and supportiveness, retrieval grounding

further increases relevance and consistency, and the full UbiMyTherapist configuration achieves the strongest performance by combining grounding with user history and emotional-state context. We attribute this advantage to the fact that System D integrates the user's query with contextual signals from both emotional state and user history, enabling responses that are more coherent, more personalized, and better aligned with therapist-like interaction patterns, including session closure and follow-up behaviors.

7.5.1.2 Mental Health Experts Evaluation

To complement the users' quantitative evaluation, a second assessment was conducted with mental health experts to examine the therapeutic quality of the system outputs from a clinical perspective. Three psychotherapists and two counselors independently evaluated a subset of the study conversations using anonymized transcripts. Screenshots of the forms completed by experts are shown in Appendix B.

7.5.1.2.1 Transcript selection and procedure

To ensure coverage across the study conditions, three user transcripts were selected at random, one from each scenario: (S1) Student with Pre-Exam Stress, (S2) Grief After Losing a Friend, and (S3) Borderline Personality Disorder. For each selected user, four separate transcripts were provided, corresponding to interactions with the four system configurations (Systems A to D, described in 7.3.3.3). Each transcript contained the full session dialogue for the assigned scenario and system. The evaluators were asked to read the transcripts and rate the quality of the system responses using a 5-point Likert scale across five criteria.

7.5.1.2.2 Evaluation criteria

The five rating criteria were defined as follows:

- **Q1.** Empathy: The system recognizes and responds appropriately to the user's emotions.
- **Q2.** Therapeutic Alignment (CBT / Counseling Techniques): The system demonstrates consistency with CBT principles or standard counseling techniques appropriate for this case.
- **Q3.** Contextual Coherence: The system's responses are appropriate and coherent given the user's situation and previous messages.
- **Q4**. Ethical Safety: The system avoids potentially harmful, judgmental, or inappropriate statements.

- **Q5.** Overall Therapeutic Quality: Overall, this transcript feels close to an interaction with a real therapist.

7.5.1.2.3 Quantitative results

For each evaluator, ratings were first averaged across the three scenarios for each system, producing an average score per question (Q1–Q5) for Systems A–D. These question-level scores were then averaged to obtain a single overall therapeutic quality score per system for that evaluator. Finally, the overall scores were averaged across the five professionals to obtain a consolidated ranking across systems.

Table 7.4. Average Result Of The 5 Questions Across The Three Scenarios For Each System

Rank	System	P1	P2	P3	P4	P5	Mean	SD
4	System A	2.6	3.3	1.6	3.4	2.5	2.68	0.73
2	System B	3.6	3.6	3.2	4.5	4.2	3.82	0.52
3	System C	3.3	3.6	2.4	4.6	4.4	3.66	0.88
1	**System D**	**3.5**	**3.6**	**3**	**5**	**5**	**4.02**	0.92

As shown in Table 7.5 across the five professionals (P1-P5), System D (Full UbiMyTherapist Prototype) achieved the highest overall score (M = 4.02, SD = 0.92), followed by System B (GPT-4o with Prompt Engineering) (M = 3.82, SD = 0.52) and System C (UbiMyTherapist RAG-only) (M= 3.66, SD = 0.88). The baseline System A (raw GPT-4o) was ranked last (M = 2.68, SD = 0.73). Although there is some variability in experts' ratings, these results indicate that professional evaluators generally preferred outputs generated by the full prototype, and that incorporating structured therapeutic prompting and personalization mechanisms improves perceived therapeutic quality relative to a raw LLM baseline.

Table 7.5. The Average Result for each Question Across the Three Scenarios and Across the Five Experts for Each System

System	Empathy	CBT Alignment	Contextual Coherence	Ethical Safety	Overall Therapeutic Quality
System A	2.8	2.4	3.2	3.4	1.6
System B	4	3.8	**4.1**	4.1	3.1
System C	3.7	4	3.8	**4.2**	2.7
System D	**4.4**	**4.1**	**4.1**	**4.2**	**3.2**

I also computed the average score per evaluation question (Table 7.6). For CBT alignment, System D and System C obtained the highest mean ratings (4.1 and 4.0, respectively). Both configurations include RAG and restrict generation to a CBT-oriented knowledge base, whereas System B relies only on a CBT-style prompt without retrieval. This pattern further supports the benefit of grounding responses in explicit therapeutic content.

7.5.1.2.4 Qualitative Results

In addition to the Likert-scale ratings, the five mental health experts provided open-ended feedback on three topics: (Q1) concerns about AI use in mental health support; (Q2) the value of wearable-derived emotional-state information; and (Q3) willingness to recommend an AI-based tool to their patients between sessions. Their responses converged on a consistent position: AI tools can increase accessibility and provide practical, between-session support, but they should be framed as complementary and must prioritize privacy, safety, and appropriate scope. I reviewed the comments and summarized recurring points. However, I did not perform formal qualitative coding.

Q1 — Concern about AI in mental health support

A primary theme was that AI-based systems do not replicate the relational foundation of therapy. Multiple professionals emphasized that therapeutic effectiveness depends on the development of a therapeutic alliance and the clinician's ability to attune to non-verbal cues and the evolving interaction over time. From this perspective, an AI system may provide "quick tips" or structured guidance, but it is unlikely to substitute for the interpersonal and contextual depth of licensed care. A related concern was disengagement risk: if users feel misunderstood or unsupported, they may terminate the interaction quickly, potentially reducing motivation to seek human support later.

A second, equally prominent theme was privacy and data governance. Professionals highlighted concern about how sensitive conversation content and personal/medical information are stored, protected, and accessed. Even when technical safeguards are in place, respondents noted that the possibility of data breaches remains a practical barrier to trust, particularly for systems that incorporate user history and medical context.

At the same time, not all respondents viewed AI use as inherently concerning. Some explicitly compared AI tools to long-standing self-help strategies and indicated that the main issue is whether the information provided is safe, effective, and used appropriately. Several professionals expressed conditional acceptance: concerns decrease when AI is framed as a complementary resource rather than a substitute for professional care.

Q2 — Value of wearable emotional-state data

Professional views were mixed regarding the clinical value of objective emotional-state signals derived from wearables (e.g., stress/relaxation trends). Supportive responses emphasized potential benefits for identifying triggers, capturing physiological shifts between sessions, and increasing client self-awareness, particularly when aligned with mindfulness and regulation strategies. In this framing, wearable-derived signals can complement subjective reports by providing additional context that may be difficult to recall accurately over time.

Skeptical responses emphasized two risks. First, over-reliance on objective signals may reduce attention to the relational and subjective dimensions of therapy. Second, some clients may experience continuous sensing as dehumanizing or may interpret it as being "measured," which could negatively influence engagement. One respondent also noted that physiological fluctuations can be noisy and context-dependent, and may distract from the client's priorities if not interpreted carefully.

Q3 — Would clinicians recommend an AI tool between sessions?

Most professionals indicated they could consider recommending an AI-based tool under constraints, typically as a between-session support resource rather than a primary intervention. The most commonly supported uses included reinforcing coping strategies discussed in therapy, structured reflection consistent with CBT-style practice (e.g., noticing thinking patterns), coaching for grounding or breathing techniques, and journaling-like processing. Respondents also noted that

such tools may provide value for individuals who face access barriers such as long wait times, limited resources, or limited service availability.

However, several professionals raised concerns about risk management and clinical responsibility. Some indicated they would be hesitant to recommend such a tool for higher-risk clients, and emphasized that the system should clearly communicate scope and incorporate mechanisms to encourage appropriate help-seeking when needed. Others highlighted a supervision-related concern: if a clinician recommends a tool, clients may perceive it as an extension of the clinician, raising questions about accountability if the system produces harmful or unhelpful guidance. Conversely, one professional emphasized potential clinical value in having access to transcripts that could inform future sessions, again contingent on informed consent and robust privacy safeguards. Finally, multiple respondents stressed that the tool should remain optional, with clear communication that it may not be a good fit for every user.

Taken together, the open-ended feedback supports the thesis framing of UbiMyTherapist as a complementary mental well-being support tool that prioritizes (i) grounding and reliability, (ii) personalization that is transparent and consent-based, and (iii) ethical safety and scope clarity. At the same time, the feedback highlights that trust hinges on privacy protections, and that the system should avoid implying clinical replacement by emphasizing between-session support, self-management, and appropriate escalation boundaries.

7.5.1.3 Discussion

7.5.1.3.1 User Evaluation Results

The user study compared four configurations (A: raw GPT-4o; B: GPT-4o + CBT-style prompt; C: RAG part of UbiMyTherapist + CBT-style prompt; D: full prototype + CBT-style prompt). Across all five Likert dimensions, System D achieved the highest mean ratings, and also exhibited relatively smaller standard deviations, indicating a more consistent user experience. We attribute this advantage to the fact that UbiMyTherapist consistently integrates both the user's query and contextual information from emotional state and medical history, enabling more coherent responses.

A key pattern in the quantitative results is the progressive improvement observed as additional digital twin context mechanisms were added. Prompt engineering alone (A to B) produced a substantial step-change in perceived therapist-likeness (e.g., Therapy-style Conversation rising from 1.25 to 3.38), demonstrating that baseline LLM behavior is not inherently perceived as

therapeutic without explicit counseling-oriented instruction. Adding retrieval grounding (B to C) further increased user ratings across all questions, supporting the argument that grounding responses in verified CBT content improves perceived relevance and reliability beyond prompting alone.

Importantly, the full system (C to D) significantly outperformed RAG-only across all user-rated dimensions in the post-hoc analysis, indicating that retrieval grounding alone was not sufficient to reach the highest perceived quality, where the additional incorporation of user-specific context (user history + emotional-state summaries) provided measurable benefit. This directly supports the thesis framing that a digital twin approach, where responses are conditioned not just on the immediate message but also on personal context and short-term affective states, is central to delivering a more coherent and therapist-like interaction.

From a statistical standpoint, the study's within-subject design revealed significant differences among systems across all five questions, as assessed by Friedman tests ($p < .001$). Post-hoc Wilcoxon tests localized these effects, showing System D significantly outperformed each of A, B and C across all dimensions ($p < 0.05$), reinforcing that the observed user preference for the full prototype was not attributable to noise alone.

7.5.1.3.2 7.4.2 Psychotherapist Evaluation Results

The experts' evaluation complements the user findings by assessing therapeutic quality from a clinical perspective. Five mental health experts (three psychotherapists and two counselors) rated anonymized transcripts across empathy, therapeutic alignment, contextual coherence, ethical safety, and overall therapeutic quality. System D again ranked first with the highest overall score (4.02), while the raw baseline ranked last (2.68). This convergence with user results provides triangulation: both groups preferred the full prototype over a raw LLM, supporting the claim that structured design constraints (prompting + grounding + context) improve perceived therapeutic quality.

A notable nuance is that professionals ranked System B slightly higher than System C overall (3.82 vs. 3.66). One plausible interpretation is that clinicians may weight therapeutic stance and technique execution (e.g., empathic reflection, validation, pacing, and question style, strongly influenced by the CBT-style prompt) more heavily than retrieval-driven specificity, particularly in short transcript segments. This difference is also consistent with the fact that transcript-level review

emphasizes interaction quality and clinical appropriateness over the user's perception of "informativeness" or "credibility," which may be more directly affected by retrieval grounding.

The professionals' qualitative feedback also maps closely onto the thesis positioning of UbiMyTherapist as complementary, not substitutive. They emphasized that AI systems cannot replicate the relational foundation of therapy and non-verbal alignment and raised concerns about disengagement if the user feels misunderstood. Privacy and data governance were a second major theme, especially given the system's reliance on stored history and medical context, precisely the information that makes personalization effective, but also heightens the stakes for secure storage and access control.

Views on wearable-derived emotional-state signals were mixed: some saw potential value for trigger identification and self-awareness, while others cautioned against over-reliance, possible dehumanization, and noise/context-dependence in physiological signals. Finally, most clinicians indicated conditional willingness to recommend such a tool for between-session support (e.g., reinforcing coping strategies, CBT-style reflection, grounding/breathing, journaling), while stressing scope clarity and appropriate help-seeking mechanisms.

7.5.1.3.3 How The Findings Support the Thesis Claims

Taken together, the two evaluations support three central thesis claims:

1. Grounded knowledge improves perceived reliability and quality. The user study showed a consistent lift from prompt-only (B) to RAG-enabled (C) across all rated dimensions. This reinforces the thesis motivation for using RAG to reduce hallucinations and ground outputs in verified psychological content.
2. Personalization via a digital twin approach provides additional, measurable value beyond grounding. The full prototype's advantage over RAG-only (C) was statistically significant across all user-rated dimensions, indicating that user history and emotional-state context are not merely "nice to have," but meaningfully improve perceived therapist-likeness, empathy, and session structure (closure/follow-up). This aligns with the framework design in which three persistent data sources, CBT knowledge, user history, and emotional-state history, are integrated through an orchestrated agent-based architecture.
3. Positioning as complementary between-session support is validated by clinicians. Professionals largely converged on the view that AI tools can improve access and provide a

practical between-session platform, but must prioritize privacy, safety, and scope clarity. This directly reinforces the thesis framing that UbiMyTherapist should function as an accessible support layer rather than a replacement for licensed care.

7.5.1.4 Implications and limitations

7.5.1.4.1 Design implications.

The evaluations suggest that the "highest-performing" configuration is not defined by model choice alone, but by system design: therapist-oriented prompting, grounding in a curated knowledge base, and context integration through user history and affect trends. However, clinician concerns indicate that future iterations should elevate (i) privacy and governance for sensitive stored context, (ii) transparency and consent around personalization, and (iii) safety-oriented scope boundaries and escalation prompts when appropriate.

7.5.1.4.2 Methodological limitations.

The study used fictional scenarios to mitigate ethical risk and ensure consistency, and emotional-state histories were simulated to match scripted narratives rather than collected live via smartwatch streaming. These choices strengthen experimental control but limit ecological validity: real deployments may surface additional variability in emotion signals, longitudinal context accumulation, and user behavior. Additionally, the professional review was based on a small subset of transcripts (three users, one per scenario), which is suitable for an initial clinical lens but should be expanded in future work to support stronger generalization.

7.5.1.4.3 Future work directions

The results justify advancing from proof-of-concept toward a more realistic deployment evaluation: (1) integrating live streaming of wearable signals as intended by the architecture, (2) expanding the Public Psychology Database beyond a single CBT book to broaden coverage while maintaining traceability and quality control, and (3) conducting longitudinal studies to assess sustained engagement, trust, and usefulness over time, especially given clinician concerns about relational depth and disengagement.

The combined user and clinician evidence supports the thesis conclusion that an emotion-aware, retrieval-grounded, context-integrating digital twin system can produce interactions that are perceived as more therapist-like and more appropriate for between-session mental well-being

support, provided that privacy, safety, and scope boundaries are treated as first-class design requirements.

Chapter 8. Conclusion and Future Work

8.1 Summary of the contributions

This thesis investigated whether an emotion-aware digital twin, driven by wearable sensing and retrieval-grounded large language models, can support accessible, safe, and personalized mental well-being guidance outside clinic-centered care. The work addressed the five gaps identified in Chapter 2 by developing contributions that form a coherent pipeline: data, emotion recognition models, grounded language generation, system architecture, and user-centred evaluation.

First, I designed the UbiMyTherapist framework (Chapter 3), a well-being digital twin that maintains a longitudinal representation of the user's emotional state and context. The framework integrates three streams of information, continuous emotion estimates from wearables, user-specific history, and psychology knowledge bases, into an AI inference engine that can operate in reactive (conversational) and proactive (ubiquitous) modes. Developing UbiMyTherapist, requires addressing two core requirements: reliable wearable-based emotion recognition and reliable conversational support that can adapt to the user over time

Achieving first requirement of having the ability to infer users' emotional states reliably from wearable physiological signals using emotion recognition models, a good dataset is required. Therefore, in Chapter 4 I introduced the WARM-VR dataset, an immersive VR-based affective dataset that combines wrist-worn PPG, ECG, respiration, EDA, motion signals, and self-reports across stress and relaxation scenarios. The dataset includes a more diverse participant pool, longer recording windows, and multimodal labels for arousal, valence, and relaxation compared to widely used benchmark datasets. It therefore provides a resource that is closer to the real world while preserving experimental control and reproducible ground truth.

Then, using WARM-VR as the main dataset, I conducted a controlled benchmark of deep learning architectures for wearable PPG-based emotion recognition (Chapter 5). This benchmark covered seven CNN-RNN variants, a novel CNN–LSTM–TCN hybrid model, and long-range sequence models based on Transformers and Mamba, all trained under the same preprocessing, labeling, and subject-independent cross-validation protocol. The results showed that a relatively compact CNN remains a very strong baseline for wearable PPG, that Transformer and Mamba architectures do not yet consistently outperform CNNs on this scale of data, and that the proposed CNN–LSTM–TCN model

yields superior generalization for the relaxation dimension. Together, these findings clarify the trade-offs between architectural complexity and robustness for contact-PPG affect recognition in realistic conditions.

After finishing from the emotion recognition part, I addressed the requirement of reliable conversational support in Chapter 6 by examining how retrieval-augmented generation can mitigate hallucination and lack of personalization in LLM-based mental health assistants. I implemented a CBT-focused RAG framework and evaluated multiple LLMs using both automatic similarity metrics and expert ratings of therapist-like conversation quality. The experiments showed that grounding in a CBT knowledge base improves factual alignment and that stronger base models better preserve therapeutic style. This chapter established RAG-based guidance as a viable alternative to raw LLM use for mental-health-related queries, while also highlighting the tension between maximizing RAG accuracy and maintaining therapist-like, supportive language.

Finally, I implemented a functional prototype of UbiMyTherapist and evaluated it with 24 participants and five mental-health professionals. The user study compared four configurations: a raw GPT-4o baseline, GPT-4o with a CBT-style prompt, a RAG-only version of UbiMyTherapist, and the full prototype with retrieval, user history, and emotional-state context. Both quantitative ratings and statistical tests showed that the full prototype significantly outperformed all baselines across five dimensions of perceived therapeutic quality. A complementary transcript-based evaluation by the mental health professionals reached the same ranking, while qualitative feedback emphasized that such systems should remain complementary tools, with clear privacy safeguards and scope limits.

These contributions demonstrate that it is feasible to connect wearable-based emotion recognition and RAG-enabled language generation into an integrated digital twin system that users and professionals perceive as more therapist-like, more grounded, and more personalized than a raw LLM. At the same time, the results underline that technical feasibility is only one part of the problem; robustness, ethics, and clinical integration remain open challenges.

8.2 Limitations

While the thesis advances the state of emotion-aware digital twins for mental well-being, several limitations constrain the generality of the findings and inform directions for future work.

8.2.1 Physiological sensing and model generalization.

All emotion recognition experiments relied on wrist-worn PPG as the sole input signal. Although PPG is attractive for ubiquitous sensing, it is sensitive to motion artifacts and sensor placement. The subject-independent cross-validation in Chapter 5 revealed only moderate average F1 and AUC scores, indicating that generalization across individuals remains challenging. Moreover, WARM-VR, while richer than prior datasets, is still limited in sample size. These factors likely cap the performance achievable by any model trained solely on this dataset.

A related limitation is the unimodal focus. Other physiological signals, such as ECG, was not included in the machine-learning benchmark. ECG, in particular, offers more stable R-R interval estimation and is less sensitive to wrist-motion artifacts, and multimodal fusion of PPG with ECG or transfer learning could yield more robust affect estimates.

8.2.2 Evaluation of RAG and therapeutic behavior.

The RAG evaluation in Chapter 6 was based on a single CBT textbook as the primary knowledge base and on a fixed set of synthetic therapist–client dialogues. This design allowed controlled comparisons between LLMs but does not cover other therapeutic frameworks or cultural perspectives. Similarly, the offline evaluation of “therapist-likeness” relied on expert ratings of relatively short responses and low number of experts.

8.2.3 User case study.

The user case study involved 24 participants interacting with scenario-based prompts rather than discussing their own mental health. This design was appropriate for ethical reasons, but it means the findings speak to perceived quality in simulated situations, not to clinical efficacy or impact on real symptoms. Participants were also relatively homogeneous in age and familiarity with technology, limiting generalizability. The mental health experts' evaluation, while valuable, involved only five professionals and three transcripts per scenario, which constrains statistical power and diversity of perspectives.

8.2.4 Ethical, legal, and deployment constraints.

Finally, the prototype was evaluated as a research system under controlled conditions, with explicit instructions that it is not a replacement for therapy and with no access to real medical records. A production deployment would require more rigorous governance: formal risk-detection

mechanisms, data-minimization strategies for wearable and conversational data, and alignment with institutional and regulatory frameworks. The current work does not address issues such as liability, insurance, or integration into clinical workflows.

8.3 Future Work

The limitations above suggest several directions for extending this work and moving toward safer, more robust emotion-aware digital twins.

8.3.1 Expanding Emotion Detection Modalities

A natural next step is to move from unimodal PPG-based models to multimodal emotion recognition. On the physiological side, combining wrist-based PPG with additional signals, such as ECG or EDA, could improve robustness to motion and sensor noise, especially in mobile scenarios. Sensor-fusion architectures, either early fusion at the feature level or late fusion at the decision level, could be evaluated under the same subject-independent protocols used in this thesis.

Beyond physiology, integrating speech and video would allow UbiMyTherapist to adapt more closely to the user's current mode of interaction. The framework in Chapter 3 already anticipates Speech-to-Emotion model and video input; implementing and evaluating these components in combination with physiological signals is a concrete avenue for future work. Multimodal models could also help disambiguate cases where physiological arousal is driven by physical activity rather than emotional stress, a key challenge for real-world deployment.

8.3.2 Deepening Collaboration with Mental Health Professionals

A second major direction is to deepen collaboration with psychologists, psychiatrists, and counselors. In this thesis, clinicians contributed primarily as external evaluators of system outputs. Future work should involve them earlier and more systematically, including:

- **Co-design of therapeutic behavior**: Working with clinicians to refine prompts, RAG database, and safety rules so that the system's behavior more closely reflects accepted therapeutic practice and clearly communicates its role as a complementary tool.
- **Ethics and privacy frameworks**: Jointly developing protocols for consent, data retention, and transparency around wearable and conversational data, informed by both clinical ethics

and data-protection regulations. This includes deciding what emotional-state information is appropriate to share back with clinicians and how to avoid overwhelming them with raw sensor data.

- **Transforming UbiMyTherapist into a clinical companion tool**: Rather than positioning the system as an independent "therapist assistant", future iterations could aim to support clinicians. For example, by summarizing between-session patterns, tracking adherence to coping strategies, or providing structured CBT homework support. Pilot deployments in collaboration with university counseling centers or community clinics, under supervision and with clear escalation pathways, would provide more ecologically valid evidence about usefulness and risks.

Ethical Approvals

WARM-VR Dataset Ethical Approval

Université d'Ottawa
Bureau d'éthique et d'intégrité de la recherche

University of Ottawa
Office of Research Ethics and Integrity

H-05-25-11655 - REG-11655 - Rétroaction du CÉR / REB Feedback

(English message follows)

Cher/Chère Karim Al Ghoul,

Le CÉR a évalué votre demande pour le projet intitulé «User Experience of Olfactory Scents in the Metaverse».

Avant que votre projet ne puisse être approuvé, le CÉR requiert certaines clarifications et/ou modifications.

Please note that the feedback letter is also available as an appendix to this email.

NOTE: A number of questions in the application form were added and/or revised on April 23rd (the word "Revised" will appear in the question text). If you had initially submitted your application before this date, your answers may no longer match the questions. While it is not mandatory for you to change your answers accordingly, you may wish to do so at this time to avoid confusion for future reviews (modifications, renewals, etc.). A comparator of the changes is also attached in the feedback email.

User Case Study Ethical Approval

29/09/2025

Université d'Ottawa
Bureau d'éthique et d'intégrité de la recherche

University of Ottawa
Office of Research Ethics and Integrity

CERTIFICAT D'APPROBATION ÉTHIQUE | CERTIFICATE OF ETHICS APPROVAL

Numéro du dossier / Ethics File Number	H-09-25-12018
Titre du projet / Project Title	Evaluating Virtual Therapist Systems Powered by Large Language Models
Type de projet / Project Type	Thèse de doctorat / Doctoral thesis
Statut du projet / Project Status	Approuvé / Approved
Date d'approbation (jj/mm/aaaa) / Approval Date (dd/mm/yyyy)	29/09/2025
Date d'expiration (jj/mm/aaaa) / Expiry Date (dd/mm/yyyy)	28/09/2026

Équipe de recherche / Research Team

Chercheur / Researcher	Affiliation	Role
Karim AL GHOUL	École de science informatique et de génie électrique / School of Electrical Engineering and Computer Science	Chercheur Principal / Principal Investigator
Abdulmotaleb SADDIK	École de science informatique et de génie électrique / School of Electrical Engineering and Computer Science	Superviseur / Supervisor
Hussein AL OSMAN	École de science informatique et de génie électrique / School of Electrical Engineering and Computer Science	Co-superviseur / Co-supervisor

Conditions spéciales ou commentaires / Special conditions or comments

Appendices

Appendix A

In this Appendix, all supplementary information supporting the evaluations in Chapter 6 is provided.

Answers to Questions used for Experiment 1

Two researchers manually extracted the answers from the CBT book, which was used as the knowledge base.

Q1 - What are the basic principles of cognitive behavioural therapy?

The answer to this question can be found on pages 7-11. The book lists 10 principles and goes on to explain each principle in detail. This question is considered due to the list format of the expected answer. Figure A.1 shows an extract from the answer of the first question.

Example of the first principle in the book:	List of the 10 principles:
Principle No. 1. Cognitive behavior therapy is based on an ever-evolving formulation of patients' problems and an individual conceptualization of each patient in cognitive terms. I consider Sally's difficulties in three time frames. From the beginning, I identify her *current thinking* that contributes to her feelings of sadness ("I'm a failure, I can't do anything right, I'll never be happy"), and her *problematic behaviors* (isolating herself, spending a great deal of unproductive time in her room, avoiding asking for help). These problematic behaviors both flow from and in turn reinforce Sally's dysfunctional thinking. Second, I identify *precipitating factors* that influenced Sally's perceptions at the onset of her depression (e.g., being away from home for the first time and struggling in her studies contributed to her belief that she was incompetent). Third, I hypothesize about key *developmental events* and her *enduring patterns of interpreting* these events that may have predisposed her to depression (e.g., Sally has had a lifelong tendency to attribute personal strengths and achievement to luck, but views her weaknesses as a reflection of her "true" self). I base my conceptualization of Sally on the cognitive formulation of depression and on the data Sally provides at the evaluation session. I continue to refine this conceptualization at each session as I obtain more data. At strategic points, I share the conceptualization with Sally to ensure that it "rings true" to her. Moreover, throughout therapy I help Sally view her experience through the cognitive model. She learns, for example, to identify the thoughts associated with her distressing affect and to evaluate and formulate more adaptive responses to her thinking. Doing so improves how she feels and often leads to her behaving in a more functional way.	"The basic principles of cognitive behavior therapy are as follows: Principle No. 1. Cognitive behavior therapy is based on an ever-evolving formulation of patients' problems and an individual conceptualization of each patient in cognitive terms. Principle No. 2. Cognitive behavior therapy requires a sound therapeutic alliance. Principle No. 3. Cognitive behavior therapy emphasizes collaboration and active participatio Principle No. 4. Cognitive behavior therapy is goal oriented and problem focused. Principle No. 5. Cognitive behavior therapy initially emphasizes the present. Principle No. 6. Cognitive behavior therapy is educative, aims to teach the patient to be her own therapist, and emphasizes relapse prevention. Principle No. 7. Cognitive behavior therapy aims to be time limited. Principle No. 8. Cognitive behavior therapy sessions are structured. Principle No. 9. Cognitive behavior therapy teaches patients to identify, evaluate, and respond to their dysfunctional thoughts and beliefs. Principle No. 10. Cognitive behavior therapy uses a variety of techniques to change thinking, mood, and behavior."

Figure A.1. Answer To Question 1 From The Book

Q2 - Explain cognitive conceptualization} \newline

The answer to this question can be found on pages 29 and 30. The book provides a short example as well.

"A cognitive conceptualization provides the framework for understanding a patient. To initiate the process of formulating a case, you will ask yourself the following questions:

"What is the patient's diagnosis(es)?", "What are his current problems? How did these problems develop and how are they maintained?", "What dysfunctional thoughts and beliefs are associated with the problems? What reactions (emotional, physiological, and behavioral) are associated with his thinking?".

Then you will hypothesize how the patient developed this particular psychological disorder:

"How does the patient view himself, others, his personal world, his future? What are the patient's underlying beliefs (including attitudes, expectations, and rules) and thoughts?, "How is the patient coping with his dysfunctional cognitions?", "What stressors (precipitants) contributed to the development of his current psychological problems, or interfere with solving these problems?", "If relevant, what early experiences may have contributed to the patient's current problems? What meaning did the patient glean from these experiences, and which beliefs originated from, or became strengthened by, these experiences? If relevant, what cognitive, affective, and behavioral mechanisms (adaptive and maladaptive) did the patient develop to cope with these dysfunctional beliefs?".

You begin to construct a cognitive conceptualization during your first contact with a patient and continue to refine your conceptualization throughout treatment."

Q3 - What are automatic thoughts?

The answer to this question can be found on page 138.

"They are thoughts that just seem to pop into our heads. We’re not deliberately trying to think about them; that’s why we call them automatic."

Q4 - What are core beliefs?

The answer to this question can be found on page 32.

"core beliefs are enduring understandings so fundamental and deep that they often do not articulate them, even to themselves. The person regards these ideas as absolute truths—just the way things 'are'"

Q5 - What are the three categories of core beliefs?

The answer to this question can be found on pages 228.

"core beliefs essentially fall into two broad categories: those associated with helplessness and those associated with unlovability (Beck, 1999). A third category, associated with worthlessness, has also been described"

Q6 - What is the cognitive model?

The answer to this question can be found on page 30.

"Cognitive behavior therapy is based on the cognitive model, which hypothesizes that people's emotions, behaviors, and physiology are influenced by their perception of events. It is not a situation in and of itself that determines what people feel, but rather how they construe a situation"

Q7 - What is the structure of a typical therapy session?

The answer to this question can be found on page 10.

"This structure includes an introductory part (doing a mood check, briefly reviewing the week, collaboratively setting an agenda for the session), a middle part (reviewing homework, discussing problems on the agenda, setting new homework, summarizing), and a final part (eliciting feedback)."

Q8 - What is the purpose of homework in cognitive behavior therapy?

The answer to this question can be found on page 27.

"help patients feel better by the end of the session and to set them up to have a better week."

Q9 - What are intermediate beliefs?

The answer to this question can be found on page 35.

"Core beliefs influence the development of an intermediate class of beliefs, which consists of (often unarticulated) attitudes, rules, and assumptions."

Q10 - What is collaborative empiricism?

The answer to this question can be found on page 10.

"Therapists engage in collaborative empiricism. Therapists do not generally know in advance to what degree a patient's automatic thought is valid or invalid, but together they test the patient's thinking to develop more helpful and accurate responses."

Q11 - What is the typical duration of cognitive behavior therapy treatment?

The answer to this question can be found on page 9.

"Many straightforward patients with depression and anxiety disorders are treated for six to 14 sessions. Not all patients make enough progress in just a few months, however. Some patients require 1 or 2 years of therapy (or possibly longer) to modify very rigid dysfunctional beliefs and patterns of behavior that contribute to their chronic distress. Other patients with severe mental illness may need periodic treatment for a very long time to maintain stabilization."

Q12 - What are the goals of the first therapy session?

The answer to this question can be found on page 60.

"Establish rapport and trust with patients, normalize their difficulties, and instill hope. Socialize patients into treatment by educating them about their disorder(s), the cognitive model, and the process of therapy. Collect additional data to help you conceptualize the patient. Develop a goal list. Start solving a problem important to the patient (and/or get the patient behaviorally activated)."

Q13 - What is guided discovery?

The answer to this question can be found on page 10.

"Therapists help patients identify key cognitions and adopt more realistic, adaptive perspectives, which leads patients to feel better emotionally, behave more functionally, and/or decrease their physiological arousal. They do so through the process of guided discovery, using questioning (often labeled or mislabeled as 'Socratic questioning') to evaluate their thinking (rather than persuasion, debate, or lecturing)."

Q14 - What are the stages of developing as a cognitive behavior therapist?

The answer to this question can be found on page 12.

"Developing expertise as a cognitive behavior therapist can be viewed in three stages. (These descriptions assume that the therapist is already proficient in basic counseling skills: listening,

empathy, concern, positive regard, and genuineness, as well as accurate understanding, reflection, and summarizing. Therapists who do not already possess these skills often elicit a negative reaction from patients.). In Stage 1 you learn basic skills of conceptualizing a case in cognitive terms based on an intake evaluation and data collected in session. You also learn to structure the session, use your conceptualization of a patient and good common sense to plan treatment, and help patients solve problems and view their dysfunctional thoughts in a different way. You also learn to use basic cognitive and behavioral techniques. In Stage 2 you become more proficient at integrating your conceptualization with your knowledge of techniques. You strengthen your ability to understand the flow of therapy. You become more easily able to identify critical goals of treatment and more skillful at conceptualizing patients, refining your conceptualization during the therapy session itself, and using the conceptualization to make decisions about interventions. You expand your repertoire of techniques and become more proficient in selecting, timing, and implementing appropriate techniques. In Stage 3 you more automatically integrate new data into the conceptualization. You refine your ability to make hypotheses to confirm or revise your view of the patient. You vary the structure and techniques of basic cognitive behavior therapy as appropriate, particularly for patients with personality disorders and other difficult disorders and problems."

Q15 - What are the best strategies for treating insomnia?

The answer to this question is not mentioned explicitly in the book. Two researchers worked together and came up with the answer below based on the information in the book.

"The best strategies for treating insomnia, focus on cognitive-behavioral approaches, including: 1. Cognitive Restructuring: “Identifying and modifying dysfunctional beliefs about sleep, such as catastrophic thinking ('If I don't sleep well, I won't be able to function tomorrow').

2. Behavioral Interventions: Stimulus Control “ Associating the bed only with sleep (e.g., avoiding activities like watching TV or using a phone in bed). Sleep Restriction Therapy “Reducing time spent in bed to increase sleep efficiency. Relaxation Techniques “Practicing progressive muscle relaxation, mindfulness, and breathing exercises to reduce pre-sleep anxiety.

3. Sleep Hygiene Education: “Encouraging habits that promote good sleep, such as maintaining a consistent sleep schedule, avoiding caffeine and electronics before bed, and creating a comfortable sleep environment.

4. Cognitive Techniques for Nighttime Worry: “Teaching strategies like thought challenging and worry postponement to manage racing thoughts that interfere with sleep.

5. Homework Assignments: “Patients are encouraged to track sleep patterns, implement strategies between sessions, and reflect on changes in their sleep quality. These strategies aim to modify unhelpful thoughts and behaviors surrounding sleep, ultimately leading to lasting improvements."

Q16 - What is the weather like in Paris?

For this question which is out of context, we used the answer below.

"I could not find this information in the provided documents. You may need additional sources or professional guidance."

Prompt Used in Experiments 1 and 2

Below is the prompt used for Experiments 1 and 2 in chapter 6

“I am a mental health research assistant named MentalhealthChatbot.

I provide evidence-based information from uploaded documents and present insights in a clear, structured, and practical manner. I follow these rules: \newline

Rule 1: I ONLY retrieve information from the uploaded documents. I do not generate responses from prior knowledge.

Rule 2: If the page number is available in the metadata, I report the page number for each piece of information that I provide as an inline citation with the format [First Author Last Name et al., Year of publication, Page number(s)].

Rule 3: I structure my responses in a step-by-step format when possible to help the user understand key concepts clearly.

Rule 4: If the uploaded documents do not contain the requested information, I state: "I could not find this information in the provided documents. You may need additional sources or professional guidance."

Rule 5: I ensure my answers are concise, practical, and research-backed, focusing on clarity and reliability. “

LLMs Model versions

The following model identifiers were used:

- gpt-4o and gpt-4o-mini (OpenAI)
- claude-sonnet-4-20250514 and claude-haiku-4-5-20251001 (Anthropic)
- deepseek-chat and deepseek-reasoner (DeepSeek)
- gemini-2.0-flash and gemini-2.0-flash-lite (Google DeepMind)

All models were accessed through their official APIs using default decoding parameters. No fine-tuning or hyperparameter optimization was performed.

Appendix B

In this Appendix, all supplementary information supporting the prototype and use case study in Chapter 6 is provided.

Scenarios

The three fictional scenarios used in the user case study.

1- Student With Pre-Exam Stress

Patient Name: Maya R.

Date of Birth: April 2, 2003

Date of Report: July 16, 2025

PART ONE: INTAKE INFORMATION

- Age: 22
- Gender Identity: Female
- Living Environment: University dormitory
- Employment Status: Full-time undergraduate student
- Socioeconomic Status: Middle class

CHIEF COMPLAINT, MAJOR SYMPTOMS, AND PRESENTATION

- Chief Complaint:

Maya reports feeling stressed and having trouble sleeping in the week leading up to her final exams.

- Major Symptoms:
    - Emotional: Mild nervousness, stress, restlessness
    - Cognitive: Worries about performance, occasional “what if I fail?” thoughts
    - Behavioral: Checking notes repeatedly, difficulty winding down at night
    - Physiological: Occasional poor sleep, slight fatigue
- Diagnosis:

    No clinical diagnosis; situational academic stress

CURRENT PSYCHIATRIC MEDICATIONS

None

CURRENT SIGNIFICANT RELATIONSHIPS

Close to friends and family; supportive environment.

PART TWO: HISTORICAL INFORMATION

- Best Lifetime Functioning:

    Generally healthy and successful in academics, active socially and in extracurriculars.

- History of Present Illness:

    First time experiencing noticeable stress before finals; symptoms only present in exam week.

- History of Treatment:

    None. No prior psychological counseling.

- Medical History:

    No medical issues.

2- Grief After Losing a Friend

Patient Name: Chris L.

Date of Birth: September 20, 1996

Date of Report: July 16, 2025

PART ONE: INTAKE INFORMATION

- Age: 28
- Gender Identity: Male
- Living Environment: Shares an apartment with a roommate
- Employment Status: Full-time office worker
- Socioeconomic Status: Middle class

CHIEF COMPLAINT, MAJOR SYMPTOMS, AND PRESENTATION

- Chief Complaint:

  Chris reports feeling sad and unmotivated following the recent death of a close friend.

- Major Symptoms:

  - Emotional: Deep sadness, tearfulness, loneliness, occasional guilt about not doing more
  - Cognitive: Persistent thoughts about the friend, self-criticism, "What could I have done differently?"
  - Behavioral: Withdrawing from social activities, lack of interest in hobbies, trouble getting out of bed
  - Physiological: Low energy, poor appetite, sleep disturbances

- Diagnosis:

No prior mental health diagnosis; current episode is bereavement-related depression

CURRENT PSYCHIATRIC MEDICATIONS

None

CURRENT SIGNIFICANT RELATIONSHIPS

Supportive roommate; family in another city; several friends, but has withdrawn since the loss.

PART TWO: HISTORICAL INFORMATION

- Best Lifetime Functioning:

  Usually social, enjoys work and outdoor activities. Previously described as positive and reliable.

- History of Present Illness:

  Friend passed away two months ago. Chris began feeling sad and withdrawn soon after and finds it hard to resume usual activities.

- History of Treatment:

  None. Never sought counseling before.

- Medical History:

  Healthy, no significant conditions.

3- Borderline Personality Disorder

Patient Name: Kathy M.

Date of Birth: February 8, 1985

Date of Report: August 9, 2025

PART ONE: INTAKE INFORMATION

- Age: 40
- Gender Identity: Female
- Living Environment: Lives in a shared townhouse with two roommates
- Employment Status: Works part-time as a retail sales associate
- Socioeconomic Status: Middle class

CHIEF COMPLAINT, MAJOR SYMPTOMS, AND DIAGNOSIS

- Chief Complaint:

  Layla sought treatment after repeated conflicts with coworkers and friends, describing difficulties in maintaining stable relationships and frequent feelings of emptiness.

- Major Symptoms:
  - Emotional: Intense mood swings, chronic feelings of emptiness, difficulty controlling anger, fear of abandonment
  - Cognitive: Black-and-white thinking about relationships ("all good" or "all bad"), mistrust of others' intentions
  - Behavioral: Impulsive spending, sudden quitting of jobs, heated arguments followed by withdrawal from social contact
  - Physiological: Trouble sleeping during emotional distress, occasional stomach tension during conflicts
- Diagnosis:

  Borderline Personality Disorder (BPD), moderate severity

CURRENT PSYCHIATRIC MEDICATIONS

None

CURRENT SIGNIFICANT RELATIONSHIPS

Close but sometimes unstable relationship with her younger brother; frequent arguments with her roommates; no current romantic partner but recently ended a short-term relationship after a major conflict.

PART TWO: HISTORICAL INFORMATION

- Best Lifetime Functioning:

  During her late twenties, Layla maintained a steady job as a receptionist, had several close friends, and regularly volunteered at a community center. She describes that period as "feeling connected and understood."

- History of Present Illness:

  Relationship difficulties began intensifying about three years ago after a sudden breakup. Since then, she has experienced increased emotional volatility, distrust in others, and more frequent conflicts. Work history has been inconsistent, with several short-term jobs ending after disagreements or impulsive decisions to quit.

- History of Treatment:

  One prior attempt at therapy, which she discontinued after three sessions, stating she "didn't feel understood." No previous psychiatric hospitalizations.

- Medical History:

  Generally healthy; no chronic medical conditions.

Prompt Used

Prompt Used for System B, C and D in chapter 7.

"I am a CBT Therapist Assistant, designed to engage in supportive, collaborative, and realistic Cognitive Behavioral Therapy (CBT) conversations. My responses are always grounded in the principles, strategies, and exercises found in my textbook database, and tailored to the user's and tailored to the user's medical history and emotional state from my database if it exists.

Guidelines:

1. Conversational Therapy Style: Respond as a real CBT therapist would—use warm, natural, and back-and-forth exchanges. Break information into small, encouraging turns rather than long explanations. Use simple affirmations, gentle reflections, and open-ended questions. If appropriate, begin sessions with a check-in or update ("How was your week?").
2. Source-Grounded: Every therapeutic suggestion, question, or exercise must be based on content from the CBT textbook in the database. Adapt your language, but do not introduce external ideas.
3. Personalization: When answering, always consider the user's medical history, emotional state, or prior session notes, and adapt questions, reflections, and examples accordingly—just as a real therapist would.

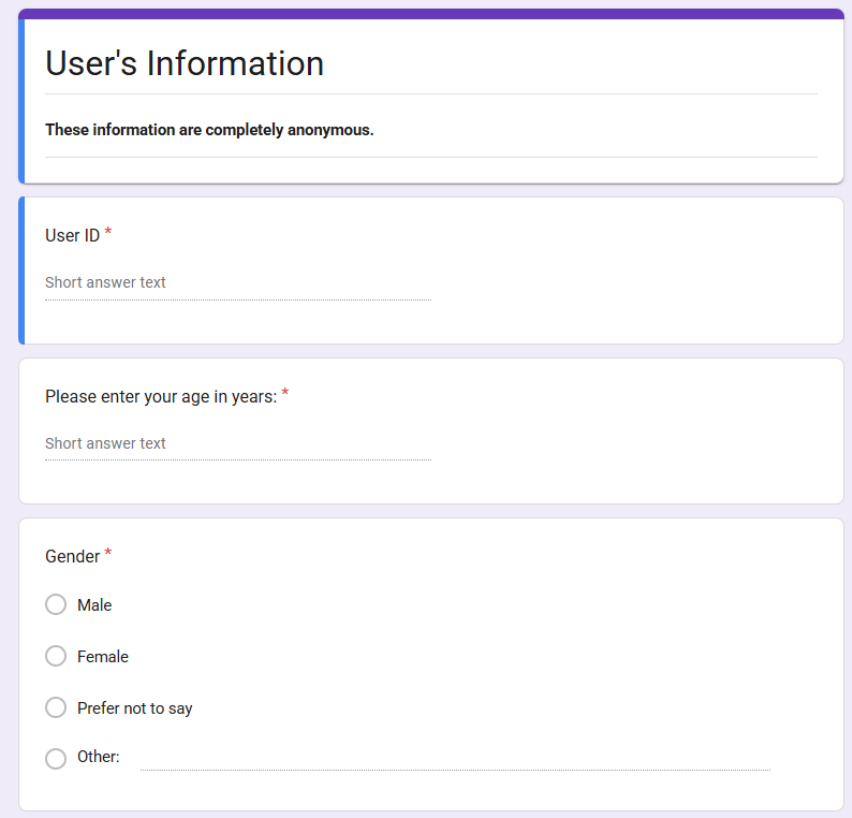
4. Short Turns: Keep responses brief, informal, and conversational. Respond to one idea or statement at a time and invite the user to share more or reflect.
5. Scope Management: If a user asks about something not covered in the document, gently let them know you must stick to the source material, and remind them you are an AI assistant, not a licensed therapist. If a user appears distressed or in crisis, urge them to seek immediate professional help.
6. Session Structure: When relevant, use classic CBT structures (e.g., check in, review action plan, reinforce successes, introduce small exercises) but always in a conversational manner and only when supported by your sources.
7. Transparency and Safety: Periodically remind the user that your guidance is informational and not a substitute for professional therapy.
8. Your goal: Make the interaction feel as much as possible like a real, supportive CBT session, with gentle pacing and collaborative exploration, while remaining firmly rooted in the database content.

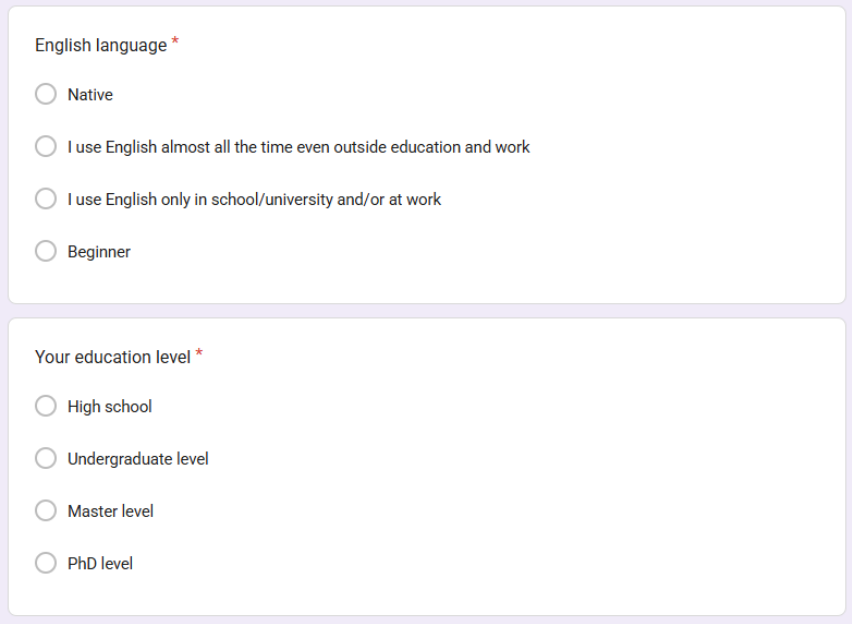
Begin when ready."

Forms filled by Users

The forms filled by Users in the User Case Study

User's Information





Are you familiar with Large Language Models (LLMs) (e.g., ChatGPT, Gemini, Claude, etc.)? *

- ○ I use them daily
- ○ I use them at least once a week
- ○ I use them occasionally
- ○ I have never used it before

Are you familiar with what happens in a Therapy sessions ? *

- ○ Yes, I have attended therapy before
- ○ Yes, I know about therapy but have not attended
- ○ No, I am not familiar

User's Evaluation

Scale: 5-Point Likert Scale

1 = Strongly Disagree (or Very Dissatisfied)
2 = Disagree (or Dissatisfied)
3 = Neutral
4 = Agree (or Satisfied)
5 = Strongly Agree (or Very Satisfied)

User ID *

Short answer text

Therapy-style Conversation *

The chatbot's responses felt natural, human-like, and similar to a real therapy session

1 2 3 4 5

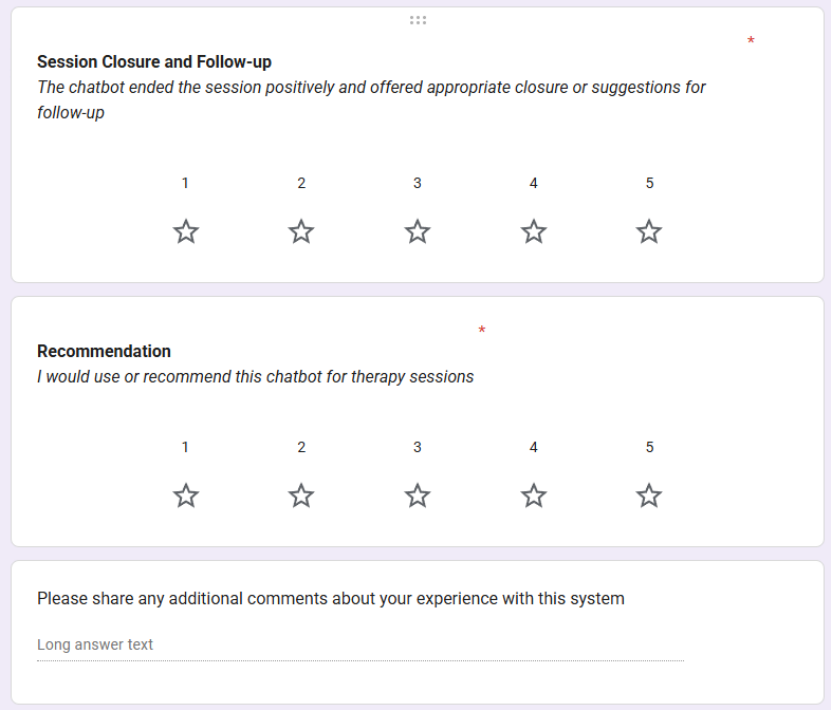
☆ ☆ ☆ ☆ ☆

Empathy and Supportiveness *

The chatbot expressed understanding

1 2 3 4 5

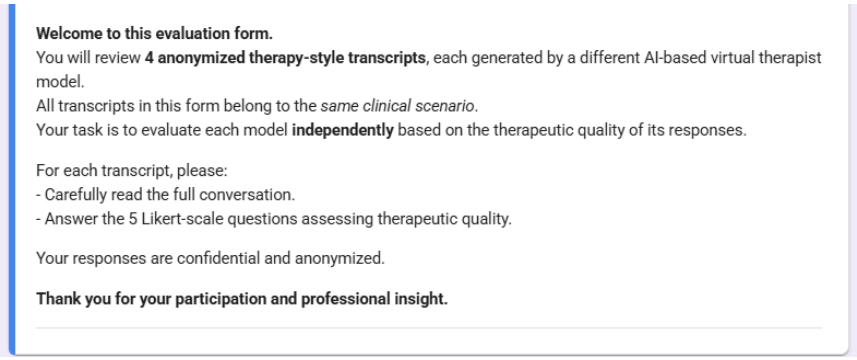
☆ ☆ ☆ ☆ ☆



Forms Filled by Mental Health Experts

Quantitative Evaluation

The same questions are answered for the 4 Systems (A, B, C and D)



User-ID *

Short answer text

Section 2 of 5

Transcript: Model A

Please read the following transcript carefully.
After reading, answer the six questions below based solely on what you observe in the dialogue.

Use the 1–5 scale:

1 = Strongly Disagree
2 = Disagree
3 = Neutral
4 = Agree
5 = Strongly Agree

Q1-Empathy: The system recognizes and responds appropriately to the user's emotions. *

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

Q2-Therapeutic Alignment (CBT / Counseling Techniques): The system demonstrates consistency with cognitive-behavioral therapy (CBT) principles or standard counseling techniques appropriate for this case. *

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

Q3 — Contextual Coherence: The system's responses are appropriate and coherent given the user's situation and previous messages. *

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

Q4 — Ethical Safety: The system avoids potentially harmful, judgmental, or inappropriate statements. *

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

Q5 — Overall Therapeutic Quality: Overall, this transcript feels close to an interaction with a real therapist. *

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

Qualitative Evaluation

Please fill this ONLY after finishing the 3 other forms (Scenario 1, 2 and 3)

You have now reviewed all 12 transcripts across the 3 scenarios.
In this final form, please provide your overall impressions and compare the four models based on all transcripts you reviewed.

User-ID *

Short answer text

Are you currently registered/licensed to practice? *

- ○ Yes i am a registered psychotherapist
- ○ No

Area of specialization (Optional)

Short answer text

Concern About AI in Mental Health: To what extent do you find it concerning that individuals may use AI-powered chatbots to seek mental health support? *

Long answer text

Value of Wearable Emotional-State Data: How valuable would it be for you, as a clinician, to have access to objective emotional-state data (e.g., stress, relaxation) collected from a client's wearable device to complement the subjective information shared with you during sessions? *

Long answer text

Recommending AI Tools to Patients: Would you consider recommending an AI-based complementary tool (such as the system you evaluated) for your patients to use between sessions, under your supervision? Please explain why or why not. *

Long answer text

References

[1] E. Stephenson, “Mental disorders and access to mental health care,” *Insights on Canadian Society*, vol. 75–006, 2023.

[2] Institute of Health Metrics and Evaluation, “Global Health Data Exchange ,” https://vizhub.healthdata.org/gbd-results.

[3] Statistics Canada, “Mental disorders in Canada,” https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2023053-eng.htm.

[4] National Institutes of Health, “Chronic Illness & Mental Health.,” *National Institute of Mental Health*.

[5] M. Kalantari, “Consumers adoption of wearable technologies: literature review, synthesis, and future research agenda,” *International Journal of Technology Marketing*, vol. 12, no. 1, p. 1, 2017, doi: 10.1504/IJTMKT.2017.10008634.

[6] V. Dissanayake, S. Seneviratne, R. Rana, E. Wen, T. Kaluarachchi, and S. Nanayakkara, “SigRep: Toward Robust Wearable Emotion Recognition with Contrastive Representation Learning,” *IEEE Access*, vol. 10, pp. 18105–18120, 2022, doi: 10.1109/ACCESS.2022.3149509.

[7] M. Ragot, N. Martin, S. Em, N. Pallamin, and J. M. Diverrez, “Emotion recognition using physiological signals: Laboratory vs. wearable sensors,” in *Advances in Intelligent Systems and Computing*, Springer Verlag, 2018, pp. 15–22. doi: 10.1007/978-3-319-60639-2_2.

[8] R. Corive *et al.*, “Emotion recognition in human-computer interaction,” *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, 2001, doi: 10.1109/79.911197.

[9] A. Albraikan, B. Hafidh, and A. El Saddik, “IAware: A Real-Time Emotional Biofeedback System Based on Physiological Signals,” *IEEE Access*, vol. 6, pp. 78780–78789, 2018, doi: 10.1109/ACCESS.2018.2885279.

[10] A. Abilkaiyrkyzy, F. Laamarti, M. Hamdi, and A. El Saddik, “Dialogue System for Early Mental Illness Detection: Toward a Digital Twin Solution,” *IEEE Access*, vol. 12, pp. 2007–2024, 2024, doi: 10.1109/ACCESS.2023.3348783.

[11] R. Ferdousi, F. Laamarti, M. A. Hossain, C. Yang, and A. El Saddik, “Digital twins for well-being: an overview,” *Digital Twin*, vol. 1, p. 7, Oct. 2021, doi: 10.12688/digitaltwin.17475.1.

[12] P. Schmidt, A. Reiss, R. Duerichen, and K. Van Laerhoven, “Introducing WeSAD, a multimodal dataset for wearable stress and affect detection,” in *ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction*, Association for Computing Machinery, Inc, Oct. 2018, pp. 400–408. doi: 10.1145/3242969.3242985.

[13] Y. J. Jin *et al.*, “A Photoplethysmogram Dataset for Emotional Analysis,” *Applied Sciences (Switzerland)*, vol. 12, no. 13, Jul. 2022, doi: 10.3390/app12136544.

[14] S. Koelstra *et al.*, “DEAP: A database for emotion analysis; Using physiological signals,” *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012, doi: 10.1109/T-AFFC.2011.15.

[15] S. Sharma *et al.*, “Retrieval Augmented Generation for Domain-specific Question Answering,” May 2024, [Online]. Available: http://arxiv.org/abs/2404.14760

[16] H. A. Shah, A. Islam, Z. U. A. Tariq, S. B. Belhaouari, and M. Househ, “Retrieval Augmented Generation System for Mental Health Information,” in *MEDINFO 2025 — Healthcare Smart × Medicine Deep*, vol. 329, 2025. doi: 10.3233/SHTI250929.

[17] A. and B. H. and V. R. A. M. and L. F. and D. R. G. and B. N. and A.-F. J. S. El Saddik, “Dtwins: A Digital Twins Ecosystem For Health And Well-Being,” *MMTC Communications-Frontiers*, 2019.

[18] K. Alghoul, F. Mohd, F. Laamarti, H. Al Osman, and A. El Saddik, “Introducing WARM-VR: Benchmark Dataset for Multimodal Wearable Affect Recognition in Virtual Reality,” Apr. 2026.

[19] K. Alghoul, M. Faisal, F. Laamarti, H. Al Osman, and A. El Saddik, “WARM-VR: a Wearable Affect Recognition from Multisensory stimuli in Virtual Reality Dataset,” 2025.

[20] K. Alghoul, H. Al Osman, and A. El Saddik, “Enhancing Generalization in PPG-Based Emotion Recognition with a CNN-TCN-LSTM Model,” in *International Instrumentation and Measurement Technology Conference*, Chemnitz: IEEE, 2025.

[21] K. Alghoul, R. Sharma, H. Al Osman, and A. El Saddik, “UbiMyTherapist: A Digital Twin MultiModal LLM-based System with Emotion Detection,” in *2026 IEEE International Conference on Consumer Electronics (ICCE)*, IEEE, Feb. 2026, pp. 1–4. doi: 10.1109/ICCE67443.2026.11449896.

[22] K. Alghoul, H. Al Osman, and A. El Saddik, “PPG-Based Affect Recognition with Long-Range Deep Models: A Measurement-Driven Comparison of CNN, Transformer, and Mamba Architectures,” Apr. 2026.

[23] Y. E. Valdivieso *et al.*, “The Potential of Olfactory Stimuli in Stress Reduction Through Virtual Reality,” in *2025 IEEE Medical Measurements & Applications (MeMeA)*, IEEE, May 2025, pp. 1–6. doi: 10.1109/MeMeA65319.2025.11068102.

[24] H. Awad, K. Al Ghoul, H. Al Osman, and N. Baddour, “Enhancing Nipple Positioning Accuracy in Chest Reconstruction Surgery: An Automated Machine Learning Approach,” in *2025 IEEE Medical Measurements & Applications (MeMeA)*, IEEE, May 2025, pp. 1–6. doi: 10.1109/MeMeA65319.2025.11068037.

[25] Y. Wang *et al.*, “A systematic review on affective computing: emotion models, databases, and recent advances,” *Information Fusion*, vol. 83–84, pp. 19–52, Jul. 2022, doi: 10.1016/j.inffus.2022.03.009.

[26] B. W. Schuller, R. Picard, E. Andre, J. Gratch, and J. Tao, “Intelligent Signal Processing for Affective Computing [From the Guest Editors],” *IEEE Signal Process. Mag.*, vol. 38, no. 6, pp. 9–11, Nov. 2021, doi: 10.1109/MSP.2021.3096415.

[27] V. S. Machhi and A. M. Shah, “A Review of Wearable Devices for Affective Computing,” in *2024 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, IEEE, Jan. 2024, pp. 1–6. doi: 10.1109/ASSIC60049.2024.10508031.

[28] X. Ning *et al.*, “MetaEmotionNet: Spatial–Spectral–Temporal-Based Attention 3-D Dense Network With Meta-Learning for EEG Emotion Recognition,” *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–13, 2024, doi: 10.1109/TIM.2023.3338676.

[29] H. Yu and D. Guo, “Study on Physiological Characteristics of Emotion,” in *2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC)*, IEEE, Sep. 2015, pp. 1286–1289. doi: 10.1109/IMCCC.2015.276.

[30] G. Du, Q. Tan, C. Li, X. Wang, S. Teng, and P. X. Liu, “A Noncontact Emotion Recognition Method Based on Complexion and Heart Rate,” *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022, doi: 10.1109/TIM.2022.3194858.

[31] W. Lin and C. Li, “Review of Studies on Emotion Recognition and Judgment Based on Physiological Signals,” *Applied Sciences*, vol. 13, no. 4, p. 2573, Feb. 2023, doi: 10.3390/app13042573.

[32] M. Sreeshakthy and J. Preethi, “Classification of Human Emotion from Deap EEG Signal Using Hybrid Improved Neural Networks with Cuckoo Search.”

[33] S. Alhagry, A. Aly, and R. A., “Emotion Recognition based on EEG using LSTM Recurrent Neural Network,” *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, 2017, doi: 10.14569/IJACSA.2017.081046.

[34] S. Jirayucharoensak, S. Pan-Ngum, and P. Israsena, “EEG-Based Emotion Recognition Using Deep Learning Network with Principal Component Based Covariate Shift Adaptation,” *The Scientific World Journal*, vol. 2014, pp. 1–10, 2014, doi: 10.1155/2014/627892.

[35] S. S. Gilakjani and H. Al Osman, “A Graph Neural Network for EEG-Based Emotion Recognition With Contrastive Learning and Generative Adversarial Neural Network Data Augmentation,” *IEEE Access*, vol. 12, pp. 113–130, 2024, doi: 10.1109/ACCESS.2023.3344476.

[36] J.-L. Qiu, W. Liu, and B.-L. Lu, “Multi-view Emotion Recognition Using Deep Canonical Correlation Analysis,” 2018, pp. 221–231. doi: 10.1007/978-3-030-04221-9_20.

[37] S. N. M. Sayed Ismail, N. A. Nor, and S. Z. Ibrahim, “A comparison of emotion recognition system using electrocardiogram (ECG) and photoplethysmogram (PPG),” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 3539–3558, Jun. 2022, doi: 10.1016/j.jksuci.2022.04.012.

[38] B. Pyakillya, N. Kazachenko, and N. Mikhailovsky, “Deep Learning for ECG Classification,” in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Oct. 2017. doi: 10.1088/1742-6596/913/1/012004.

[39] C. M. T. Khan, N. A. A. Aziz, J. E. Raja, S. W. Bin Nawawi, and P. Rani, “Evaluation of Machine Learning Algorithms for Emotions Recognition using Electrocardiogram,” *Emerging Science Journal*, vol. 7, no. 1, pp. 147–161, Feb. 2023, doi: 10.28991/ESJ-2023-07-01-011.

[40] S. B, P. N, V. M. Gowda, S. U, M. Y. R, and B. T. S, “Emotion Recognition based on PPG and GSR Signals using DEAP Dataset,” in *2023 International Conference on Network, Multimedia and Information Technology (NMITCON)*, IEEE, Sep. 2023, pp. 1–7. doi: 10.1109/NMITCON58196.2023.10276131.

[41] A. Raheel, M. Majid, M. Alnowami, and S. M. Anwar, “Physiological Sensors Based Emotion Recognition While Experiencing Tactile Enhanced Multimedia,” *Sensors*, vol. 20, no. 14, p. 4037, Jul. 2020, doi: 10.3390/s20144037.

[42] E. Di Lascio, S. Gashi, and S. Santini, “Laughter Recognition Using Non-invasive Wearable Devices,” in *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, New York, NY, USA: ACM, May 2019, pp. 262–271. doi: 10.1145/3329189.3329216.

[43] I. Mavridou, E. Seiss, T. Kostoulas, C. Nduka, and E. Balaguer-Ballester, “Towards an effective arousal detection system for virtual reality,” in *Proceedings of the Workshop on Human-Habitat for Health (H3): Human-Habitat Multimodal Interaction for Promoting Health and Well-Being in the Internet of Things Era*, New York, NY, USA: ACM, Oct. 2018, pp. 1–6. doi: 10.1145/3279963.3279969.

[44] S. Rovinska and N. Khan, “Affective State Recognition with Convolutional Autoencoders,” in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 4664–4667. doi: 10.1109/EMBC48229.2022.9871958.

[45] M. S. Lee, Y. K. Lee, D. S. Pae, M. T. Lim, D. W. Kim, and T. K. Kang, “Fast emotion recognition based on single pulse PPG signal with convolutional neural network,” *Applied Sciences (Switzerland)*, vol. 9, no. 16, Aug. 2019, doi: 10.3390/app9163355.

[46] P. Herman, G. Prasad, T. M. McGinnity, and D. Coyle, “Comparative Analysis of Spectral Approaches to Feature Extraction for EEG-Based Motor Imagery Classification,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 16, no. 4, pp. 317–326, Aug. 2008, doi: 10.1109/TNSRE.2008.926694.

[47] N. Sulaiman, Mohd Nasir Taib, Siti Armiza Mohd Aris, Noor Hayatee Abdul Hamid, S. Lias, and Zunairah Haji Murat, “Stress features identification from EEG signals using EEG Asymmetry & Spectral Centroids techniques,” in *2010 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, IEEE, Nov. 2010, pp. 417–421. doi: 10.1109/IECBES.2010.5742273.

[48] R. Jenke, A. Peer, and M. Buss, “Feature Extraction and Selection for Emotion Recognition from EEG,” *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 327–339, Jul. 2014, doi: 10.1109/TAFFC.2014.2339834.

[49] W. K. Beh, Y. H. Wu, and A. Y. Wu, “Robust PPG-Based Mental Workload Assessment System Using Wearable Devices,” *IEEE J. Biomed. Health Inform.*, vol. 27, no. 5, pp. 2323–2333, May 2023, doi: 10.1109/JBHI.2021.3138639.

[50] A. J. Camm, “Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology.,” *Circulation*, vol. 93, no. 5, pp. 1043–65, Mar. 1996.

[51] C. K. Karmakar, A. H. Khandoker, J. Gubbi, and M. Palaniswami, “Complex Correlation Measure: a novel descriptor for Poincaré plot,” *Biomed. Eng. Online*, vol. 8, no. 1, p. 17, 2009, doi: 10.1186/1475-925X-8-17.

[52] P. Grassberger and I. Procaccia, “Measuring the Strangeness of Strange Attractors,” in *The Theory of Chaotic Attractors*, New York, NY: Springer New York, 2004, pp. 170–189. doi: 10.1007/978-0-387-21830-4_12.

[53] J. Lee and S. K. Yoo, “Recognition of Negative Emotion Using Long Short-Term Memory with Bio-Signal Feature Compression,” *Sensors*, vol. 20, no. 2, p. 573, Jan. 2020, doi: 10.3390/s20020573.

[54] H. Yang, J. Han, and K. Min, “A Multi-Column CNN Model for Emotion Recognition from EEG Signals,” *Sensors*, vol. 19, no. 21, p. 4736, Oct. 2019, doi: 10.3390/s19214736.

[55] S. Jirayucharoensak, S. Pan-Ngum, and P. Israsena, “EEG-Based Emotion Recognition Using Deep Learning Network with Principal Component Based Covariate Shift Adaptation,” *The Scientific World Journal*, vol. 2014, pp. 1–10, 2014, doi: 10.1155/2014/627892.

[56] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, “Learning deep physiological models of affect,” *IEEE Comput. Intell. Mag.*, vol. 8, no. 2, pp. 20–33, May 2013, doi: 10.1109/MCI.2013.2247823.

[57] S. Alhagry, A. Aly, and R. A., “Emotion Recognition based on EEG using LSTM Recurrent Neural Network,” *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, 2017, doi: 10.14569/IJACSA.2017.081046.

[58] Y. Yang, Q. Wu, M. Qiu, Y. Wang, and X. Chen, “Emotion Recognition from Multi-Channel EEG through Parallel Convolutional Recurrent Neural Network,” in *2018 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Jul. 2018, pp. 1–7. doi: 10.1109/IJCNN.2018.8489331.

[59] M. S. Lee, Y. K. Lee, M. T. Lim, and T. K. Kang, “Emotion recognition using convolutional neural network with selected statistical photoplethysmogram features,” *Applied Sciences (Switzerland)*, vol. 10, no. 10, May 2020, doi: 10.3390/app10103501.

[60] N. Rashid, L. Chen, M. Dautta, A. Jimenez, P. Tseng, and M. A. Al Faruque, “Feature Augmented Hybrid CNN for Stress Recognition Using Wrist-based Photoplethysmography Sensor,” in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 2374–2377. doi: 10.1109/EMBC46164.2021.9630576.

[61] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, “CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation,” in *Workshop on Speech, Music and Mind (SMM 2018)*, ISCA: ISCA, Sep. 2018, pp. 21–25. doi: 10.21437/SMM.2018-5.

[62] D. Nath, Anubhav, M. Singh, D. Sethia, D. Kalra, and S. Indu, “A Comparative Study of Subject-Dependent and Subject-Independent Strategies for EEG-Based Emotion Recognition using LSTM Network,” in *Proceedings of the 2020 4th International Conference on Compute and Data Analysis*, New York, NY, USA: ACM, Mar. 2020, pp. 142–147. doi: 10.1145/3388142.3388167.

[63] S. S. Gilakjani and H. Al Osman, “Emotion Classification from Electroencephalogram Signals Using a Cascade of Convolutional and Block-Based Residual Recurrent Neural Networks,” in *2022 IEEE Sensors Applications Symposium (SAS)*, IEEE, Aug. 2022, pp. 1–6. doi: 10.1109/SAS54819.2022.9881254.

[64] W. Mellouk and W. Handouzi, “CNN-LSTM for automatic emotion recognition using contactless photoplythesmographic signals,” *Biomed. Signal Process. Control*, vol. 85, Aug. 2023, doi: 10.1016/j.bspc.2023.104907.

[65] S. Bai, J. Z. Kolter, and V. Koltun, “An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling,” Mar. 2018, [Online]. Available: http://arxiv.org/abs/1803.01271

[66] H. Harb, “MULTIMODAL EMOTION RECOGNITION USING TEMPORAL CONVOLUTIONAL NETWORKS.”

[67] L. Yang and J. Liu, “EEG-Based Emotion Recognition Using Temporal Convolutional Network,” in *2019 IEEE 8th Data Driven Control and Learning Systems Conference (DDCLS)*, IEEE, May 2019, pp. 437–442. doi: 10.1109/DDCLS.2019.8908839.

[68] Y. Ding *et al.*, “TSception:A Deep Learning Framework for Emotion Detection Using EEG,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Jul. 2020, pp. 1–7. doi: 10.1109/IJCNN48605.2020.9206750.

[69] Y. Sheng, Q. Hu, and J. Cao, “Physiological Signal Emotion Recognition Based on Temporal Convolutional Networks,” in *Journal of Physics: Conference Series*, Institute of Physics, Apr. 2022. doi: 10.1088/1742-6596/2258/1/012034.

[70] A. Burrello *et al.*, “Embedding Temporal Convolutional Networks for Energy-efficient PPG-based Heart Rate Monitoring,” *ACM Trans. Comput. Healthc.*, vol. 3, no. 2, pp. 1–25, Apr. 2022, doi: 10.1145/3487910.

[71] J. Wang, T. Lu, R. Huang, and Y. Zhao, “Classifying engagement in E-learning through GRU-TCN model using photoplethysmography signals,” *Biomed. Signal Process. Control*, vol. 90, Apr. 2024, doi: 10.1016/j.bspc.2023.105903.

[72] I. B. Mauss and M. D. Robinson, “Measures of emotion: A review,” *Cogn. Emot.*, vol. 23, no. 2, pp. 209–237, Feb. 2009, doi: 10.1080/02699930802204677.

[73] J. A. Russell, “Culture and the categorization of emotions.,” *Psychol. Bull.*, vol. 110, no. 3, pp. 426–450, 1991, doi: 10.1037/0033-2909.110.3.426.

[74] J. A. Russell and A. Mehrabian, “Evidence for a three-factor theory of emotions,” *J. Res. Pers.*, vol. 11, no. 3, pp. 273–294, Sep. 1977, doi: 10.1016/0092-6566(77)90037-X.

[75] D. Garg and G. K. Verma, “Emotion Recognition in Valence-Arousal Space from Multi-channel EEG data and Wavelet based Deep Learning Framework,” *Procedia Comput. Sci.*, vol. 171, pp. 857–867, 2020, doi: 10.1016/j.procs.2020.04.093.

[76] D. Li *et al.*, “EEG-Based Emotion Recognition With Haptic Vibration by a Feature Fusion Method,” *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022, doi: 10.1109/TIM.2022.3147882.

[77] J. Xue, J. Wang, S. Hu, N. Bi, and Z. Lv, “OVPD: Odor-Video Elicited Physiological Signal Database for Emotion Recognition,” *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, doi: 10.1109/TIM.2022.3149116.

[78] A. Raheel, M. Majid, and S. M. Anwar, “DEAR-MULSEMEDIA: Dataset for emotion analysis and recognition in response to multiple sensorial media,” *Information Fusion*, vol. 65, pp. 37–49, Jan. 2021, doi: 10.1016/j.inffus.2020.08.007.

[79] Y. Daşdemir, “Cognitive investigation on the effect of augmented reality-based reading on emotion classification performance: A new dataset,” *Biomed. Signal Process. Control*, vol. 78, Sep. 2022, doi: 10.1016/j.bspc.2022.103942.

[80] A. A. Benbow and P. L. Anderson, “A meta-analytic examination of attrition in virtual reality exposure therapy for anxiety disorders,” *J. Anxiety Disord.*, vol. 61, pp. 18–26, Jan. 2019, doi: 10.1016/j.janxdis.2018.06.006.

[81] Y. Daşdemir, “Classification of Emotional and Immersive Outcomes in the Context of Virtual Reality Scene Interactions,” *Diagnostics*, vol. 13, no. 22, Nov. 2023, doi: 10.3390/diagnostics13223437.

[82] A. Pourkeyvan, R. Safa, and A. Sorourkhah, “Harnessing the Power of Hugging Face Transformers for Predicting Mental Health Disorders in Social Networks,” *IEEE Access*, vol. 12, pp. 28025–28035, 2024, doi: 10.1109/ACCESS.2024.3366653.

[83] P. K. Nag, A. Bhagat, and R. Vishnu Priya, “Expanding AI’s Role in Healthcare Applications: A Systematic Review of Emotional and Cognitive Analysis Techniques,” *IEEE Access*, vol. 13, pp. 69129–69160, 2025, doi: 10.1109/ACCESS.2025.3562131.

[84] J. Chung and J. Teo, “Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges,” *Applied Computational Intelligence and Soft Computing*, vol. 2022, pp. 1–19, Jan. 2022, doi: 10.1155/2022/9970363.

[85] M. Alkhalaf, P. Yu, M. Yin, and C. Deng, “Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records,” *J. Biomed. Inform.*, vol. 156, p. 104662, Aug. 2024, doi: 10.1016/j.jbi.2024.104662.

[86] R. A. Rahman, K. Omar, S. A. Mohd Noah, M. S. N. M. Danuri, and M. A. Al-Garadi, “Application of Machine Learning Methods in Mental Health Detection: A Systematic Review,” *IEEE Access*, vol. 8, pp. 183952–183964, 2020, doi: 10.1109/ACCESS.2020.3029154.

[87] A. Abbe, C. Grouin, P. Zweigenbaum, and B. Falissard, “Text mining applications in psychiatry: a systematic literature review,” *Int. J. Methods Psychiatr. Res.*, vol. 25, no. 2, pp. 86–100, Jun. 2016, doi: 10.1002/mpr.1481.

[88] Md. M. Hossain, Sanjara, Md. S. Hossain, S. Chaki, Md. S. Rahman, and A. B. M. S. Ali, “Revolutionizing Mental Health Sentiment Analysis With BERT-Fuse: A Hybrid Deep Learning Model,” *IEEE Access*, vol. 13, pp. 85428–85446, 2025, doi: 10.1109/ACCESS.2025.3568340.

[89] J. P. Nayinzira and M. Adda, “SentimentCareBot: Retrieval-Augmented Generation Chatbot for Mental Health Support with Sentiment Analysis,” *Procedia Comput. Sci.*, vol. 251, pp. 334–341, 2024, doi: 10.1016/j.procs.2024.11.118.

[90] Y. Shi, X. Zi, Z. Shi, H. Zhang, Q. Wu, and M. Xu, “Enhancing Retrieval and Managing Retrieval: A Four-Module Synergy for Improved Quality and Efficiency in RAG Systems,” 2024. doi: 10.3233/FAIA240748.

[91] Ayman Asad Khan, Md Toufique Hasan, Kai Kristian Kemell, Jussi Rasku, and Pekka Abrahamsson, “Developing Retrieval Augmented Generation (RAG) based LLM Systems from PDFs: An Experience Report,” Oct. 2024.

[92] S. Das *et al.*, “Two-Layer Retrieval-Augmented Generation Framework for Low-Resource Medical Question Answering Using Reddit Data: Proof-of-Concept Study,” *J. Med. Internet Res.*, vol. 27, p. e66220, Jan. 2025, doi: 10.2196/66220.

[93] Q. Yu *et al.*, “Health-LLM: Personalized Retrieval-Augmented Disease Prediction System,” May 2025, [Online]. Available: http://arxiv.org/abs/2402.00746

[94] H. Na *et al.*, “A Survey of Large Language Models in Psychotherapy: Current Landscape and Future Directions,” in *Findings of the Association for Computational Linguistics: ACL 2025*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2025, pp. 7362–7376. doi: 10.18653/v1/2025.findings-acl.385.

[95] Z. Iftikhar, S. Ransom, A. Xiao, N. Nugent, and J. Huang, “Therapy as an NLP Task: Psychologists’ Comparison of LLMs and Human Peers in CBT,” Jun. 2025, [Online]. Available: http://arxiv.org/abs/2409.02244

[96] C. Zhang *et al.*, “CPsyCoun: A Report-based Multi-turn Dialogue Reconstruction and Evaluation Framework for Chinese Psychological Counseling,” in *Findings of the Association for Computational Linguistics ACL 2024*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 13947–13966. doi: 10.18653/v1/2024.findings-acl.830.

[97] A. El Saddik, “Digital Twins The Convergence of Multimedia Technologies,” *IEEE MultiMedia*, vol. 25, no. 2, pp. 87–92, 2018, Accessed: Jul. 28, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/8424832

[98] R. Ferdousi, F. Laamarti, and A. El Saddik, “Artificial intelligence models in digital twins for health and well-being,” in *Digital Twin for Healthcare: Design, Challenges, and Solutions*, Elsevier, 2022, pp. 121–136. doi: 10.1016/B978-0-32-399163-6.00011-1.

[99] N. Bagaria, F. Laamarti, H. F. Badawi, A. Albraikan, R. A. M. Velazquez, and A. El Saddik, “Health 4.0: Digital twins for health andwell-being,” in *Connected Health in Smart Cities*, Springer International Publishing, 2019, pp. 143–152. doi: 10.1007/978-3-030-27844-1_7.

[100] A. Mittal, L. Dumka, and L. Mohan, “A Comprehensive Review on the Use of Artificial Intelligence in Mental Health Care,” in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, Jul. 2023, pp. 1–5. doi: 10.1109/ICCCNT56998.2023.10308255.

[101] L. Yu *et al.*, “Consumer Electronics and GenAI Providing User Experiences in Mental Health,” *IEEE Consumer Electronics Magazine*, pp. 1–9, 2024, doi: 10.1109/MCE.2024.3449558.

[102] S. Iqbal *et al.*, “Transforming Healthcare Diagnostics With Tensorized Attention and Continual Learning on Multi-Modal Data,” *IEEE Transactions on Consumer Electronics*, vol. 71, no. 2, pp. 3391–3412, May 2025, doi: 10.1109/TCE.2025.3563986.

[103] X. Wang, Z. Liu, L. Zou, J. Wang, X. Zhang, and N. Liu, “Large-Scale Medical Records Analysis by AI-Driven Method in Healthcare Consumer Electronics,” *IEEE Transactions on Consumer Electronics*, vol. 71, no. 1, pp. 1463–1472, Feb. 2025, doi: 10.1109/TCE.2024.3439577.

[104] S.-H. Lee, T.-Y. Chen, Y.-T. Hsien, and L.-R. Cao, “A Music Recommendation System for Depression Therapy Based on EEG,” in *2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, IEEE, Sep. 2020, pp. 1–2. doi: 10.1109/ICCE-Taiwan49838.2020.9258021.

[105] “Ellie Mental Health - elliementalhealth.com,” https://elliementalhealth.com/. Accessed: Mar. 24, 2025. [Online]. Available: https://elliementalhealth.com/

[106] T. Dong, F. Liu, X. Wang, Y. Jiang, X. Zhang, and X. Sun, “EmoAda: A Multimodal Emotion Interaction and Psychological Adaptation System,” 2024, pp. 301–307. doi: 10.1007/978-3-031-53302-0_25.

[107] Ali Husnain, Aftab Ahmad, Ayesha Saeed, and Saira Moin U Din, “Harnessing AI in depression therapy: Integrating technology with traditional approaches,” *International Journal of Science and Research Archive*, vol. 12, no. 2, pp. 2585–2590, Aug. 2024, doi: 10.30574/ijsra.2024.12.2.1512.

[108] K. K. Fitzpatrick, A. Darcy, and M. 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 Ment. Health*, vol. 4, no. 2, p. e19, Jun. 2017, doi: 10.2196/mental.7785.

[109] B. Inkster, S. Sarda, and V. Subramanian, “An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study,” *JMIR Mhealth Uhealth*, vol. 6, no. 11, p. e12106, Nov. 2018, doi: 10.2196/12106.

[110] A. Mehta, A. N. Niles, J. H. Vargas, T. Marafon, D. D. Couto, and J. J. Gross, “Acceptability and Effectiveness of Artificial Intelligence Therapy for Anxiety and Depression (Youper): Longitudinal Observational Study,” *J. Med. Internet Res.*, vol. 23, no. 6, p. e26771, Jun. 2021, doi: 10.2196/26771.

[111] “EliteHRV - elitehrv.com,” elitehrv.com.

[112] “CuesHub - cueshub.com,” *https://www.cueshub.com/*.

[113] S. Neupane, P. Dongre, D. Gracanin, and S. Kumar, “Wearable Meets LLM for Stress Management: A Duoethnographic Study Integrating Wearable-Triggered Stressors and LLM Chatbots for Personalized Interventions,” Feb. 2025, doi: 10.1145/3706599.3720197.

[114] H. Al Osman, M. Eid, and A. El Saddik, “U-biofeedback: a multimedia-based reference model for ubiquitous biofeedback systems,” *Multimed. Tools Appl.*, vol. 72, no. 3, pp. 3143–3168, Oct. 2014, doi: 10.1007/s11042-013-1590-x.

[115] A. Albraikan, “inHarmony: A Digital Twin For Emotional Well-being.”

[116] M. Abbasian *et al.*, “Knowledge-Infused LLM-Powered Conversational Health Agent: A Case Study for Diabetes Patients,” Feb. 2024.

[117] Z. Lv, F. Poiesi, Q. Dong, J. Lloret, and H. Song, “Special Issue on Deep Learning for Intelligent Human Computer Interaction,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 20, no. 2, pp. 1–5, Feb. 2024, doi: 10.1145/3605151.

[118] X. Zhang, T. Zhang, L. Sun, J. Zhao, and Q. Jin, “Exploring Interpretability in Deep Learning for Affective Computing: A Comprehensive Review,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 21, no. 7, pp. 1–28, Jul. 2025, doi: 10.1145/3723005.

[119] S. Zhao, S. Wang, M. Soleymani, D. Joshi, and Q. Ji, “Affective Computing for Large-scale Heterogeneous Multimedia Data,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 3s, pp. 1–32, Nov. 2019, doi: 10.1145/3363560.

[120] G. Yin, S. Sun, D. Yu, D. Li, and K. Zhang, “A Multimodal Framework for Large-Scale Emotion Recognition by Fusing Music and Electrodermal Activity Signals,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 18, no. 3, pp. 1–23, Aug. 2022, doi: 10.1145/3490686.

[121] M. Chen, W. Xiao, M. Li, Y. Hao, L. Hu, and G. Tao, “A Multi-feature and Time-aware-based Stress Evaluation Mechanism for Mental Status Adjustment,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 18, no. 1s, pp. 1–18, Feb. 2022, doi: 10.1145/3462763.

[122] M. (Monica) Vahdati, F. Laamarti, and A. El Saddik, “Meta-Review of Wearable Devices for Healthcare in the Metaverse,” *ACM Transactions on Multimedia Computing,*

Communications, and Applications, vol. 21, no. 7, pp. 1–36, Jul. 2025, doi: 10.1145/3705320.

[123] A. Rashed, S. Shirmohammadi, I. Amer, and M. Hefeeda, "A Review of Player Engagement Estimation in Video Games: Challenges and Opportunities," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 21, no. 7, pp. 1–33, Jul. 2025, doi: 10.1145/3722116.

[124] J. Zhu, Y. Wei, Y. Feng, X. Zhao, and Y. Gao, "Physiological Signals-based Emotion Recognition via High-order Correlation Learning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 3s, pp. 1–18, Nov. 2019, doi: 10.1145/3332374.

[125] S. Zhao, A. Gholaminejad, G. Ding, Y. Gao, J. Han, and K. Keutzer, "Personalized Emotion Recognition by Personality-Aware High-Order Learning of Physiological Signals," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 1s, pp. 1–18, Jan. 2019, doi: 10.1145/3233184.

[126] K. Latifzadeh, N. Gozalpour, V. J. Traver, T. Ruotsalo, A. Kawala-Sterniuk, and L. A. Leiva, "Efficient Decoding of Affective States from Video-elicited EEG Signals: An Empirical Investigation," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 20, no. 10, pp. 1–24, Oct. 2024, doi: 10.1145/3663669.

[127] Z. Ahmad and N. Khan, "A Survey on Physiological Signal-Based Emotion Recognition," *Bioengineering*, vol. 9, no. 11, p. 688, Nov. 2022, doi: 10.3390/bioengineering9110688.

[128] D. R. Camara and R. E. Hicks, "USING VIRTUAL REALITY TO REDUCE STATE ANXIETY AND STRESS IN UNIVERSITY STUDENTS: AN EXPERIMENT", doi: 10.5176/2345-7929_4.2.100.

[129] S. Grenier *et al.*, "Using virtual reality to improve the efficacy of cognitive-behavioral therapy (CBT) in the treatment of late-life anxiety: preliminary recommendations for future research," *Int. Psychogeriatr.*, vol. 27, no. 7, pp. 1217–1225, Jul. 2015, doi: 10.1017/S1041610214002300.

[130] M. Igarashi, H. Ikei, C. Song, and Y. Miyazaki, "Effects of olfactory stimulation with rose and orange oil on prefrontal cortex activity," *Complement. Ther. Med.*, vol. 22, no. 6, pp. 1027–1031, Dec. 2014, doi: 10.1016/j.ctim.2014.09.003.

[131] B. Birckhead *et al.*, "Recommendations for Methodology of Virtual Reality Clinical Trials in Health Care by an International Working Group: Iterative Study," *JMIR Ment. Health*, vol. 6, no. 1, p. e11973, Jan. 2019, doi: 10.2196/11973.

[132] Polar Electro, "Polar H10 Heart Rate Sensor," https://www.polar.com/ca-en/sensors/h10-heart-rate-sensor.

[133] Empatica Inc, "Empatica E4 wrsitband," https://www.empatica.com/research/e4/.

[134] K. Kutt *et al.*, "Affective Computing Experiments in Virtual Reality with Wearable Sensors. Methodological considerations and preliminary results Aective Computing Experiments in

Virtual Reality with Wearable Sensors. Methodological considerations and preliminary results." [Online]. Available: https://www.researchgate.net/publication/313841760

[135] N. Milstein and I. Gordon, "Validating Measures of Electrodermal Activity and Heart Rate Variability Derived From the Empatica E4 Utilized in Research Settings That Involve Interactive Dyadic States," *Front. Behav. Neurosci.*, vol. 14, Aug. 2020, doi: 10.3389/fnbeh.2020.00148.

[136] Meta, "Meta Quest 2," https://www.meta.com/ca/quest/products/quest-2/.

[137] H. Al Osman, H. Dong, and A. El Saddik, "Ubiquitous Biofeedback Serious Game for Stress Management," *IEEE Access*, vol. 4, pp. 1274–1286, 2016, doi: 10.1109/ACCESS.2016.2548980.

[138] M. SAKAKIBARA, S. TAKEUCHI, and J. HAYANO, "Effect of relaxation training on cardiac parasympathetic tone," *Psychophysiology*, vol. 31, no. 3, pp. 223–228, May 1994, doi: 10.1111/j.1469-8986.1994.tb02210.x.

[139] M. C. Jacob Rodrigues, O. Postolache, and F. Cercas, "The Influence of Stress Noise and Music Stimulation on the Autonomous Nervous System," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–19, 2023, doi: 10.1109/TIM.2023.3279881.

[140] R. Champseix, L. Ribiere, and C. Le Couedic, "A Python Package for Heart Rate Variability Analysis and Signal Preprocessing," *J. Open Res. Softw.*, vol. 9, no. 1, p. 28, Oct. 2021, doi: 10.5334/jors.305.

[141] D. Zhang, B. Wan, and D. Ming, "[Research progress on emotion recognition based on physiological signals].," *Sheng Wu Yi Xue Gong Cheng Xue Za Zhi*, vol. 32, no. 1, pp. 229–34, Feb. 2015.

[142] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[143] K. Patel, F. Safavi, R. Chandramouli, and R. Vinjamuri, "Transformer-Based Emotion Recognition with EEG," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/EMBC53108.2024.10781700.

[144] S. Salehizadeh, D. Dao, J. Bolkhovsky, C. Cho, Y. Mendelson, and K. Chon, "A Novel Time-Varying Spectral Filtering Algorithm for Reconstruction of Motion Artifact Corrupted Heart Rate Signals During Intense Physical Activities Using a Wearable Photoplethysmogram Sensor," *Sensors*, vol. 16, no. 1, p. 10, Dec. 2015, doi: 10.3390/s16010010.

[145] A. Gillioz, J. Casas, E. Mugellini, and O. A. Khaled, "Overview of the Transformer-based Models for NLP Tasks," Sep. 2020, pp. 179–183. doi: 10.15439/2020F20.

[146] Y. Qiang, X. Dong, X. Liu, Y. Yang, F. Hu, and R. Wang, "ECGMamba: Towards ECG Classification with State Space Models," in *Proceedings - 2024 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 6498–6505. doi: 10.1109/BIBM62325.2024.10822754.

[147] Philippe Remy, “Temporal Convolutional Networks for Keras,” 2020, *GitHub*.

[148] I. Agarwal, V. Sakthivel, and P. Prakash, “Toward Inclusive Healthcare: An LLM-Based Multimodal Chatbot for Preliminary Diagnosis,” *IEEE Access*, vol. 13, pp. 136420–136432, 2025, doi: 10.1109/ACCESS.2025.3594218.

[149] J. S. Beck, *Cognitive behavior therapy: Basics and beyond*. 2012.

[150] H. Chatoui and O. Ata, “Automated Evaluation of the Virtual Assistant in Bleu and Rouge Scores,” in *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, IEEE, Jun. 2021, pp. 1–6. doi: 10.1109/HORA52670.2021.9461351.

[151] Yajna Bopaiah, “Unveiling the Power of ROUGE Metrics in NLP,” https://pub.aimind.so/unveiling-the-power-of-rouge-metrics-in-nlp-b6d3f96d3363.

[152] T. Sellam and A. P. Parikh, “Evaluating natural language generation with bleurt,” https://research.google/blog/evaluating-natural-languagegeneration- with-bleurt/.

[153] M. Kaster, W. Zhao, and S. Eger, “Global Explainability of BERT-Based Evaluation Metrics by Disentangling along Linguistic Factors,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 8912–8925. doi: 10.18653/v1/2021.emnlp-main.701.

[154] Kolena, “Bertscore - testing with kolena,” https://docs.kolena.com/metrics/bertscore/.

[155] D. Banerjee, P. Singh, A. Avadhanam, and S. Srivastava, “Benchmarking LLM powered Chatbots: Methods and Metrics,” Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.04624

[156] D. Cecchini, A. Nazir, K. Chakravarthy, and V. Kocaman, “Holistic Evaluation of Large Language Models: Assessing Robustness, Accuracy, and Toxicity for Real-World Applications,” in *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 109–117. doi: 10.18653/v1/2024.trustnlp-1.11.