Multimodal Emotional State Recognition for Personalized Responses

Avishka Sathyanjana

2025

Multimodal Emotional State Recognition for Personalized Responses

Avishka Sathyanjana
Index No: 20001681

Supervisor: Dr. Enosha Hettiarachchi
Co-Supervisor: Mr. Amod Pathirana

May 2025

Submitted in partial fulfillment of the requirements of the B.Sc in Computer Science Final Year Project (SCS4224)

Declaration

I certify that this dissertation does not incorporate, without acknowledgement, any material previously submitted for a degree or diploma in any university and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, be made available for photocopying and for interlibrary loans, and for the title and abstract to be made available to outside organizations.

Student Name : B.A. Sathyanjana

2025 / 06 / 30

Signature & Date

This is to certify that this thesis is based on the work of Mr.B.A. Sathyanjana under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Supervisor Name : Dr. K.H.E.L.W. Hettiarachchi

2025/06/30

Signature & Date

Co Supervisor Name : Mr. Amod Pathirana

2025/06/30

Signature & Date

Abstract

Human emotions are complex, personal, and vary often across individuals, situations, and cultures. Many existing emotion recognition systems focus only on identifying emotional states at a given moment. This research aims to address that limitation by developing a personalized multimodal emotion recognition framework that identifies and adapts to each user's emotional baseline over time.

The framework combines facial and vocal signals using Decision Level Fusion, where Mean Squared Error (MSE) is used to assign personalized weights based on how close each modality's prediction is to user reported emotions. Kernel Density Estimation (KDE) method introduced to estimate the initial emotional baseline in the arousal-valence space. This baseline is further refined through reinforcement learning, using user feedbacks through emoji-based mechanism. Experiments were conducted across five emotional categories (Happy, Angry, Sad, Boredom, and Calm) using a group of 10 participants.

The fused method yields an average improvement of 33.92% over the facial method and 6.52% over the vocal method. Emotionally enhanced responses using personalized emotional inputs showed improvements of Empathy (75.3%) and Emotional Alignment (69.5%), followed by Satisfaction (37.6%). Most participants (66.67%) agreed with the computed refinement baseline values.

This research makes three main contributions: a personalized emotion fusion method, baseline identification, and an iterative refinement process. While the system currently supports a limited set of emotions and uses only facial and vocal inputs, it opens pathways for including more emotional categories, physiological data, and advanced context aware fusion techniques in future work.

Acknowledgment

I would like to express my heartfelt gratitude to my research supervisor, Dr. Enosha Hettiarachchi, and my co-supervisor, Mr. Amod Pathirana, for their invaluable guidance, encouragement, and continuous support throughout the course of this research. Their insights, expertise, and constructive feedback greatly contributed to the quality and direction of my study. Their encouragement have been instrumental in shaping this work and helping me navigate the challenges of academic research.

I am also sincerely grateful to Dr. Kasun Karunanayaka, Dr. Lasanthi De Silva, and Ms. Sanjani Gunathilaka of the University of Colombo School of Computing for their valuable feedback during the research proposal and interim evaluations, which helped me improve and refine my work.

A special thanks goes to my fellow colleagues and friends for their constant motivation and collaborative spirit, which kept me focused and inspired during challenging times. I would also like to acknowledge the participants of my study who generously volunteered their time and shared their experiences, without whom this research would not have been possible.

Above all, I am deeply thankful to my family for their unwavering support, patience, and love. Their encouragement has been a cornerstone of my academic journey.

It is with great pleasure that I acknowledge the support and contributions of all those who have helped me in numerous ways to successfully complete this research.

Contents

Abstract **ii**

Acknowledgment **iii**

1 Introduction **2**

1.1 Introduction 2

2 Literature Review **4**

2.1 Background 4

2.2 Facial Emotion Recognition 6

2.3 Speech Emotion Recognition 9

2.4 Multi-modal Fusion 13

2.5 Reinforcement Learning Approaches for Personalized Applications . 14

3 Motivation and Research Questions **16**

3.1 Motivation 16

3.2 Research Gap 16

3.3 Significance of the Research 17

3.4 Research Aim, Questions and Objective 17

3.4.1 Research Aim 17

3.4.2 Research Questions and Objectives 18

3.5 Research Methodology 19

3.5.1 Research Approach 19

3.6 Scope 19

3.6.1 In Scope 19

3.6.2 Delimitations 20

4 Design and Implementation 22

4.1 Methodology Overview 22

4.2 Phase 1 – Selecting Suitable Emotion Recognition Models 24

4.2.1 Facial Expression Experiment Setup 26

4.2.2 Vocal Emotion Experiment Setup 26

4.3 Phase 2 – Multimodal fusion with personalized emotion recognition 28

4.3.1 Emotion Eliciting Tasks 28

4.3.2 Emotion Mapping to A-V Plan 41

4.3.3 Personalized Multimodal Fusion 44

4.4 Phase 3 - Initial Baseline Identification 45

4.4.1 Initial Baseline Identification Using Kernel Density Estimation (KDE) 46

4.5 Phase 4 - Evaluating Responses from LLM with Emotional State . . 48

4.6 Phase 5 - Refining the Initial Baseline using Reinforcement Learning 54

4.6.1 Relationship with Emojis and Arousal-Valence Values 55

4.6.2 Emotion Refining Using Reinforcement Learning 56

5 Results and Evaluation 62

5.1 Participant Selection 62

5.2 Phase 1: Selecting Suitable Emotion Recognition Models 63

5.2.1 Facial Expression Experiment 63

5.2.2 Vocal Emotion Recognition Models Analysis 65

5.3 Phase 2: Personalized Multimodal Fusion 68

5.3.1 Personalized Weights 68

5.3.2 Multimodal Fusion Analysis 72

5.4 Phase 3: Initial Baseline Identification 76

5.5 Phase 4: LLM Response Evaluation 83

5.6 Phase 5: Baseline Refinement 84

6 Conclusion 87

6.1 Conclusions about the Research Questions and Aim 87

6.2 Research Contributions 89

6.3 Limitations and Future Work 90

A Appendix **99**

List of Figures

2.1 Emotion categorization 5

3.1 Action Plan 20

4.1 Main stages of the research methodology 22

4.2 Experimental flow of Phase 1 24

4.3 Experiment snapshot for facial expression recognition 26

4.4 Experimental flow of Phase 2 28

4.5 Arousal-Valence space with diverse emotion regions 29

4.6 Eliciting tasks for Happy Emotion 31

4.7 Eliciting tasks for Anger Emotion 34

4.8 Eliciting tasks for Sad Emotion 35

4.9 Eliciting tasks for Boredem emotion 37

4.10 Screenshots from the three tasks used to elicit calmness and the evaluation page 39

4.11 Snapshots of participants engaged in the emotion elicitation tasks. . 39

4.12 Circumplex Model of Affect Russell (1980). This circular model shows how emotions are distributed in a two-dimensional space using arousal and valence. 42

4.13 Emotion coordinates adapted from Paltoglou & Thelwall (2012). These values are used in this study to map categorical emotions to A-V values. 42

4.14 Experimental flow of Phase 2 49

4.15 Example of LLM prompt with emotional baseline data 52

4.16 Experimental flow of Phase 5 55

4.17 Emoji feedback interface used for lightweight user input. 57

5.1 Emotion categorization results for facial expression experiment . . . 63

5.2 Emotion intensity identification results for facial expression experiment 64

5.3 Vocal emotion categorization results: HUME vs Wave2Vec2 66

5.4 Vocal emotion intensity identification: HUME vs Wave2Vec2 66

5.5 Average fusion weights per emotion category 69

5.6 Weight distribution boxplot across all participants 69

5.7 Fusion weight visualization for each participant 70

5.8 Performance comparison by Euclidean distance for each emotion . . 74

5.9 Improvement heatmaps: Fused vs Facial and Fused vs Vocal by Emotion and Intensity 75

5.10 Statistical improvement (percentage) by emotion 76

5.11 Initial Emotional data mapping for all participants 77

5.12 3D surface plots of KDE-based emotional distributions for all participants 79

5.13 2D mapping of KDE-estimated baseline zones for each participant . 80

5.14 Participant agreement frequency with identified baseline values . . . 81

5.15 Scatter plot comparing identified and participant-proposed baseline coordinates 82

5.16 Comparison of average ratings for LLM responses 83

5.17 Initial and refined baseline regions for selected participants. 85

A.1 Baseline evaluation Questionnaire Screenshot 1 103

A.2 Baseline evaluation Questionnaire Screenshot 2 104

A.3 Baseline evaluation Questionnaire Screenshot 3 105

A.4 Baseline Refinement Questionnaire Screenshot 1 106

A.5 Results for the questionnaire 107

A.6 Informed Consent Form Page 1 108

A.7 Informed Consent Form Page 2 109

List of Tables

2.1 Comparison of Suitable Datasets 7

2.2 Comparison of Emotion Recognition Models 9

2.3 Comparison of Speech Emotion Recognition Datasets 10

2.4 Comparison of Speech Emotion Recognition Studies 13

2.5 Summary of Reinforcement Learning Approaches for Personalized Applications 15

4.1 Summary of Emotion Eliciting Experiments 40

4.2 Emotion Intensity Coordinates Based on Paltoglou & Thelwall (2012) 43

4.3 Emotional Intelligence Benchmarking of LLMs (adapted from Wang et al. (2023)) 50

4.4 Clusters of emojis and their valence-arousal values from Kutsuzawa et al. (2022). 56

5.1 Mean Euclidean Distances for Each Method 74

5.2 T-test Results for Fused Method Comparisons 74

5.3 Participant baseline values estimated from KDE peak density . . . 78

5.4 Participant Agreement and Baseline Discrepancy Summary 82

5.5 Comparison of participant evaluations for control vs emotionally enhanced responses 84

5.6 Participant Agreement on Refined Baselines 86

A.1 Self-reported values for **Happy** emotion 99

A.2 Self-reported values for **Angry** emotion 100

A.3 Self-reported values for **Sad** emotion 100

A.4 Self-reported values for **Boredom** emotion 100

A.5 Self-reported values for **Calm** emotion 101

A.6 Modality Weights and MSE for Participants (Left: 1–5, Right: 6–10) 102

Listings

4.1 Emotion-specific fusion weights per user 44
4.2 Gaussian KDE implementation 46
4.3 Baseline region identification with KDE 47
4.4 Example of emotional data sent to LLM 50
A.1 Model Parameters 110
A.2 State and Coordinate Conversion 110
A.3 Action Selection and State Transition 111
A.4 Reward Calculation 112
A.5 Placeholder Functions for Data Input 112
A.6 Q-Learning with Eligibility Traces 113

1

Introduction

1.1 Introduction

Envision a scenario where your devices not only understand your words but also your emotions, responding with empathy and personalized interactions tailored to your emotional state. This vision drives research in multi-modal emotional state recognition, a field at the intersection of affective computing, human-computer interaction, and artificial intelligence. Affective computing is a multidisciplinary research area where computer science bridges the gap between cognitive science, psychology, and social science. It empowers intelligent systems to recognize, conclude, and interpret human emotions, facilitating better human-machine interaction by responding to humans based on their emotional state (Picard 2000).

The emotional state of individuals varies significantly from person to person (Lim, 2016), making it essential to create customized and personalized models for emotion recognition. This is particularly important for applications like personal assistants, where user satisfaction and engagement are paramount (Salama AbdELminaam et al. 2020). Personalization in emotion identification is essential for building truly adaptable and empathetic systems that can transform industries like healthcare, customer service, education, and entertainment. Research by Gelbrich et al. (2021), Mariacher et al. (2021), and Inkster et al. (2018) has expanded this area, showing the potential of emotionally aware AI systems.

This study focuses on building a personalized emotion recognition system by first selecting the most effective facial and speech emotion models. These models are then fused to capture emotions more accurately while considering individual differences. An emotional baseline is established for each user, which is later refined through reinforcement learning. The system also explores how incorporating emotional context can improve responses from large language models during user interactions.

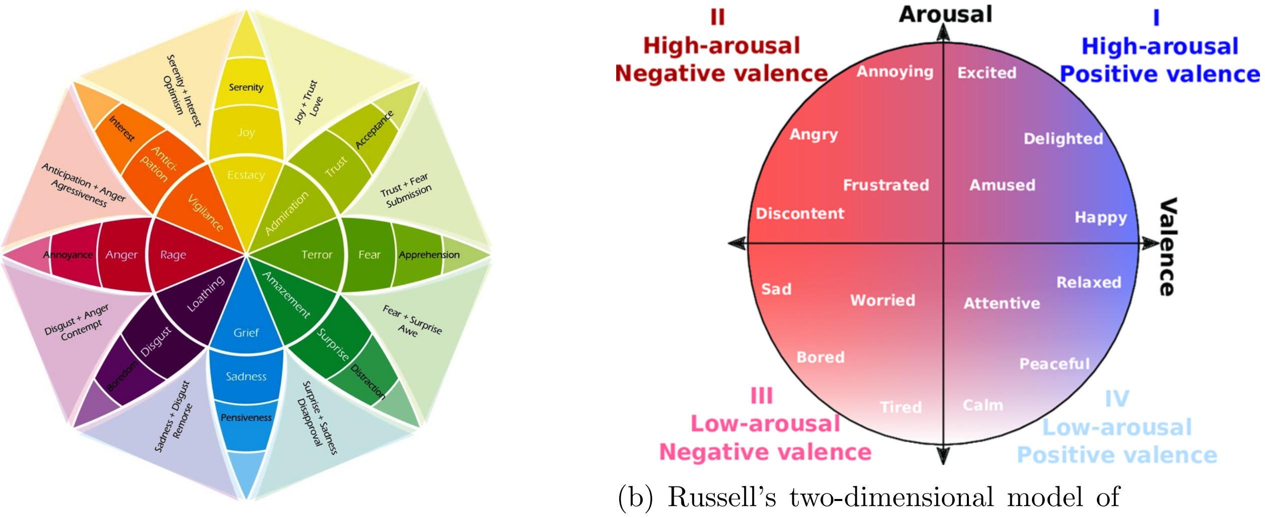
2

Literature Review

2.1 Background

Emotions are complex experiences of consciousness, bodily sensation, and behavior that reflects the personal significance of a thing, or an event. Plutchik introduced the Wheel of Emotions, a model that identifies eight fundamental emotions: joy, sadness, trust, disgust, fear, anger, anticipation, and surprise. According to this model, these basic emotions can combine in varying intensities to form more complex emotional states. Other emotions are overlapping those fundamental emotions which can be seen in Figure 2.1a. This research is the base for upcoming research in emotion categorization (Plutchik 1980). Valence-Arousal-Dominance (VAD) model represents emotions along three dimensions: pleasantness (valence), activation (arousal), and control (dominance). It provides a simple and intuitive way to understand the different components of emotion Figure 2.1b. (Oberlander & Klinger 2018).

These emotions are expressed using both verbal and nonverbal channels. As Ekman & Friesen (1971) showed, facial expressions universally convey six basic emotions across cultures. Our voices carry emotional information through tone, pitch, and pace, with high-arousal emotions often involving higher pitch and faster pace (Barrett, 2004). Body language, such as posture and gestures, also communicates emotional states (Ekman & Friesen 1971). Verbally, we directly express our feelings using linguistic patterns and exact words reflecting higher emotional granularity



(a) Wheel of emotions

(b) Russell's two-dimensional model of valence and arousal

Figure 2.1: Emotion categorization

(Smidt & Suvak 2015). In today's digital age, we frequently express emotions in text, a focus of sentiment analysis in natural language processing (Lim 2016). Physiological changes like increased heart rate also signal emotions, particularly their arousal levels (Barrett et al. 2004).

However, the expression of the emotions varies significantly from person to person and across different cultures. Emotional granularity, which is a factor in this variability, is the ability to make fine-grained distinctions between similar emotions (Smidt & Suvak 2015). The cultural differences profoundly influence emotional arousal and expression. For instance, research by (Lim 2016) highlights that Western cultures tend to value and promote high-arousal emotions such as excitement and anger, whereas Eastern cultures prioritize low-arousal emotions like calmness and contentment. Because of that, it's important to establish a baseline behavior of individuals in advance to identify their emotional state in a more precise manner.

Baseline behavior refers to an individual's typical or normal pattern of behavior, thoughts, and emotions when they are not experiencing any specific external stimuli or circumstances that would significantly influence their state. It represents the default or resting state from which deviations or changes can be measured. Emotional baseline, also known as emotional homeostasis or emotional equilibrium, is the relatively stable state of emotional experience that an individual tends to return to after experiencing temporary emotional fluctuations. It is the individual's characteristic or typical level of emotional arousal and emotional experience when not influenced by external events or stimuli. (Davidson 1998)

2.2 Facial Emotion Recognition

Facial expressions occur by contracting and releasing of the muscles under the skin. Most existing applications use the movements of facial muscles considering action units and this uses supervised learning approach which is known as Facial Action Coding System (FACS) (Kantharia & P. 2015). By using the CNN, the emotion recognition can be done in real time. This Technology is known as Facial Expression Recognition using CNN (FERC). FERC has two parts, first it removes background and noises from the source and second part entirely focuses on extracting facial features. FERC model uses Expressional Vector to identify the basic emotions, and this was able to achieve 96% accuracy (Mehendale 2020)

Our research interest is mostly on the continuous emotion prediction models rather than distinct emotion classification models. Study done by Savchenko (2024) proposes EmotiEffNet family of models (a series of pre-trained convolutional networks) for valence-arousal prediction. These networks extract frame-level features which are fed into a Multi Layer Perceptron (MLP) and a LightAutoML classifier ensemble, with a post-processing step using smoothing to stabilize results across sequential frames. The Aff-Wild2 (Kollias & Zafeiriou 2018) dataset is used in this study. Model achieved a valence CCC of 0.5603 and an arousal CCC of 0.5597 on the Aff-Wild2 dataset. This model is efficient in terms of computational requirements, reaching 45 FPS on a GPU and 12 FPS on a CPU, with a model size of 15MB, making it adaptable for mobile deployment.

The MaxViT model (Wagner et al. 2024) utilizes a hybrid approach, combining continuous valence-arousal labels with discrete emotion categories to train a transformer-based architecture. This circumplex model-guided inference provides a more nuanced approach by learning from both types of labels, thus enhancing expression recognition accuracy. The model is trained with AffectNet (Mollahosseini et al. 2017) and EMOTIC (Kosti et al. 2019) datasets. CAGE achieved a CCC of 0.716 for valence and 0.642 for arousal on AffectNet, with a root mean square error (RMSE) reduction of 7% for valence and 6.4% for arousal. This model has a larger footprint, with a size of 86MB, and processes at 30 FPS on a GPU and 8 FPS on a CPU, making it suitable for high-end applications where computational resources are available.

Study done by Savchenko (2023) introduce EmotiEffNet model which leverages the EfficientNet-B0 architecture for frame-level feature extraction, followed by an MLP and LightAutoML classifier ensemble for downstream emotion analysis tasks. This approach is applied to predict VA, FER, and AU tasks within the video frames. Aff-Wild2 is again the primary dataset and the model achieves a valence CCC of 0.494 and an arousal CCC of 0.607. For FER, the F1 score reaches 0.433, while the AU detection F1 score is 0.486. This balance between accuracy and computational efficiency results in an inference speed of 40 FPS on GPU and 15 FPS on CPU, with a model size of 23MB, providing a suitable option for applications needing moderate computational efficiency.

Table 2.1: Comparison of Suitable Datasets

Aspect	**AffWild2**	**AffectNet**	**EMOTIC**	**Hume Facial Dataset**
Total Frames/Images	2.8M frames (545 videos)	320,739 (train) + 41,406 (val)	23,266 (train) + 7,203 (test)	452,783 mimic images + 534,459 judgments
Annotations	VA (-1 to +1), 8 expressions, 12 AUs	8 expressions, VA (-1 to +1)	26 expressions, VA, arousal, dominance	28+ continuous emotional dimensions
Resolution	Varying (in-the-wild)	High-resolution	Context-rich, full-body	160x160 pixels (standardized)
Annotators	Expert-labeled	12 professionals	Multi-annotator settings	5,833 participants (6 countries)
Key Features	Video-based dynamic expressions	Large-scale static images	Contextual full-body analysis	Cross-cultural mimicry, controls for demographic bias

A comparison of suitable datasets for our study is shown in Table 2.1. These datasets vary in terms of labeled expressions, complexity, and real-world applicability. While AffWild2 and AffectNet focus on detailed face annotations, EMOTIC extends beyond facial analysis to include body language in contextual settings.

Hume.ai has developed sophisticated facial expression analysis technology centered on "semantic space theory,"(Cowen & Keltner 2021) which enables a high-dimensional, data-driven understanding of human emotional expressions. This approach transcends traditional models by capturing hundreds of dimensions of human expression, allowing for the identification of subtle emotional nuances. Their Facial Expression Model identifies over 28 distinct facial expressions by analyzing millions of natural facial expressions collected from a diverse global population across six countries (USA, China, India, Venezuela, Ethiopia, and South Africa), comprising over 452,783 participant-generated mimic images and 534,459 emotion judgments from 5,833 participants. This extensive cross-cultural dataset deliberately controls for demographic variables to isolate genuine emotional signals, addressing limitations of culturally homogeneous samples in previous research.(Brooks et al. 2024)

The technology relies on a DNN model built on a FaceNet Inception ResNet v1 architecture pretrained on the VGGFace2 dataset that specifically analyzes facial features. Their FACS 2.0 Model provides an enhanced automated version of the Facial Action Coding System that measures 26 facial action units and 29 other facial features, offering detailed breakdown of facial movements that contribute to emotional expressions. The company employs principal preserved components analysis (PPCA) and generalized PPCA (G-PPCA) to extract 28 shared or culture-specific emotional dimensions from facial data. Unlike traditional models constrained to basic emotion categories, Hume.ai's system evaluates performance via correlation with human ratings of facial expressions rather than conventional metrics, ensuring alignment with real-world emotional interpretations while minimizing demographic biases(Brooks et al. 2024).

Hume.ai's expression measurement technology is particularly valuable for practical applications through its WebSocket-based streaming capabilities, which facilitate real-time data processing without burdening local machines. This API-based approach enables continuous data flow between applications and Hume's models, providing immediate feedback on facial expressions through persistent two-way communication optimized for high throughput and low latency (Hume AI 2025*b*). The system can process various media formats with reasonable size limits (images up to 3,000 x 3,000 pixels, video up to 5 seconds) and offers both REST endpoints for batch processing and WebSocket endpoints for real-time predictions from sources like webcam streams. This infrastructure makes the sophisticated facial analysis technology highly accessible for applications requiring instant processing such as live customer service tools with the computational complexity handled on Hume.ai's servers rather than client devices (Hume AI 2025*a*).

Table 2.2: Comparison of Emotion Recognition Models

Aspect	**MT-EmotiDDA MFN**	**MaxViT**	**EmotiEffNet**	**Hume Facial Expression**
Dataset	Aff-Wild2	AffectNet, EMOTIC	Aff-Wild2	Hume Facial Dataset
Model Architecture	MobileNetV3 (lightweight MTL)	MaxViT (transformer-based hybrid)	EfficientNet-B0 with MLP ensemble	FaceNet Inception ResNet v1
Post-processing	Gaussian/Box filters	Not specified	Box filtering for smoothing	MTCNN face detection, 160x160 pixel standardization
Suitability	Real-time, mobile deployment	High-end applications	Moderate real-time applications	Cross-cultural, Real time

The table 2.2 provides a summarized, side-by-side comparison of each model's architecture, performance metrics, inference efficiency, and suitability for specific use cases, highlighting key findings.

2.3 Speech Emotion Recognition

Our study focusing on identify the arousal valence values of speech emotions. Although there are many categorical datasets in SER, Only few has data with arousal, valence labeled and also in English language. Table 2.3 shows a detailed comparison on available SER datasets.

Dataset	**Year**	**Content**	**Emotions**
RAVDESS	2018	7,356 recordings by 24 actors	7 emotions: calm, happy, sad, angry, fearful, surprise, disgust
MuSe-CAR	2021	40 hours, 6,000+ recordings of 25,000+ sentences by 70+ speakers	Continuous dimensions: valence, arousal, trustworthiness
Morgan Emotional Speech Set	2019	999 spontaneous voice messages from 100 speakers	Valence, arousal, 4 emotions: happiness, anger, sadness, calm
OMG Emotion	2018	420 videos, avg. length 1 min	7 emotions categories; valence, arousal
IEMOCAP	2007	12 hours, 5 sessions, 10 actors	15 emotion categories; valence, arousal and dominance
HUME-VB	2023	282,906 vocalizations from 4,080 participants across 5 countries	48 emotion categories and 24 emotional dimensions
HUME-Prosody	2023	5,000+ "seed" samples and 282,906 trials of crowd-sourced mimicry responses across English, Mandarin, Spanish, Hindi	48 emotion categories with continuous values

Table 2.3: Comparison of Speech Emotion Recognition Datasets

The study Zhang et al. (2017) utilizes the IEMOCAP dataset with CNNs for emotion recognition, extracting MFCCs, pitch, and prosodic features. The model achieved better performance on high-arousal emotions like anger and happiness but had challenges with neutral, low-arousal states. CNNs effectively captured emotional features relevant to both arousal and valence dimensions, highlighting the importance of carefully selected acoustic features in SER.

Another study done by Martinez-Lucas et al. (2020) Leveraging the MSP-Podcast corpus with time-continuous annotations, this study applied RNNs, particularly LSTMs, to capture sequential dependencies in speech emotions. By modeling continuous changes in emotions, the model achieved high correlation scores in arousal and valence prediction, making it well-suited for applications where emotions evolve over time.

Using the RAVDESS dataset, the study Jalal et al. (2019) explored LSTMs to enhance emotion detection, especially for low-arousal states. Focusing on features like pitch, energy, and spectral elements, the LSTM-based model demonstrated effectiveness in classifying nuanced emotions, achieving significant accuracy gains on emotions with subtle arousal shifts. The study supports the use of LSTMs in capturing temporal patterns, particularly for more subtle, low-arousal emotions.

This recent study Wagner et al. (2023) applied transformer architectures like wav2vec 2.0 and HuBERT, pre-trained on large audio datasets, for SER across MSP-Podcast, IEMOCAP, and MOSI. Transformers achieved state-of-the-art performance in valence recognition, with results revealing robust performance across diverse conditions and fairness in gender representation. The study showed that fine-tuning transformers with continuous annotations allows them to implicitly capture linguistic cues, which significantly improves valence prediction.

Hume.ai's Vocal Burst dataset (HUME-VB) represents a groundbreaking resource for emotion recognition research, comprising 282,906 vocalizations from 4,080 participants across five culturally diverse countries (USA, China, India, Venezuela, South Africa) spanning multiple languages (English, Mandarin, Spanish, Hindi). This extensive dataset captures emotional expressions in real-world conditions with varied recording environments, making it the largest of its kind. The dataset has been leveraged in two significant ways: first, done by Brooks et al. (2023) to train DNNs that predict 48 emotion categories from vocal bursts, revealing that nonverbal vocalizations express 24 distinct emotional dimensions with 79% cross-cultural consistency, supporting Semantic Space Theory which conceptualizes emotions as continuous multi-dimensional states rather than discrete categories. Second done by Tzirakis et al. (2023), used the dataset to develop transformer-based models, particularly Whisper architectures for detecting and classifying 67 vocalization types in audio streams, with the best-performing models achieving F1-scores of 96.2% even in challenging noisy environments. These studies demonstrate the HUME-VB dataset's value in advancing understanding of cross-cultural emotional communication while providing practical applications for affective computing technology.

Hume.ai's Speech Prosody dataset (Hume-Prosody Corpus, HP-C) represents another significant contribution to emotional expression research, containing over 5,000 "seed" samples of emotional vocalizations and 282,906 trials of crowd-sourced mimicry responses collected across multiple languages (English, Mandarin, Spanish, Hindi) and cultures (USA, China, India, Venezuela, South Africa). This dataset was prominently featured in the 2023 Computational Paralinguistics Challenge (ComParE), where researchers tackled the "Emotion Share" task of predicting continuous emotion proportions across 48 emotion categories in speech segments Schuller et al. (2023). The challenge evaluated models using Spearman's rank correlation metrics, with baseline approaches including both modern transformer-based systems (Wav2Vec2) and traditional acoustic feature engineering (OpenSMILE). Research findings revealed that models struggled particularly with low-prevalence emotions, highlighting the need for balanced datasets, while also demonstrating significant cross-cultural variability in emotion expression. This work established important benchmarks for emotion share prediction while suggesting that future advances could come from combining acoustic and linguistic features, extending the dataset's utility for developing robust speech-based affective computing systems.

In summary, emotion recognition research has evolved from foundational psychological theories like Plutchik's wheel and the VAD model to sophisticated multimodal systems that analyze facial, vocal, and textual cues. Modern FER models such as EmotiEffNet and MaxViT demonstrate strong performance in valence-arousal prediction, balancing accuracy with computational efficiency. Hume.ai's approach offers a culturally diverse and highly detailed understanding of emotional expression. Similarly, SER leverages CNNs, LSTMs, and transformers to model emotional variations in audio, with recent advances achieving state-of-the-art results using large, diverse datasets. Together, these developments underscore the

Study	Dataset	Methodology	Key Results	Performance
CNN on IEMOCAP	IEMOCAP	CNN, MFCC, pitch, prosodic	High accuracy on high-arousal	82% for anger/happiness, 65% for neutral
RNN on MSP-Podcast	MSP-Podcast	RNN, LSTM, time-continuous annotation	Continuous tracking of emotions	Correlation: 0.403 (arousal), 0.196 (valence)
LSTM on RAVDESS	RAVDESS	LSTM, pitch, energy, spectral features	Low-arousal detection improved	75% accuracy for primary emotions, 68% for low-arousal
Transformer-based Models on MSP-Podcast	MSP-Podcast, IEMOCAP, MOSI	Transformer (wav2vec 2.0, HuBERT)	Robust valence recognition, state-of-the-art on valence	CCC of 0.638 on MSP-Podcast for valence

Table 2.4: Comparison of Speech Emotion Recognition Studies

importance of multimodal, culturally-aware, and continuously annotated datasets for advancing emotion recognition technologies.

2.4 Multi-modal Fusion

Multimodal analysis leverages input data from various channels like video, audio, and text to enhance the performance and accuracy of emotion recognition systems. The fusion of this multi-modal data is crucial, with techniques including feature-level fusion (combining features into one vector), decision-level fusion (independently classifying features and fusing outcomes), hybrid fusion (combining feature and decision-level approaches), model-level fusion (using correlations between models), rule-based fusion (assigning normalized weights), classification-based fusion (employing algorithms like SVMs and neural networks), and estimation-based fusion (useful for real-time audio and visual data, with filters like Kalman and particle filters). These fusion methods aim to effectively combine the data gathered from multiple modalities, enabling better emotional classification and recognition (Poria et al. 2017).

In summary, emotion recognition research has evolved from foundational psychological theories like Plutchik's wheel and the VAD model to sophisticated multimodal systems that analyze facial, vocal, and textual cues. FER models such as EmotiEffNet and MaxViT demonstrate strong performance in valence-arousal prediction, balancing accuracy with computational efficiency. Hume.ai's approach offers a culturally diverse and highly detailed understanding of emotional expression. Similarly, SER leverages CNNs, LSTMs, and transformers to model emotional variations in audio, with recent advances achieving state-of-the-art results using large, diverse datasets. Together, these developments underscore the importance of multimodal, culturally-aware, and continuously annotated datasets for advancing emotion recognition technologies.

2.5 Reinforcement Learning Approaches for Personalized Applications

Reinforcement Learning (RL) is a powerful method for building personalized systems. It helps systems to learn and improve based on user interactions over time. Many RL techniques have been used in different fields like entertainment, healthcare, education, and e-commerce. This section explains some popular RL approaches that are used for personalization.

Contextual Bandits

This method is used by Netflix to show different artwork (thumbnails) for movies and series to different users. It learns from the user's past behaviour and decides which artwork will attract the user more. It balances between trying new options (exploration) and using known successful options (exploitation) Blog (2018).

Proximal Policy Optimization (PPO)

In healthcare, PPO has been used to suggest personalized cancer treatments. It works by ranking different drugs based on the data from each patient. This helps doctors to choose better treatments that match each patient's unique condition Liu et al. (2022).

Deep Q-Learning (DQN)

DQN has been applied to recommend projects to users based on their interests. It improves the accuracy of recommendations and helps users to trust the system more Wang et al. (2019).

Q-Learning

This is a simple and popular RL method. It can be used for many general recommendation tasks and is good at adapting to changes in user preferences Edirisinghe (2020).

A study done by Moise et al. Moise et al. (2020) reviewed 166 research papers and found that Q-learning is the most commonly used RL method (used in 60 studies). It has been used in many areas like healthcare, entertainment, education, and commerce. According to the study, commerce and entertainment domains often use realistic experiments, while healthcare and communication have limitations due to safety and data issues.

Approach	**Application**	**Domain**	**Key Features**
Contextual Bandits	Artwork Personalization	Entertainment	Balances exploration/-exploitation, scalable
Proximal Policy Optimization (PPO)	Treatment Recommendation	Healthcare	Ranks drugs using DRL, handles high-dimensional data
Deep Q-Learning (DQN)	Project Recommendations	Recommendation	Learns user preferences, boosts trust
Q-Learning	General Recommendations	Recommendation	Adapts to changing preferences, versatile

Table 2.5: Summary of Reinforcement Learning Approaches for Personalized Applications

3

Motivation and Research Questions

3.1 Motivation

While progress in affective computing has enabled machines to recognize emotions in the moment, a critical challenge remains: capturing the dynamic, personalized nature of emotional experiences. This leads to frustration and dissatisfaction, as these systems provide generic, one-size-fits-all responses that fail to account for user's emotional context (Kim et al. 2024).

When emotions are triggered, they emerge from a foundational **"Baseline"** or mood, representing our stable emotional state. Emotional changes can be understood as shifts or escalations from this baseline (Davidson 1998). By disregarding the baseline, these systems lack an essential element of emotional understanding, leading to responses that may feel disconnected or inadequate. For instance, in urgent situations, long responses can increase frustration, while in joyful moments, generic responses can reduce engagement.

3.2 Research Gap

Most current systems overlook the significance of longitudinal analysis and fail to capture individual baseline behavior. Neglecting these unique emotional baselines can compromise accuracy of emotion identification, as readings may not align with an individual's natural emotional biases. This research aims to bridge this gap by developing a framework that collects long-term multimodal data to understand each person's emotional baseline, enabling adaptive shifts in these baselines over time. By considering individuals' baseline behavior alongside a multimodal approach combining facial expressions and interaction patterns, the proposed system strives to achieve precise, personalized, and context-aware emotion recognition.

3.3 Significance of the Research

This research addresses a critical gap: the lack of emotional awareness and personalized responses in current AI systems. By developing systems that accurately recognize users' emotional states through multimodal data in a more personalized way using fine-tuned weights according to user emotions.

Also, it captures the baseline of users' emotion in context, and system continuously monitors for deviations, identifying subtle emotional shifts that represent changes from user's typical emotional state. Through an iterative process, the system will adjust the baseline of the user from time to time.

This personalized emotional insight is crucial in domains where ongoing and accurate understanding of user emotions is essential, such as mental health support, customer service, and adaptive learning. By focusing on both baseline and changes in emotional states, our approach paves the way for interactions that are genuinely responsive, supportive, and adaptive to users' evolving emotional needs.

3.4 Research Aim, Questions and Objective

3.4.1 Research Aim

To develop and evaluate a personalized emotion recognition framework that uses facial expressions and audio data to identify an individual's emotional baseline, with the goal of enhancing the emotional intelligence of LLM responses.

3.4.2 Research Questions and Objectives

Objective 1 - Develop a multi-modal emotional recognition framework with Personalized Arousal-Valence Identification

- **RQ 1.1:** What are suitable pre-implemented models that can be used to get a higher accuracy for emotion recognition?
- **RQ 1.2:** How the recognized emotion values from different modalities fused together in order to get more personalized arousal-valence value?
- **Approach:** Conduct a comprehensive literature review to identify suitable pre-trained ML models along with the suitable datasets for each modality. Participants will engage in emotion-eliciting tasks to gather real-time data, enabling fine-tuning of the fusion technique. User feedback will guide adjustments, refining the model to accurately capture personalized arousal-valence values, thereby enhancing the system's adaptability to individual emotional responses.

Objective 2 - Identify Initial Baseline and Implement Dynamic Personalized Baseline Identification Using Reinforcement Learning Iterations

- **RQ 2.1:** What techniques are most suitable for establishing an initial emotional baseline and how can this baseline be dynamically adjusted over time to reflect changes in the user's emotional responses and self-reported feedback?
- **Approach:** Begin by evaluating various techniques to determine the best fit for identifying an initial emotional baseline using data from emotional-eliciting tasks. Implement an iterative adjustment process where participant feedback and ongoing data inputs help refine the baseline over time. This approach enables the model to dynamically adapt to individual users.

Objective 3 - Integrate the personalized emotional state information with user queries and measure the impact on the quality of responses generated by a LLM, compared to a control condition without emotional state input

- **RQ 3.1:** How does integrating personalized emotional state information with user queries affect the relevance and user satisfaction of responses from LLMs?
- **Approach:** Participants will interact with a LLM by providing queries and receiving two responses: one from passing the raw query to the LLM (control condition), and another one combining the personalized emotional state information along with the user query (experimental condition).

3.5 Research Methodology

3.5.1 Research Approach

The planned approach for this research is a combined approach using **Action research and mixed methods**.

Action research is suitable for this research because we plan to use iterative cycles of action, observations, and the participatory nature of the research. Participants feedback and information will continuously refine the emotion baseline recognition system, ensuring the research addresses practical challenges and generates actionable knowledge.(figure 3.1)

The mixed-methods approach integrates quantitative data with qualitative insights, providing a comprehensive evaluation of the framework's effectiveness. Quantitative data will be used to train and validate the machine learning model, while qualitative feedback will offer deeper insights into the participants' experiences and the system's practical applicability.

3.6 Scope

3.6.1 In Scope

1. Use existing pre-trained models for multimodal emotion recognition
   - Our key contribution lies in effectively combining facial and audio emotion recognition modalities. By fusing data from facial and audio inputs, we can cross-validate and enhance overall emotion detection accuracy.

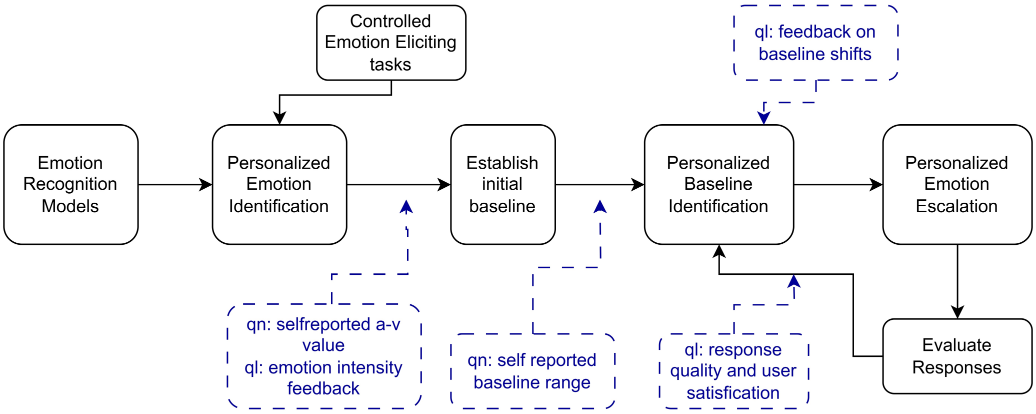

Controlled Emotion Eliciting tasks
ql: feedback on baseline shifts
Emotion Recognition Models
Personalized Emotion Identification
Establish initial baseline
Personalized Baseline Identification
Personalized Emotion Escalation
Evaluate Responses
qn: selfreported a-v value
ql: emotion intensity feedback
qn: self reported baseline range
ql: response quality and user satisfaction

Figure 3.1: Action Plan

2. Develop personalized, adaptive emotional baselines model

    - Many emotion recognition systems use universal, static baselines, leading to inaccuracies as individual behaviors change over time. Our innovative approach is to establish and continuously update personalized baselines for each user.

3. Develop an application that captures data to recognize the user's emotional state and facilitates interaction with large language models.

    - This application will use prompt engineering to combine the user's emotional state with their query, enhancing the relevance and personalization of the responses from the LLM.

3.6.2 Delimitations

1. Develop new facial and audio emotion modals from scratch

    - Given the abundance of highly accurate, well-established models in each domain, reimplementing these is unnecessary. Our focus is on multimodal integration and personalization, not on improving individual modalities.

2. Modify core AI models (GPT architecture)

    - Our scope is to enhance LLM responses by providing emotional context, not to alter the fundamental architecture of LLMs.

3. Making a dataset suited to the cultural context of Sri Lankans.

4

Design and Implementation

4.1 Methodology Overview

The research methodology is designed to address the research objectives mentioned in Section 3.4.2. This chapter first presents a general idea of the research process and then provides a detailed explanation of each stage, including how it was implemented and the experiments conducted at each step.

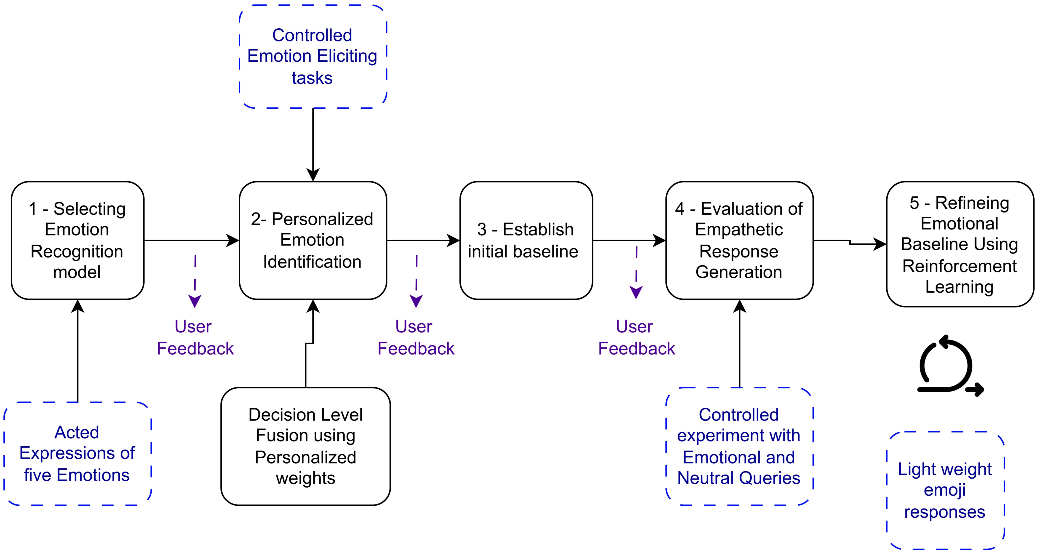

Controlled Emotion Eliciting tasks
1 - Selecting Emotion Recognition model
2- Personalized Emotion Identification
3 - Establish initial baseline
4 - Evaluation of Empathetic Response Generation
5 - Refineing Emotional Baseline Using Reinforcement Learning
User Feedback
User Feedback
User Feedback
Acted Expressions of five Emotions
Decision Level Fusion using Personalized weights
Controlled experiment with Emotional and Neutral Queries
Light weight emoji responses

Figure 4.1: Main stages of the research methodology

The whole research is divided into five main stages. The overall design of these stages is illustrated in Figure 4.1. Each stage is designed in a way that supports the next stage and contributes to the final objective of the research. A short overview of each stage is given below, and a more detailed explanation is provided in the following sections of this chapter.

1. **Selecting emotion recognition model:** In this stage, the main focus is to identify effective models that can recognize human emotions using both facial expressions and speech signals. Several existing models were reviewed and tested to choose the most suitable ones to address Research Question 1.1.

2. **Personalized emotion identification:** After selecting the models, the next step is to combine both facial and speech data for better emotion recognition. Also, the system is adjusted to consider individual differences by using dynamic weighting machanism by addressing Research Question 1.2, making it more personalized.

3. **Establish initial Baseline:** This stage involves setting up the Emotional Baseline which can be used to identify the emotional mood of the user to address Research Question 2.1. The baseline is established using data collected from participants during emotion-eliciting tasks. The aim is to create a reference point for each participant that can be used to measure deviations in their emotional state.

4. **Evaluation of Empathyic Response Generation:** In this stage, participants interact with a LLM by submitting their queries. Two types of responses are collected: one using the original query, and another combining the query with the user's current emotional state using prompt engineering. The goal of this stage is to observe how emotional context can influence the quality and personalization of responses given by the LLM. This stage is designed to address Research Question 3.1.

5. **Refining the Emotional Baseline using Reinforcement Learning:** Finally, the initial baseline is improved using reinforcement learning techniques by addressing Research Question 2.1. This allows the system to learn and improve over time based on feedback and performance.

Each of these stages will be discussed in detail in the following sections. Experiments, implementation methods, tools, and results related to each stage will also be described.

4.2 Phase 1 – Selecting Suitable Emotion Recognition Models

After conducting a literature review as discussed in Sections 2.2 and 2.3, the next task involved identifying and setting up suitable pre-trained models for emotion recognition. The experimental flow is illustrated in Figure 4.2.

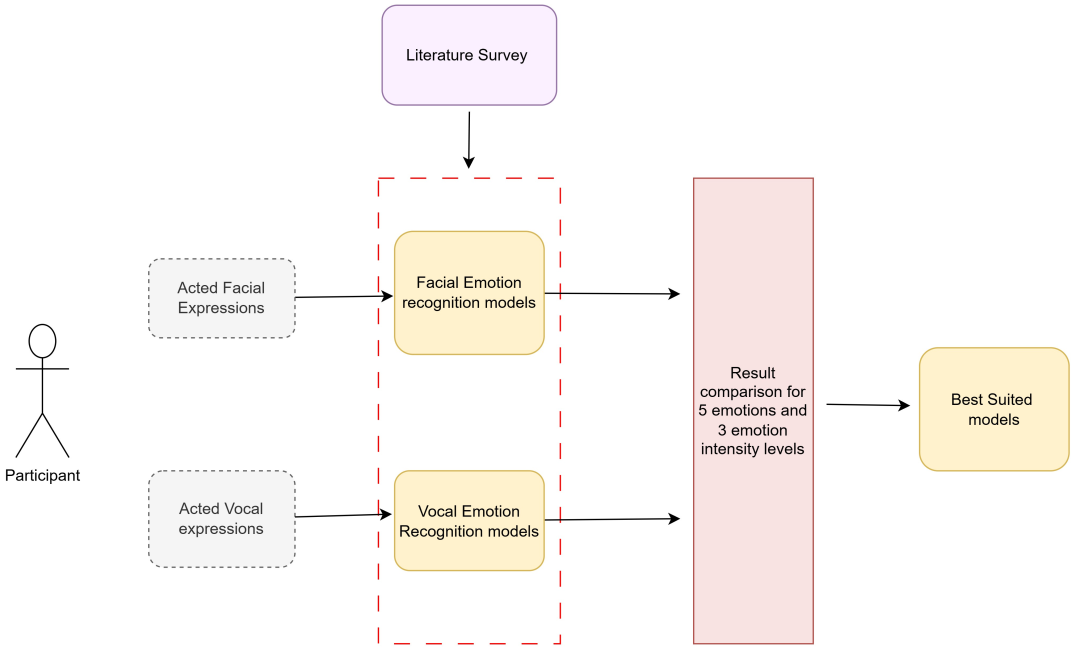

Literature Survey
Acted Facial Expressions
Facial Emotion recognition models
Participant
Acted Vocal expressions
Vocal Emotion Recognition models
Result comparison for 5 emotions and 3 emotion intensity levels
Best Suited models

Figure 4.2: Experimental flow of Phase 1

We initially explored the EmoNet model (Toisoul et al. 2021), available at https://github.com/face-analysis/emonet, which is well-regarded for arousal-valence detection. However, it presented several technical challenges due to its outdated dependencies, as it was developed in 2021. Although We was able to configure the environment, the model took over four seconds to process a single frame on a CPU, making it unsuitable for real-time applications.

To overcome these issues, We explored alternative options and identified the CAGE expression inference model (Wagner et al. 2024), which is optimized for real-time arousal-valence detection.

Additionally, We found Hume AI's expression recognition API to be valuable for practical applications. Hume provides real-time facial expression predictions through WebSocket-based streaming, allowing continuous data flow without overloading local resources. It supports a variety of media formats and offers both REST and WebSocket endpoints for batch and live processing, respectively. This makes it a strong candidate for integration with interactive applications.

Based on model performance, practicality, and ease of integration, We selected the CAGE model and the Hume facial expression model for experimental evaluation.

For discrete emotion classification in speech, several datasets are widely available, including RAVDESS (Livingstone & Russo 2018), Emo-DB, and MSP-IMPROV. However, for continuous speech emotion assessment with arousal-valence labeling in English, only a few datasets meet the criteria, such as OMG Emotion, IEMOCAP (Busso et al. 2008), and MSP-Podcast (Lotfian & Busso 2017). These datasets provide audio clips labeled with arousal and valence scores, enabling a more granular emotional analysis.

Among the available pre-trained models, we identified the wav2vec2 model (Wagner et al. 2023) on Hugging Face, which outputs arousal-valence values from speech signals and showed strong compatibility with the project requirements.

Additionally, Hume.ai's speech emotion recognition system, trained on the large-scale HUME-VB dataset, provides high-performance real-time prediction capabilities. The HUME-VB dataset contains over 280,000 vocalizations from 4,000+ participants across five culturally diverse countries, including the USA, China, India, Venezuela, and South Africa, making it a valuable addition for multilingual and cross-cultural emotion recognition scenarios.

For the experiment, we selected both the **wav2vec2** model and the **Hume vocal expression model** to assess and compare their performance in recognizing emotional states from speech input.

4.2.1 Facial Expression Experiment Setup

To evaluate the accuracy of the selected facial models (Hume and CAGE), we conducted an experiment using acted facial expressions. Participants were asked to express five emotions, **Happy, Angry, Sad, Boredom, and Calm** across three intensity levels: **Low**, **Medium**, and **High**. Each emotion was performed intentionally, and recordings were captured under similar lighting conditions using the webcam. An experiment snapshot is shown in Figure 4.3.

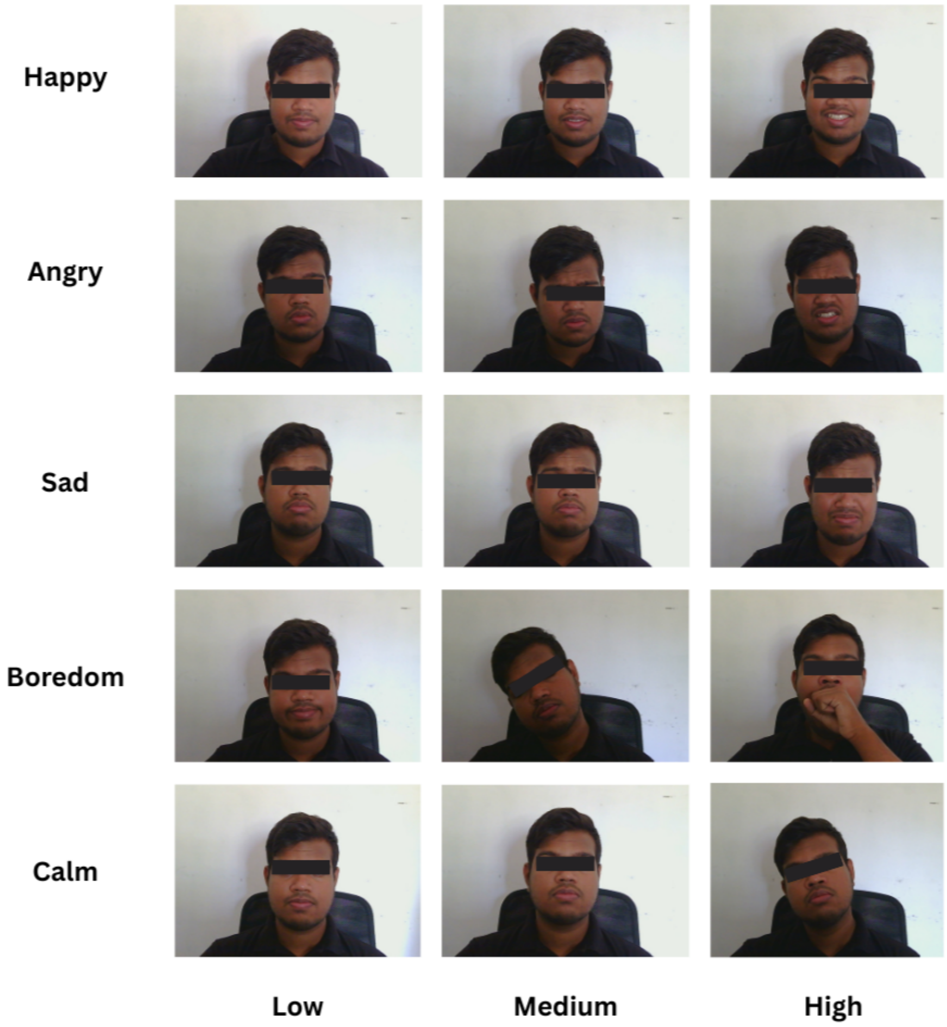


Figure 4.3: Experiment snapshot for facial expression recognition

4.2.2 Vocal Emotion Experiment Setup

For vocal emotion recognition, participants were asked to read emotion-evoking phrases that are commonly used in scientific emotion corpora such as RAVDESS and IEMOCAP. A total of 15 phrases were used, covering all five target emotions. These phrases were selected to be emotionally neutral in content so that the emotion would be expressed purely through vocal tone and prosody.

Emotion phrases used in the experiment:

- **Happy:**
  - "I'm glad you're here."
  - "That was a fantastic surprise."
  - "You made my day!"
- **Angry:**
  - "This is completely unacceptable."
  - "I've told you this before!"
  - "Why didn't you listen to me?"
- **Sad:**
  - "I miss you so much."
  - "Everything feels so heavy today."
  - "I just want to be alone."
- **Boredom:**
  - "There's nothing to do."
  - "Same thing every day."
  - "I don't care anymore."
- **Calm:**
  - "Everything is going to be okay."
  - "Let's take a deep breath."
  - "I'm feeling at peace."

4.3 Phase 2 – Multimodal fusion with personalized emotion recognition

This phase focuses on combining the selected models from Phase 1 to create a multimodal emotion recognition system. The goal is to enhance the accuracy and personalization of emotional state detection by integrating both facial and vocal data. The experimental flow for this phase is illustrated in Figure 4.4.

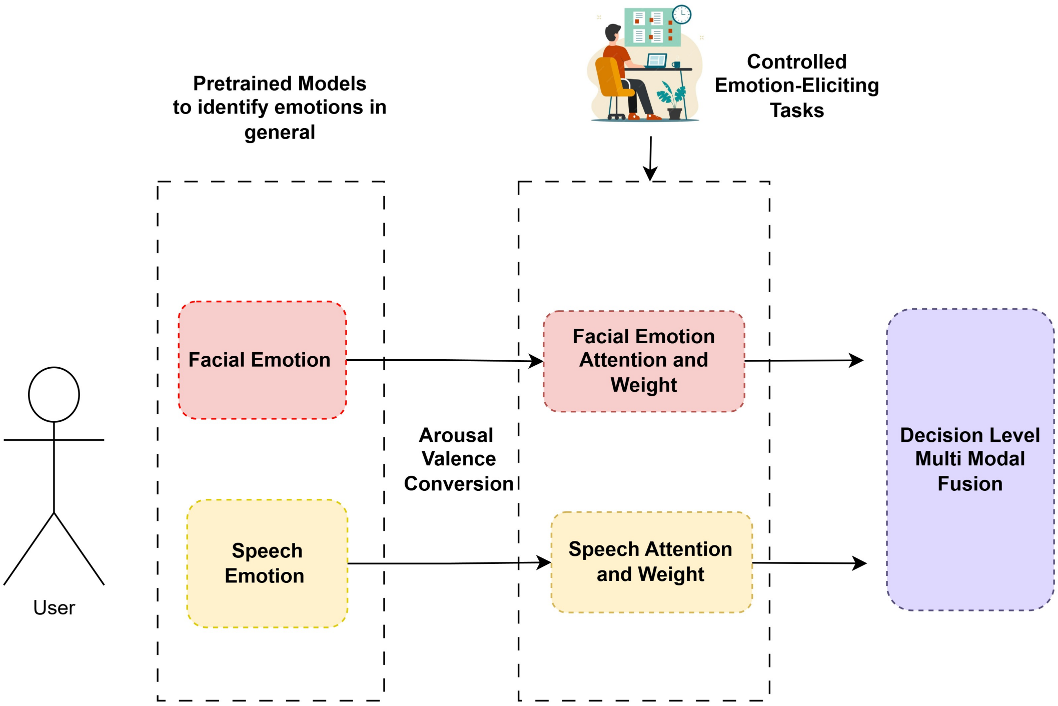

Pretrained Models to identify emotions in general
Controlled Emotion-Eliciting Tasks
Facial Emotion
Speech Emotion
User
Arousal Valence Conversion
Facial Emotion Attention and Weight
Speech Attention and Weight
Decision Level Multi Modal Fusion

Figure 4.4: Experimental flow of Phase 2

4.3.1 Emotion Eliciting Tasks

In this phase, a set of emotion-eliciting tasks were designed to capture both facial and vocal expressions for five target emotions: **Happy, Angry, Sad, Boredom, and Calm**. These emotions were specifically chosen because they cover diverse directions in the Arousal-Valence (A-V) space, which helps in achieving better separation between emotional states. This separation allows the system to identify a clearer origin point for emotional detection, as illustrated in Figure 4.5.

During the tasks, participants were asked to perform actions or respond to situa-

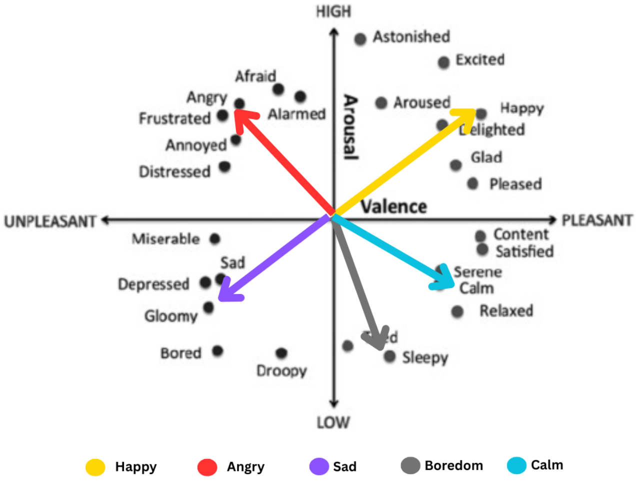


Figure 4.5: Arousal-Valence space with diverse emotion regions

tions that naturally induce each of the five emotions. While they were doing the tasks, both facial expressions and vocal responses were recorded. The emotional points collected in this phase are later used in the next step to identify the user's emotional baseline.

As discussed in Section 2.1, emotions are highly personalized, and different people express them in unique ways. Therefore, to make the system more personalized, we combined the model outputs with user feedback. After each task, participants were asked to report how they actually felt and to rate the intensity of the emotion they experienced as shown in figure 4.10. This provided quantitative data directly from the user, which was used to adjust the emotion recognition output.

In addition, after collecting vocal expressions during each task, participants were briefly asked to describe how the experience felt in their own words. These voice recordings were used as extra vocal samples corresponding to that emotion.

At the end of the entire experiment, we also collected qualitative feedback from participants. While this feedback was not used to measure emotional levels, it gave valuable insights into how the system and tasks could be improved in future iterations of the research.

4.3.1.1 Experiments for Eliciting Happy Emotion

To induce the emotion of happiness across different intensity levels, three separate tasks were designed. Each task corresponds to one of the three intensity levels: low, medium, and high. The tasks were chosen based on findings from psychological and affective computing literature that demonstrate the effectiveness of various stimuli for emotion elicitation.

Low-Intensity Happiness – Watching a Funny Video

The first task involved participants watching a short prank video clip titled "Circus Elephant Prank," sourced from YouTube (https://youtu.be/ZwJfXgTO7J4) (Just For Laughs Gags 2011). This video features a light-hearted, humorous interaction involving a hidden elephant costume, designed to surprise and entertain pedestrians.

Studies show that visual stimuli, especially comedic videos, are highly effective for inducing happiness, as they generate strong self-reported emotions and physiological responses such as smiling and increased heart rate (Siedlecka & Denson 2019). Light-hearted content like animal pranks is typically perceived as safe and amusing, making it ideal for gently elevating mood without causing overstimulation.

The emotional intensity expected from such clips is generally low. Research indicates that cute or humorous videos involving animals or children tend to trigger emotions of low motivational intensity, which means the emotional state is pleasant but not highly arousing or action-driven (Wang & Chen 2022).

Medium-Intensity Happiness – Autobiographical Recall

For medium-intensity happiness, participants were asked to recall and verbally describe a recent event that made them feel happy. This task relies on autobiographical recall, a well-established emotion elicitation technique in psychological research.

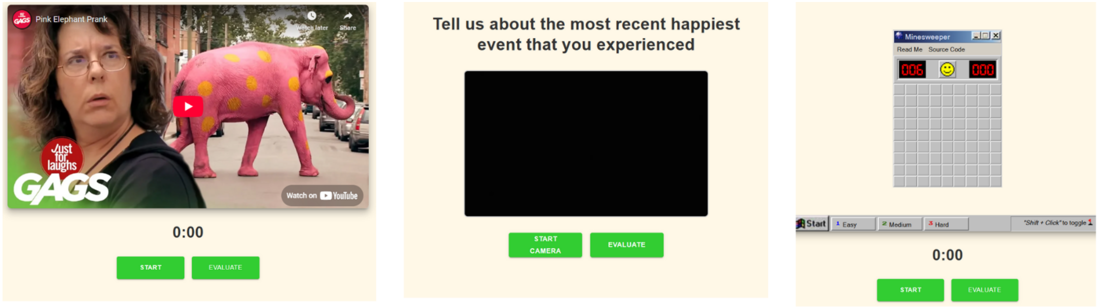
Recollecting personal happy memories has been shown to increase both self-reported happiness and physiological responses such as heart rate (Siedlecka & Denson 2019). Moreover, autobiographical recall engages brain regions linked to reward processing, reinforcing the positive emotional experience (Speer & Delgado 2017). This method is particularly meaningful because it draws from the participant's own experiences, often resulting in a more personalized and emotionally moderate response compared to passive methods like video watching.

High-Intensity Happiness – Playing a Game (Minesweeper)

The third task was designed to evoke high-intensity happiness through a controlled gaming experience. Participants played a version of the classic puzzle game Minesweeper, which was modified to have a high probability of success—though this was not disclosed to the participant. Upon winning, the system delivered verbal praise and congratulations to enhance the reward experience.

Games, especially those that involve cognitive effort and strategic thinking, have been shown to elicit strong feelings of accomplishment and positive emotion (Jones et al. 2014). Puzzle-based games like Minesweeper can be particularly rewarding when players succeed after investing effort, making them suitable for eliciting high-arousal positive emotions such as excitement and joy. By ensuring a high chance of winning, the task aimed to maximize the user's feeling of success, thereby producing a more intense emotional response.

The screenshots related to happiness eliciting tasks are shown in Figure 4.6. The first image shows the video clip, the second image shows the participant describing their happy memory, and the third image shows the Minesweeper game interface. The tasks were designed to be engaging and varied, ensuring that participants could experience happiness at different intensity levels.



Tell us about the most recent happiest event that you experienced

Happy: Test 1

Happy: Test 2

Happy: Test 3

Figure 4.6: Eliciting tasks for Happy Emotion

4.3.1.2 Experiments for Eliciting Anger Emotion

To study the emotion of anger across various intensity levels, three custom-designed gameplay tasks were developed. All tasks introduce frustration through interruptions, distractions, or interface issues—elements proven to induce negative emotional responses like anger in controlled experiments.

Low-Intensity Anger – Chrome Dino Game with Noise Distraction

The first task involves participants playing the well-known Chrome Dino game. However, to introduce a frustrating element, the game is played while the participant listens to honking vehicle sounds and other urban noise distractions via audio playback (`https://www.youtube.com/watch?v=dOk1JFAAMCo`) (soundsforyou 2021).

Research has shown that auditory noise, particularly when irrelevant to the task, can significantly interfere with attentional control and lead to feelings of frustration and mild anger (Choi et al. 2013). Even non-hazardous noise levels can trigger emotional stress, especially during cognitive tasks. In gaming scenarios, background noise is found to negatively affect concentration and cognitive performance, making it a useful tool to induce low levels of anger or irritation.

Medium-Intensity Anger – Number Game with Pop-Up Interruptions

The second task involves a number selection game where participants must click numbers in ascending order under a time constraint. As the game progresses and the participant nears completion, random pop-up ads begin appearing more frequently. In the last 10 seconds, a countdown starts flashing and loud beep sounds are introduced to further stress the player. Ultimately, due to these disruptions, the game becomes unwinnable.

Scientific literature supports the idea that unexpected interruptions such as pop-up ads can be strong triggers for anger and frustration. Such ads are perceived as highly intrusive and break the player's sense of immersion and control (Hanbazazh & Reeve 2021). The psychological theory of reactance further suggests that when individuals feel their freedom to act is restricted (e.g., by pop-ups interfering with gameplay), they experience negative emotions such as frustration and anger. These effects are even more pronounced when the interruptions occur during critical moments, which is intentionally designed in this task.

High-Intensity Anger – Flappy Bird Variant with Broken Controls

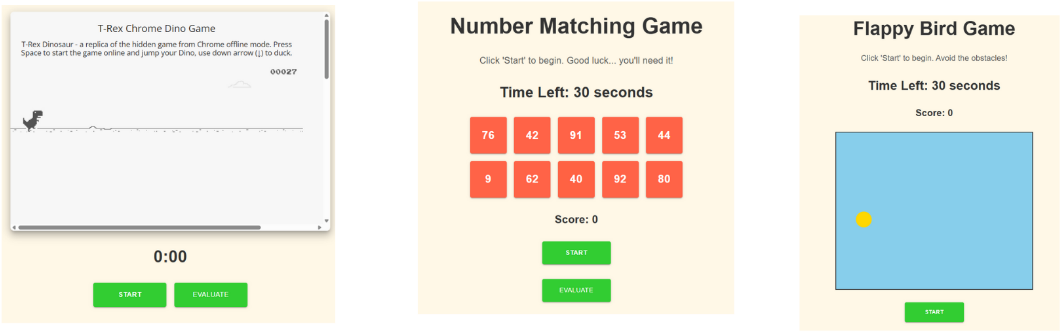
The third task is based on the classic Flappy Bird game. However, participants are given no instructions on how to play. Furthermore, the game mechanics behave unpredictably: pressing the spacebar does not always trigger a jump, and sometimes causes the window to scroll instead. As in the first task, distracting honking sounds play in the background to further irritate the participant.

This task is designed to provoke a high level of anger by breaking the user's mental model of how the game should behave. When participants feel they are failing not due to their own skill but because of poor or broken game mechanics, frustration increases dramatically. Studies have shown that unclear game objectives and malfunctioning controls significantly reduce players' sense of competence and autonomy, which are essential psychological needs (Bevilacqua et al. 2019). When those needs are blocked, it can lead to intense anger and even aggressive reactions, including what's commonly known as "rage quitting."

Screenshots related to the anger eliciting tasks are shown in Figure 4.7. The fourth image shows the Chrome Dino game with noise distractions, the fifth image shows the number game with pop-up interruptions, and the last image shows the Flappy Bird variant with broken controls. These tasks were designed to be engaging and varied, ensuring that participants could experience anger at different intensity levels.

4.3.1.3 Experiments for Eliciting Sad Emotion

To capture sad emotional responses at different intensities, three distinct tasks were designed. These tasks use music and emotionally powerful videos, which are known to effectively evoke sadness through auditory and visual pathways.



T-Rex Chrome Dino Game

T-Rex Dinosaur - a replica of the hidden game from Chrome offline mode. Press Space to start the game online and jump your Dino, use down arrow (↓) to duck.

Anger: Test 1

Number Matching Game

Click 'Start' to begin. Good luck... you'll need it!

Time Left: 30 seconds

Anger: Test 2

Flappy Bird Game

Click 'Start' to begin. Avoid the obstacles!

Time Left: 30 seconds

Anger: Test 3

Figure 4.7: Eliciting tasks for Anger Emotion

Low-Intensity Sadness – Listening to "Little Motel" by Modest Mouse

The first task involves participants listening to the song "Little Motel" by Modest Mouse https://youtu.be/zqQTODR3kR8?si=yqNkeW1bUYPNUEXb (Mouse 2007), accompanied by a short explanation highlighting the lyrical themes of heartbreak, regret, and emotional distance. Sad music has been shown to be an effective tool for inducing low-level sadness, although the emotional response may include feelings of nostalgia, aesthetic pleasure, or being emotionally moved (Ribeiro et al. 2019).

"Little Motel" is commonly described by listeners as emotionally evocative and melancholic, largely due to its subdued instrumentation and reflective lyrics. Anecdotal listener feedback, such as that found in public forums Reddit users (n.d.), reinforces this emotional interpretation. Providing context before playing the song helps guide participants' emotional attention, making the sadness more focused and easier to observe.

Medium-Intensity Sadness – Viewing a Clip of a Young Person in Distress

For medium-intensity sadness, participants were shown a video clip of a young person crying while overwhelmed by everyday struggles. Research supports the use of video stimuli for eliciting sadness, particularly when they depict realistic, relatable struggles and human vulnerability. Visual and auditory cues together create a more immersive emotional experience compared to music or images alone (Nojavanasghari et al. 2016).

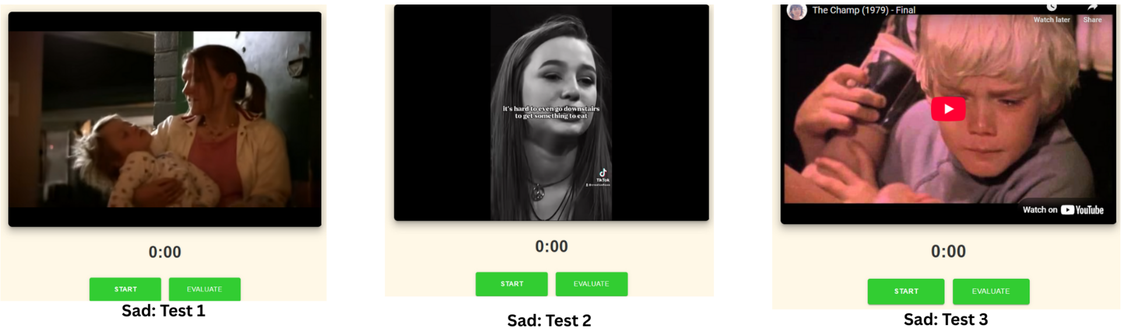
Scenes involving children or youth in distress can be especially powerful due to their perceived innocence and relatability, increasing empathy from the viewer. As a result, this type of content tends to generate a moderate level of sadness while remaining emotionally grounded.

High-Intensity Sadness – "The Champ" (1979) Death Scene

The third task used a well-known emotionally intense video: the final scene from the 1979 film "The Champ", in which a young boy reacts to the death of his father (https://youtu.be/b5qwTeCj4jc?si=2HJe3G0KL6U6azXY) (Lootens 2013). This particular clip is considered a gold standard in emotion research for eliciting high-intensity sadness. It has been cited in numerous studies, including foundational work by Gross and Levenson (1995), and appears in validated databases like E-MOVIE (Kuijsters et al. 2016, Maffei & Angrilli 2019).

The scene evokes intense grief and helplessness, making it a powerful tool for eliciting high levels of sadness. Its use in many research studies confirms its reliability and emotional impact across different participant groups.

Figure 4.8 shows the screenshots related to the sadness eliciting tasks.



Sad: Test 1

Sad: Test 2

Sad: Test 3

Figure 4.8: Eliciting tasks for Sad Emotion

4.3.1.4 Experiments for Eliciting Boredom Emotion

To explore boredom at different intensity levels, three carefully designed tasks were conducted. These tasks are grounded in psychological literature that identifies monotony, repetition, and lack of engagement as key contributors to boredom.

Low-Intensity Boredom – Repetitive Button Clicking Task

In the first task, participants were instructed to click a button at exact intervals over a duration of two minutes. This task is intentionally repetitive and minimally demanding, with no variation or challenge to stimulate cognitive engagement.

Boredom frequently arises from repetitive, purposeless actions that fail to engage mental resources (Bench & Lench 2013). This low-intensity task aligns with such conditions, offering minimal novelty and no meaningful outcome, which is expected to produce a mild but noticeable feeling of boredom due to understimulation.

Medium-Intensity Boredom – Watching a Dripping Water Video

The second task involves watching a video of water dripping from a tap for two minutes (https://youtu.be/lVrYV0odeFY) (Filmmaker 2013). The visual stimulus is static and unchanging, offering no narrative or engaging content, which has been shown to reliably induce medium levels of boredom in controlled settings.

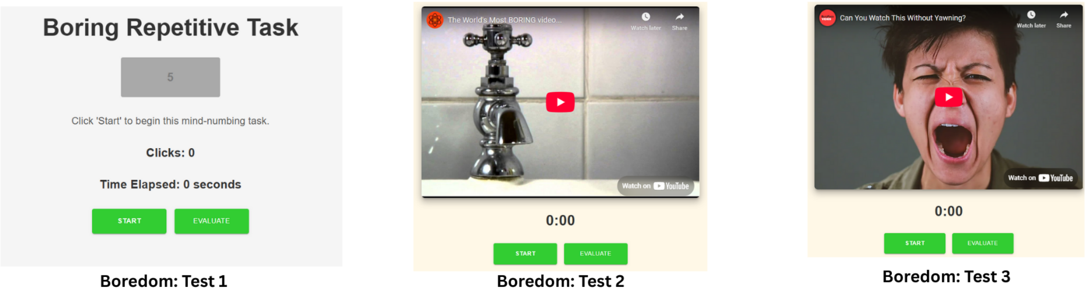
The use of unchanging visual stimuli is a validated method for boredom induction (Markey et al. 2014). The slow, repetitive nature of the dripping water presents a low-information environment that sustains passive attention without mental engagement. Compared to the motor activity of button clicking, this passive viewing requires continuous focus but offers no new stimulation, making it suitable for inducing a moderate level of boredom (Bench & Lench 2013).

High-Intensity Boredom – Watching a Yawning Video

For high-intensity boredom, participants watched a two-minute video of someone yawning repeatedly (https://youtu.be/M3QYDtSbhrA) BuzzFeedVideo (2018). Yawning is closely associated with boredom, low arousal, and disengagement. Watching someone yawn repeatedly is likely to trigger similar responses in viewers, including physiological reactions like contagious yawning and a reduction in alertness.

Research has linked the act of yawning with increased feelings of boredom, drowsiness, and mind-wandering (Norscia et al. 2020). The lack of narrative or engaging visual stimuli in the video is expected to amplify this effect. Additionally, contagious yawning may create a feedback loop of low arousal, further reinforcing the emotional state of boredom and making this task suitable for inducing high levels of boredom in participants.

Screenshots of the videos used in the experiments are shown in Figure 4.9.



Boring Repetitive Task

Click 'Start' to begin this mind-numbing task.

Clicks: 0

Time Elapsed: 0 seconds

Boredom: Test 1

Boredom: Test 2

Boredom: Test 3

Figure 4.9: Eliciting tasks for Boredem emotion

4.3.1.5 Experiments for Eliciting Calm Emotion

To elicit the emotional state of calmness across three intensities, participants were exposed to video and audio stimuli known to promote relaxation. These tasks include a combination of soothing visual elements, relaxing music, and breathing techniques that are widely used in stress reduction and mindfulness practices.

Low-Intensity Calmness – Rain and Relaxing Music Video

The first task aimed at inducing a low level of calmness involved watching a video featuring soft rain visuals accompanied by relaxing ambient music (https://youtu.be/PjUZbgZfMOo) (Professor 2024). This combination is often used in meditative settings and is known for its ability to promote mild tranquility.

Research indicates that auditory stimuli such as slow-tempo, gentle music, when paired with natural sounds like rainfall, can work synergistically to reduce physiological arousal and promote emotional relaxation. These types of stimuli are frequently used in guided meditations and background soundtracks for relaxation, reflecting their general effectiveness in inducing calmness at a subtle level.

Medium-Intensity Calmness – Music with Guided Breathing Exercise

The second task combines relaxing music with a breathing exercise guided through video (https://youtu.be/uxayUBd6T7M) (Calm 2020). This task requires the participant to actively engage in slow, deep breathing synchronized with musical rhythm, which helps regulate physiological responses and reduce stress.

Scientific evidence supports the role of deep breathing in stimulating the parasympathetic nervous system and lowering stress levels. When breathing is consciously controlled and paired with soothing music, it can lead to a more focused and deeper relaxation experience than passive listening alone. This form of active engagement is shown to promote a stronger state of calmness by influencing heart rate variability and promoting emotional self-regulation.

High-Intensity Calmness – Listening to "Weightless" by Marconi Union

For high-intensity calmness, participants listened to the song "Weightless" by Marconi Union while watching its official visual accompaniment (https://youtu.be/UfcAVejslrU) (JustMusicTV 2015). This song has been labeled as "the world's most relaxing song" and has been scientifically validated for its profound impact on stress and anxiety levels.

Studies have reported that listening to "Weightless" can reduce anxiety by up to 65% and lower physiological resting states by up to 35% (Cooper 2011). The song was developed in collaboration with sound therapists, featuring a gradually slowing tempo from 60 to 50 beats per minute, low-frequency tones, and ambient instrumentation including piano, guitar, chimes, and subtle vocals.

Comparative studies suggest that listening to this track can be more relaxing than receiving a massage, and in some cases, it has shown effects comparable to anxiety medications (Today 2020). These scientifically designed elements make it a powerful tool for eliciting deep calmness and emotional stillness.

Screenshots from the three tasks are shown in Figure 4.10. The first task features a serene rain scene, the second task includes a guided breathing exercise with calming music, and the third task showcases the ambient visuals accompanying "Weightless." Figure 4.11 presents some of the participants engaged in the experimental tasks.

Summary of all the emotion eliciting experiments is shown in Table 4.1, which is reproduced below for convenience.

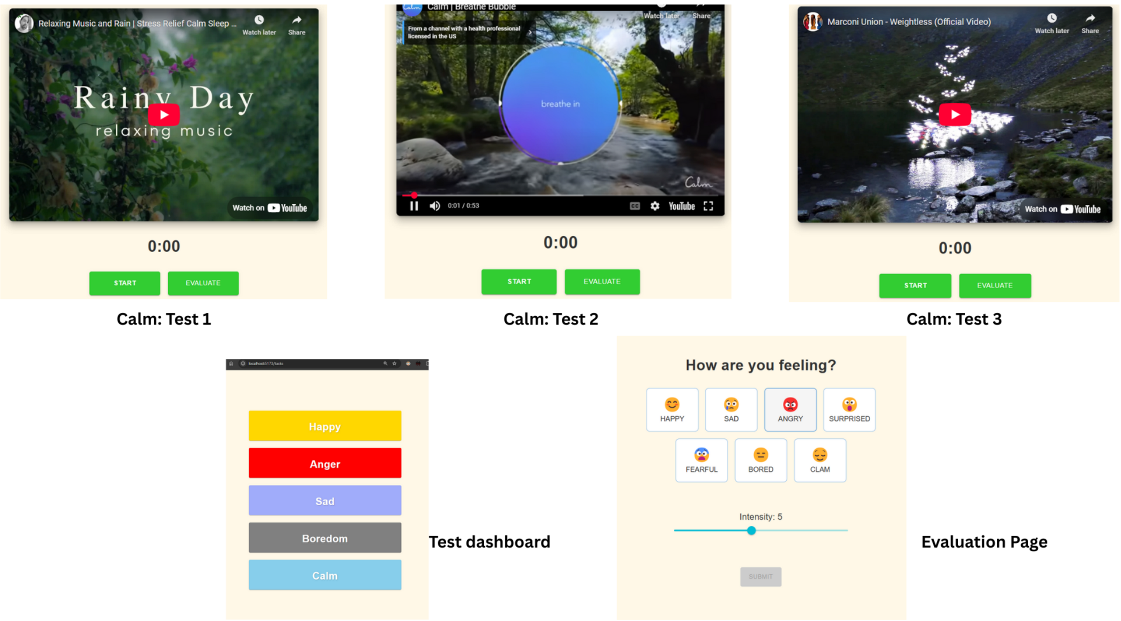


Figure 4.10: Screenshots from the three tasks used to elicit calmness and the evaluation page

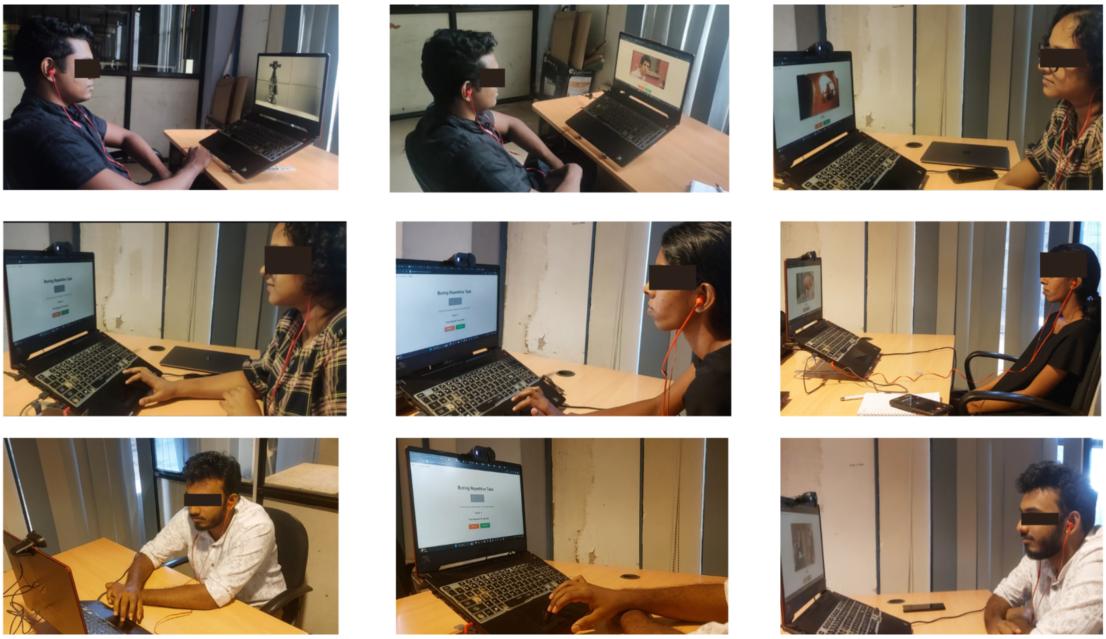


Figure 4.11: Snapshots of participants engaged in the emotion elicitation tasks.

Table 4.1: Summary of Emotion Eliciting Experiments

Emotion	**Intensity**	**Stimulus**	**Description**
Happiness	Low	Funny prank video (https://youtu.be/ZwJfXgTO7J4)	Comedy video to induce mild amusement and smiling through light-hearted content
	Medium	Autobiographical recall	Participants describe recent happy personal memory to invoke moderate happiness
	High	Minesweeper game (high chance of winning)	Designed to trigger joy through accomplishment and reward feedback
Anger	Low	Chrome Dino game + city noise (https://youtu.be/d0k1JFAAMCo)	Simple game with distracting honking sounds to create mild frustration
	Medium	Number-clicking game + pop-ups + countdown	Ads and timers disrupt gameplay, increasing frustration and loss of control
	High	Flappy Bird variant with broken controls	Unclear mechanics + noise designed to frustrate and break user expectation
Sadness	Low	”Little Motel” song (https://youtu.be/zqQTODR3kR8)	Sad music with contextual explanation to invoke reflective sadness
	Medium	Clip of young person in distress	Emotionally relatable video showing personal struggles and crying
	High	”The Champ” (1979) death scene (https://youtu.be/b5qwTeCj4jc)	Highly validated emotional clip used to induce intense grief
Boredom	Low	Button clicking task	Repetitive motor task with no variation or challenge
	Medium	Dripping water video (https://youtu.be/lVrYV0odeFY)	Monotonous, slow-paced video with low informational content
	High	Yawning video (https://youtu.be/M3QYDtSbhrA)	Triggers low arousal and disengagement, possibly contagious yawning
Calmness	Low	Rain + relaxing music (https://youtu.be/PjUZbgZfM0o)	Gentle visuals and soft ambient music for mild relaxation
	Medium	Relaxing music + breathing (https://youtu.be/uxayUBd6T7M)	Guided breathing exercise synchronized with music to enhance calm
	High	”Weightless”(https://youtu.be/UfcAVejslrU)	Scientifically validated song with strong calming effect

4.3.2 Emotion Mapping to A-V Plan

After collecting the emotional data, we need to represent these emotions in the arousal-valence (A-V) space. To do this, we convert the categorical emotion labels (like happy, sad, angry) into numerical coordinates on the A-V plane.

Arousal and valence are two important dimensions in emotion research. Valence shows how pleasant or unpleasant a feeling is, while arousal shows how calm or excited the feeling is. The **Circumplex Model of Affect**, proposed by Russell (1980), is widely used to represent this concept. It places emotions in a circular 2D space, where each emotion is mapped according to its valence and arousal values. For example, happiness is usually high in both valence and arousal, while sadness is low in both.

However, the original model does not give exact numbers for each emotion. So, for this research, we use the numerical coordinates for emotions proposed by Paltoglou & Thelwall (2012). Their work gives us empirically validated A-V values for a set of common emotions. These values are especially suitable for computational models and were derived from large-scale emotion analysis in text data, which makes them practical and tested.

We selected these values for a few key reasons:

- **Empirical Validation:** The values are based on real data and experiments, not just theory.
- **Computational Use:** The values are already tested in affective computing systems.
- **Standardization:** Using known values makes our system easier to compare with other studies.
- **Dimensional Mapping:** They fit well with our aim to measure emotions on a scale, not just labels.

Figure 4.12 shows the general circular A-V model proposed by Russell, and Figure 4.13 presents the exact coordinates used in this study, based on the work of Paltoglou and Thelwall.

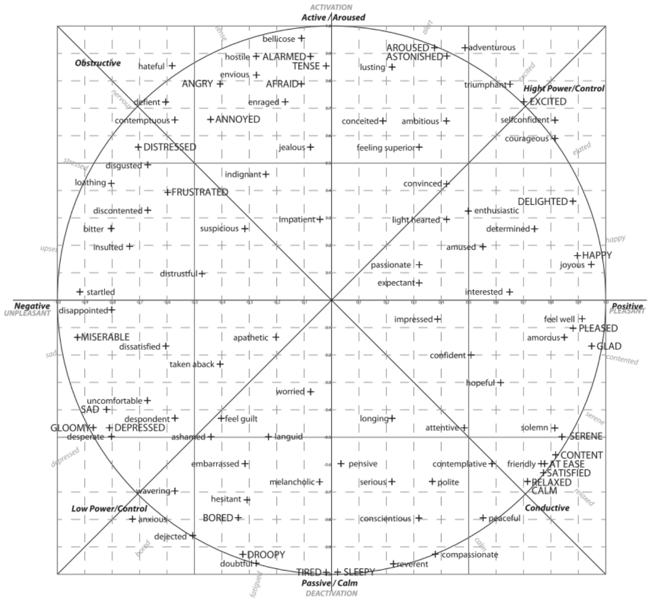


Figure 4.12: Circumplex Model of Affect Russell (1980). This circular model shows how emotions are distributed in a two-dimensional space using arousal and valence.

mood	*valence*	*arousal*	*#posts*	*mood*	*valence*	*arousal*	*#posts*
sleepy	0.01	−1.00	11,549	*bored*	−0.35	−0.78	12,784
tired	−0.01	−1.00	20,308	*annoyed*	−0.44	0.76	8,247
afraid	−0.12	0.79	20	*enraged*	−0.18	0.83	1,016
angry	−0.40	0.79	3,152	*excited*	0.70	0.71	11,074
calm	0.78	−0.68	10,040	*melancholy*	−0.05	−0.65	2,274
relaxed	0.71	−0.65	3,545	*satisfied*	0.77	−0.63	2,629
content	0.81	−0.55	11,177	*distressed*	−0.71	0.55	2,233
depressed	−0.81	−0.48	6,383	*uncomfortable*	−0.68	−0.37	1,763
discontent	−0.68	−0.32	2,490	*worried*	−0.07	−0.32	3,252
determined	0.73	0.26	3,360	*amused*	0.55	0.19	24,231
happy	0.89	0.17	16,518	*apathetic*	−0.20	−0.12	2,732
anxious	−0.72	−0.80	7,039	*peaceful*	0.55	−0.80	2,513
good	0.90	−0.08	5,062	*contemplative*	0.58	−0.60	10,718
pensive	0.03	−0.60	2,322	*embarrassed*	−0.31	−0.60	1,092
impressed	0.39	−0.06	1,610	*sad*	−0.81	−0.40	6,119
frustrated	−0.60	0.40	4,356	*hopeful*	0.61	−0.30	5,312
disappointed	−0.80	−0.03	3,328	*pleased*	0.89	−0.10	3,517

Figure 4.13: Emotion coordinates adapted from Paltoglou & Thelwall (2012). These values are used in this study to map categorical emotions to A-V values.

- **Happy**: Valence = 0.89, Arousal = 0.17
- **Angry**: Valence = -0.40, Arousal = 0.79
- **Sad**: Valence = -0.81, Arousal = -0.40
- **Boredom**: Valence = -0.35, Arousal = -0.78
- **Calm**: Valence = 0.78, Arousal = -0.68

To show different emotional intensities, we applied a **linear scaling method**. Here, each emotion starts at the neutral point $(0, 0)$, and intensity levels from 1 to 10 move the emotion point closer to its full coordinate. This follows the idea that emotional intensity changes smoothly in a straight line from neutral to peak intensity, which is also supported in literature Posner et al. (2005).

To calculate the valence and arousal coordinates at any intensity level i, where $i \in \{1, 2, ..., 10\}$, the following formula is used:

$$\text{Valence}_i = \frac{\text{Valence}_{\text{max}} \times i}{10}, \quad \text{Arousal}_i = \frac{\text{Arousal}_{\text{max}} \times i}{10} \tag{4.1}$$

Where:

- $\text{Valence}_{\text{max}}$ is the valence value at maximum intensity (level 10)
- $\text{Arousal}_{\text{max}}$ is the arousal value at maximum intensity (level 10)
- i is the current intensity level

This formula creates a straight path from the neutral point $(0, 0)$ to the maximum intensity point for each emotion. For example:

- At intensity level 1, the coordinates are 10% of the maximum
- At level 5, they are 50%
- At level 10, the full intensity is reached

Table 4.2 presents the standardized valence and arousal coordinates for five testing emotions across different intensity levels scale from 1 to 10.

Table 4.2: Emotion Intensity Coordinates Based on Paltoglou & Thelwall (2012)

Int.	**Happy**		**Angry**		**Sad**		**Boredom**		**Calm**	
	Val	Aro	Val	Aro	Val	Aro	Val	Aro	Val	Aro
1	0.089	0.017	-0.040	0.079	-0.081	-0.040	-0.035	-0.078	0.078	-0.068
2	0.178	0.034	-0.080	0.158	-0.162	-0.080	-0.070	-0.156	0.156	-0.136
3	0.267	0.051	-0.120	0.237	-0.243	-0.120	-0.105	-0.234	0.234	-0.204
4	0.356	0.068	-0.160	0.316	-0.324	-0.160	-0.140	-0.312	0.312	-0.272
5	0.445	0.085	-0.200	0.395	-0.405	-0.200	-0.175	-0.390	0.390	-0.340
6	0.534	0.102	-0.240	0.474	-0.486	-0.240	-0.210	-0.468	0.468	-0.408
7	0.623	0.119	-0.280	0.553	-0.567	-0.280	-0.245	-0.546	0.546	-0.476
8	0.712	0.136	-0.320	0.632	-0.648	-0.320	-0.280	-0.624	0.624	-0.544
9	0.801	0.153	-0.360	0.711	-0.729	-0.360	-0.315	-0.702	0.702	-0.612
10	0.890	0.170	-0.400	0.790	-0.810	-0.400	-0.350	-0.780	0.780	-0.680

4.3.3 Personalized Multimodal Fusion

In the multimodal fusion stage, we selected decision-level fusion over feature-level and hybrid fusion. This choice allows each modality to output its results independently, which we can then combine without altering the internal architecture of pre-trained models. Decision-level fusion is particularly effective here, as it maintains modularity while still leveraging each modality's unique strengths.

Decision-Level Fusion Techniques: Support Vector Regression (SVR), Adaptive Weighted Fusion Using Reinforcement Learning, Hierarchical Fusion Using Neural Networks, Dynamic Multi-Head Attention Mechanism, Fuzzy Logic Fusion, Weighted Functions.

Due to limited user-specific data, we selected a non-machine-learning-based approach, specifically the **MSE-Based Fusion Method**. Although we initially considered fuzzy logic rules—which dynamically adjust weights,like prioritizing facial modality if facial arousal is high and vocal valence is low, we faced challenges in rule definition due to data scarcity. Instead, our framework employs Mean Squared Error (MSE) calculations to derive emotion-specific weights for each modality.

A unique **Fusion Matrix** is generated for each user using data collected during the emotion elicitation task. This matrix includes weights for five emotion, where different modalities may be more reliable for different emotional states:

Listing 4.1: Emotion-specific fusion weights per user

```
fusion_weights = {
    "Happy":   {"facial": W_facial, "vocal": W_vocal},
    "Sad":     {"facial": W_facial, "vocal": W_vocal},
    "Angry":   {"facial": W_facial, "vocal": W_vocal},
    "Boredom": {"facial": W_facial, "vocal": W_vocal},
    "Calm":    {"facial": W_facial, "vocal": W_vocal}
}
```

The fusion strategy is grounded in statistical theory: lower error indicates higher reliability. Thus, weights are assigned **inversely proportional to the modality's**

MSE when compared to self-reported values.

Step 1: Error Calculation

$$MSE_{facial}(e) = \frac{(A_{facial} - A_{self})^2 + (V_{facial} - V_{self})^2}{2}$$

$$MSE_{vocal}(e) = \frac{(A_{vocal} - A_{self})^2 + (V_{vocal} - V_{self})^2}{2}$$

Step 2: Weight Generation

$$W_{facial}(e) = \frac{1/MSE_{facial}(e)}{1/MSE_{facial}(e) + 1/MSE_{vocal}(e)}, \quad W_{vocal}(e) = \frac{1/MSE_{vocal}(e)}{1/MSE_{facial}(e) + 1/MSE_{vocal}(e)}$$

Step 3: Multimodal Fusion Output

$$A_{fused}(e) = W_{facial}(e) \times A_{facial} + W_{vocal}(e) \times A_{vocal}$$

$$V_{fused}(e) = W_{facial}(e) \times V_{facial} + W_{vocal}(e) \times V_{vocal}$$

This fusion method allows the system to dynamically adapt to each individual's emotion expression style, yielding improved recognition accuracy over single-modality and non-personalized fusion approaches.

The results from the personalized fusion approach and individual modality approaches were evaluated and compared to understand whether the fusion process provides a significant improvement in emotion recognition accuracy and reliability. This comparison helps to justify the effectiveness of using adaptive, user-specific weighting in our framework.

4.4 Phase 3 - Initial Baseline Identification

In this section, we present the methods and techniques used to identify the initial emotional baseline for each participant. We explore two primary methods: **Kernel Density Estimation (KDE)** and a simpler **Histogram-Based method**. Both methods aim to identify the region of highest density in the A-V space, which corresponds to the participant's emotional baseline. The emotional states often follow a emotion escalation cycle, beginning from a neutral or baseline state, escalating to a peak, and then returning back to baseline. This implies that the highest density region in the emotional data distribution corresponds to the user's emotional baseline.

4.4.1 Initial Baseline Identification Using Kernel Density Estimation (KDE)

After thorough research and experiments, we identified **Kernel Density Estimation (KDE)** with a Gaussian kernel as a suitable method to identify the baseline emotional region in the arousal-valence (A-V) space.

Gaussian KDE Equation and Implementation

The 2D Gaussian KDE function is defined as:

$$\hat{f}_h(x, y) = \frac{1}{n \cdot h^2 \cdot 2\pi} \sum_{i=1}^{n} \exp\left(-\frac{1}{2}\left[\frac{(x - x_i)^2 + (y - y_i)^2}{h^2}\right]\right) \tag{4.2}$$

where:

- n is the number of data points,
- (x_i, y_i) are the arousal-valence coordinates of each data point,
- h is the bandwidth parameter controlling the smoothness of the density estimate.

The KDE was implemented in Python using `scipy.stats.gaussian_kde` as shown below:

Listing 4.2: Gaussian KDE implementation

```
from scipy import stats
kernel = stats.gaussian_kde([arousal, valence])
f = np.reshape(kernel(positions), xx.shape)
```

The bandwidth h is automatically selected using **Scott's Rule**, a well-known heuristic for KDE bandwidth estimation. For $d = 2$ dimensions (arousal and valence), it is computed as:

$$h = n^{-\frac{1}{d+4}} = n^{-\frac{1}{6}} \tag{4.3}$$

Baseline Region Identification and Implementation

Once the density function $\hat{f}_h(x, y)$ is estimated, we identify the square region with the highest total density. We define a 0.1×0.1 square sliding over the KDE grid. For each position (x_0, y_0), the integral over the region is computed as:

$$D(x_0, y_0) = \int_{x_0}^{x_0+0.1} \int_{y_0}^{y_0+0.1} \hat{f}_h(x, y)\, dy\, dx \tag{4.4}$$

The center of the baseline region is identified by finding the square with the maximum integrated density:

$$(x_b, y_b) = \underset{(x_0, y_0)}{\arg\max}\, D(x_0, y_0) \tag{4.5}$$

This integral is approximated numerically using grid sums. The implementation is shown below:

Listing 4.3: Baseline region identification with KDE

```
for i in range(len(x_grid) - grid_points_in_square):
    for j in range(len(y_grid) - grid_points_in_square):
        square_density = np.sum(
            f[j:j+grid_points_in_square,
              i:i+grid_points_in_square
            ]
        )
        if square_density > max_density:
            max_density = square_density
            center_x = x_grid[i] + square_size / 2
            center_y = y_grid[j] + square_size / 2
            baseline_center = (center_x, center_y)
```

Parameter Values Used in Analysis

- **Grid Size** (g): 0.01

  A smaller grid provides higher precision but requires more computation time.

- **Square Size** (s): 0.1

  Defines the size of the candidate region for emotional baseline. Assumes baseline remains stable in a small area of the A-V space.

- **Bandwidth** (h): Automatically calculated using Scott's Rule: $h \approx n^{-1/6}$

This KDE-based approach provides a robust and data-driven way to estimate a participant's emotional baseline, which can then be used as a reference point for detecting deviations in real-time emotion tracking.

Baseline Representation and Evaluation

For KDE method, the baseline region is represented as the center point of the region with highest density. Each baseline center was then compared with emotional values recorded during emotion eliciting tasks, where participants experienced *Happy, Angry, Sad, Boredom*, and *Calm* states.

These emotional states were recorded alongside their intensity and converted into arousal-valence values using the method discussed in Subsection 4.3.2. This allowed a direct comparison between the estimated baseline and observed emotional states, forming the foundation for evaluating the accuracy and reliability of both KDE and histogram-based methods.

4.5 Phase 4 - Evaluating Responses from LLM with Emotional State

In this stage of the research, participants interacted with a Large Language Model (LLM) to evaluate how emotional context can influence the personalization and quality of generated responses as shown in Figure 4.14. Two types of responses were collected for each participant query:

- **Control Response:** Generated using the original raw query.

- **Emotion-Aware Response:** Generated by combining the raw query with the participant's baseline emotional state using prompt engineering techniques.

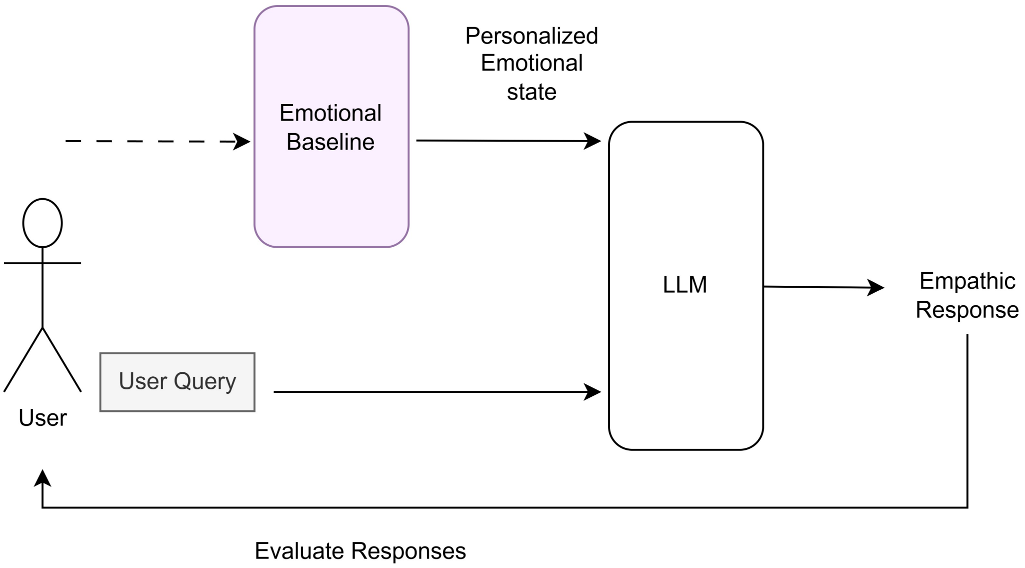


Figure 4.14: Experimental flow of Phase 2

Use of LLMs in Emotionally Intelligent Response Generation

To maintain consistency and reliability in emotional response analysis, we selected GPT-4o for the experimental phase due to its proven capabilities in emotional understanding and language generation. According to the study by Wang et al. (2023) Wang et al. (2023), GPT-4 demonstrated:

- **Superior Emotional Intelligence:** Achieved an EQ score of 117, outperforming 89% of human participants.
- **Human-Like Emotional Pattern Recognition:** Showed a pattern similarity of $r = 0.28$ (aligning with 67% of humans), indicating it processes emotions in a manner similar to human cognition.
- **High Consistency and Scale:** The newer GPT-4.1 offers improved consistency, which is essential for valid response comparisons.

These characteristics make GPT-4o-mini an excellent choice for our task, where nuanced understanding of user emotions is required to personalize the generated text. Using a model with benchmarked emotional intelligence also enhances the research's credibility and reproducibility.

Comparison with Alternative LLMs

For reference, Table 4.3 summarizes other potential LLMs and their performance based on EQ benchmarks from the same study:

Model	**EQ Score**	**Pattern Similarity**	**Recommendation**
GPT-4o	117 (89%)	0.28 (67%)	**Excellent – Selected Model**
Claude	106 (61%)	0.11 (28%)	Good alternative
GPT-3.5-turbo	103 (52%)	0.04 (17%)	Acceptable but limited
Vicuna	105 (59%)	-0.02 (10%)	Not recommended
Alpaca	104 (56%)	0.03 (15%)	Not recommended
ChatGLM	94 (28%)	0.09 (24%)	Not recommended
Koala	83 (13%)	0.43 (93%)	Only for analysis comparison

Table 4.3: Emotional Intelligence Benchmarking of LLMs (adapted from Wang et al. (2023))

While models like DeepSeek and EmoLLM were considered due to their open-source and emotion-specific capabilities, their relatively small parameter sizes and lack of academic benchmarking in emotional intelligence made them unsuitable for this study's goals.

Prompt Engineering with Emotional Context and Chain-of-Thought

In order to personalize the LLM's response, we incorporated the user's emotional state directly into the prompt. This approach not only injects context but also uses a **chain-of-thought (CoT) mechanism** to encourage the LLM to reason about the user's emotion before generating the answer.

The emotional data structure used for this task is shown below:

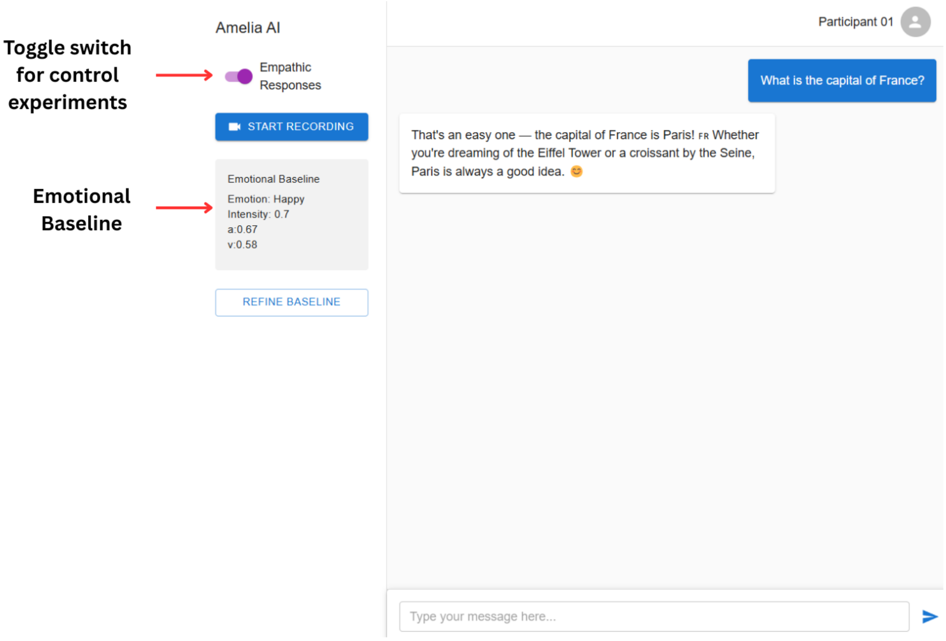
Listing 4.4: Example of emotional data sent to LLM

```
# Define the emotional data
emotional_data = {
    "emotion": "happy",
    "intensity": 7,
    "arousal": 0.67,
    "valence": 0.58
}
```

The following is how the emotional data is embedded into the prompt using a multi-step instruction structure to guide the LLM's reasoning process:

```
# Define the user's query
user_query = "What is the capital of France?"

# Create the prompt with emotional data and Chain-of-Thought
prompt = f"""
SYSTEM: You are an AI assistant designed to provide helpful
responses while considering the user's emotional state.
Before responding, analyze the provided emotional baseline
data to inform your response approach.

EMOTIONAL BASELINE:
- Emotion: {emotional_data.get('emotion', 'neutral')}
- Intensity: {emotional_data.get('intensity', 5)}/10
- Arousal: {emotional_data.get('arousal', 0)} (scale -1 to 1)
- Valence: {emotional_data.get('valence', 0)} (scale -1 to 1)

INSTRUCTIONS:
1. First, analyze the user's emotional state based on
the data provided
2. Consider how this emotional state might influence
their needs or expectations
3. Craft a response that addresses both the content of
their query and is appropriate for their emotional context

USER QUERY: {user_query}
"""
```

This method ensures that the LLM reasons step-by-step before answering, making the output more aligned with the user's emotional state and potentially improving satisfaction and trust during interaction. Screenshot of the Chat interface with the prompt and emotional baseline data is shown in Figure 4.15.



Emotional Baseline
Emotion: Happy
Intensity: 0.7
a:0.67
v:0.58

What is the capital of France?

That's an easy one — the capital of France is Paris! FR Whether you're dreaming of the Eiffel Tower or a croissant by the Seine, Paris is always a good idea. 😊

Figure 4.15: Example of LLM prompt with emotional baseline data

Proposed Questions and Rationale

To evaluate the influence of emotional context on LLM responses, a set of seven user questions was developed across a variety of categories. These questions were selected to reflect common day-to-day topics where emotional tone could realistically influence the response. The categories included explanatory, advice-seeking, recommendation, general knowledge, practical, historical explanation, and skill-building advice.

Ealuation Questions asking for LLM responses were as follows:

- **Explanatory:** Can you explain what blockchain technology is and how it works?
- **Advice-seeking:** What are some effective ways to improve my memory?

- **Recommendation:** Recommend a book for someone who enjoys mystery novels.
- **General Knowledge:** How does the stock market work?
- **Practical:** What should I consider when buying a new laptop?
- **Historical Explanation:** Tell me about the history of the Olympic Games.
- **Skill-building Advice:** How can I start learning to play the guitar?

These questions were chosen to allow for emotional influence on tone, content, and empathy. For example:

- For the memory improvement question, a happy emotional state might result in more engaging or playful suggestions, while a sad state may yield serious, science-backed advice.
- For book recommendations, positive emotions may lead to light-hearted choices, while negative emotions could suggest deeper, more reflective books.
- In technical explanations like blockchain, the tone may shift based on emotional state, more enthusiastic with high arousal, or more formal and concise with low arousal.

Factual questions such as "What is the capital of France?" were intentionally excluded since emotional data is unlikely to affect the outcome. This design decision aligns with findings from *EmoBench* Sabour et al. (2024), which highlight that LLMs show greater emotional adaptation in open-ended or subjective queries.

This stage aims to determine whether incorporating emotional state into the LLM prompt results in improved user satisfaction, emotional alignment, and response relevance.

Evaluation Considerations

The goal of this phase is to determine whether the inclusion of emotional context improves the perceived relevance, tone, and empathy of responses. To do this, each query is submitted both with and without emotional context. Responses are then compared using the following criteria:

- **Tone and Language:** Does the emotionally-aware response include more enthusiastic, calming, or supportive language when appropriate?
- **Content Adjustment:** Are suggestions or recommendations tailored to the user's emotional state (e.g., energetic vs. relaxing activities)?
- **Empathy and Relevance:** Does the response reflect awareness of the user's needs or mood in how it addresses the question?

These metrics are used for both qualitative review and participant feedback, helping assess how effectively the LLM incorporates emotional awareness into real-world queries.

4.6 Phase 5 - Refining the Initial Baseline using Reinforcement Learning

As explained earlier in Section 2.1, the emotion baseline of a user is not something fixed. It can change over time depending on the person's mood, situation, and other conditions. Because of that, our system needs a way to update this baseline from time to time. But we also want to do this with minimal inputs from the user.

To solve this problem, we are using a Reinforcement Learning (RL) framework. More specifically, we use Q-learning to refine the emotional baseline in the arousal-valence plane. This method helps the system to learn how to adjust the baseline gradually, based on the feedback it gets and how well it performs. It also works with very little direct input from the user.

Sometimes, the user gives direct feedback using a simple emoji-based system. These emojis help the system understand how the user is feeling in a lightweight and non-intrusive way. The overall idea of this stage is shown in Figure 4.16.

Prior to examining the structure and functioning of the RL model, the operation of the emoji feedback mechanism is first considered.

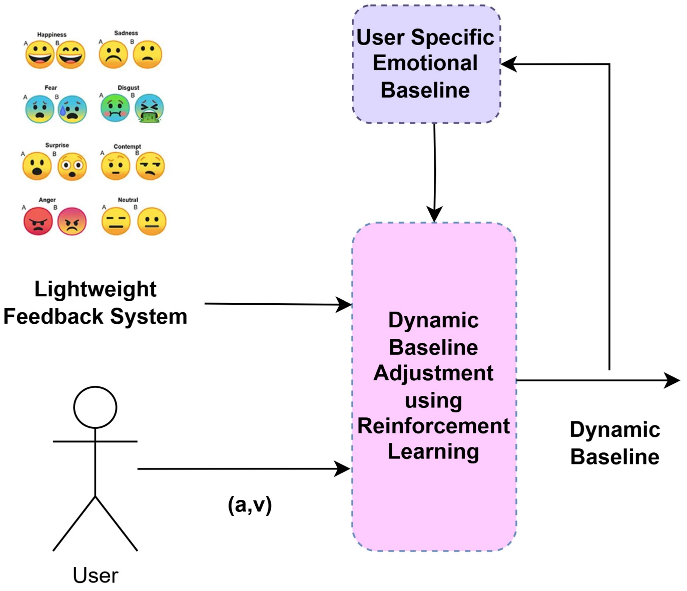

Happiness
Sadness
Fear
Disgust
Surprise
Contempt
Anger
Neutral
Lightweight Feedback System
User
(a,v)
User Specific Emotional Baseline
Dynamic Baseline Adjustment using Reinforcement Learning
Dynamic Baseline

Figure 4.16: Experimental flow of Phase 5

4.6.1 Relationship with Emojis and Arousal-Valence Values

Several studies have tried to map emojis into the arousal-valence space. One important example is a study published in 2022 titled *Classification of 74 facial emoji's emotional states on the valence-arousal axes* Kutsuzawa et al. (2022). This study involved 1,082 participants who rated 74 facial emojis using a nine-point scale for both valence and arousal.

The researchers used cluster analysis and one-way ANOVA to group the emojis into six main clusters. Each cluster represents a certain emotional state, going from very negative to very positive. The clusters also show how strong or intense each emotion is by using arousal values.

The table below shows the main results from the study:

This table helps us understand how different emojis can be used to reflect emotional states in both valence (positive or negative) and arousal (intensity) dimensions.

Using this idea, we built our emoji feedback mechanism. It allows users to give lightweight feedback about how they feel. This feedback is then used in the reinforcement learning stage. The overall design of this emoji feedback interface is

Cluster Label	**N**	**Valence (SD)**	**Arousal (SD)**
Strong positive sentiment	12	7.42 (0.40)	7.19 (0.34)
Moderately positive sentiment	9	6.57 (0.62)	5.98 (0.29)
Neutral with positive bias	12	5.49 (0.38)	5.19 (0.29)
Neutral with negative bias	19	4.27 (0.38)	4.83 (0.27)
Moderately negative sentiment	10	3.59 (0.37)	5.84 (0.30)
Strong negative sentiment	12	2.74 (0.40)	6.91 (0.37)

Table 4.4: Clusters of emojis and their valence-arousal values from Kutsuzawa et al. (2022).

shown in Figure 4.17.

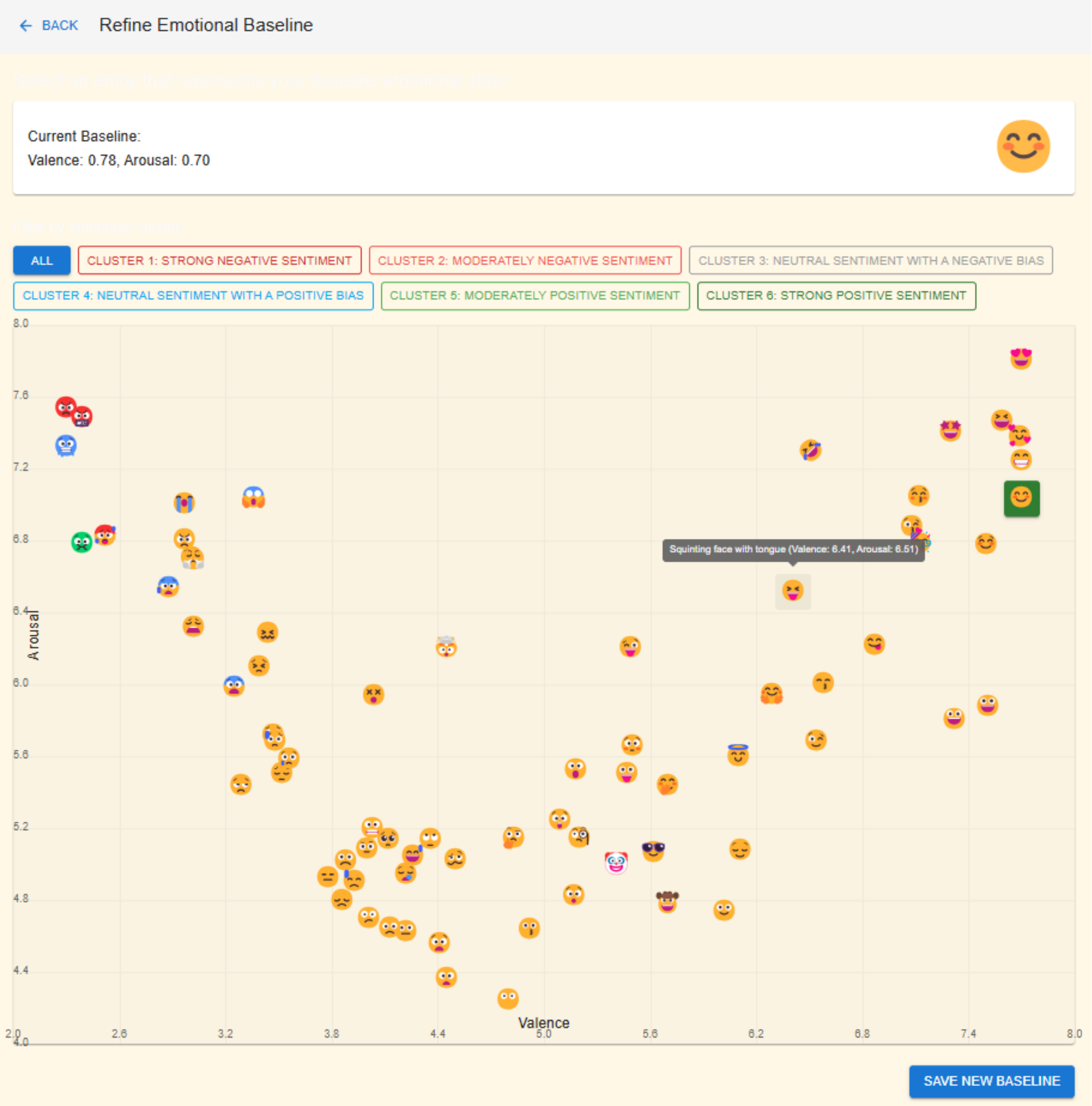
4.6.2 Emotion Refining Using Reinforcement Learning

Initially, RL model was trained using data gathered during the Phase 2 baseline identification stage. Data points close to the inferred baseline and those confirmed by user feedback were used as representative values for the baseline emotional state.

To ensure the reliability of the baseline over time, an additional data collection session was conducted one week later. During this session, each participant engaged in a 10-minute interaction involving tasks designed to elicit minimal emotional response. These tasks were chosen specifically to observe the participant's natural, non-elicited expressions. Those tasks are:

- Describing their daily routine
- Counting from 1 to 20 accending and decending out loud
- Listing the items they see in a random image
- Request explanations of photosynthesis from LLM and reading the response
- Request explanations of how ballpoint pens work from LLM and reading the response.

after end of each task, the participants were asked to rate their emotional state using the emoji feedback system.



Refine Emotional Baseline

Select an emoji that represents your baseline emotional state:

Current Baseline:
Valence: 0.78, Arousal: 0.70

Filter by emotional cluster:

Squinting face with tongue (Valence: 6.41, Arousal: 6.51)

← BACK
ALL
CLUSTER 1: STRONG NEGATIVE SENTIMENT
CLUSTER 2: MODERATELY NEGATIVE SENTIMENT
CLUSTER 3: NEUTRAL SENTIMENT WITH A NEGATIVE BIAS
CLUSTER 4: NEUTRAL SENTIMENT WITH A POSITIVE BIAS
CLUSTER 5: MODERATELY POSITIVE SENTIMENT
CLUSTER 6: STRONG POSITIVE SENTIMENT
Arousal
8.0
7.6
7.2
6.8
6.4
6.0
5.6
5.2
4.8
4.4
4.0
Valence
2.0
2.6
3.2
3.8
4.4
5.0
5.6
6.2
6.8
7.4
8.0
SAVE NEW BASELINE

Figure 4.17: Emoji feedback interface used for lightweight user input.

Why Q-Learning?

Q-learning is a model-free RL algorithm that learns which actions give the best long-term rewards. It is a good fit for this task because of the following reasons:

- The baseline may move slowly over time. Q-learning can learn from experience and adjust its values to keep up with these changes.
- Direct user input is not given all the time. With eligibility traces, Q-learning can still learn from delayed rewards.
- The epsilon-greedy method in Q-learning helps the model to explore new emotional states, while also using what it has already learned.

Inputs to the Model

The model uses three main types of data:

- **Emotional Data:** Arousal and valence values collected when the user is calm (not emotionally elevated). These are close to the real baseline.
- **Direct Feedback:** Occasionally, the user gives feedback using a lightweight emoji system. These are assumed to be accurate.
- **Initial Baseline:** We start with a known estimate of the user's baseline as the prior knowledge.

State and Action Spaces

State Space:

The arousal-valence space is divided into a 10x10 grid, resulting in 100 possible states. Each state is written as $s = (i, j)$, where:

- $i \in \{0, 1, \ldots, 9\}$ for arousal
- $j \in \{0, 1, \ldots, 9\}$ for valence

Each cell center is calculated as:

$$\text{Arousal} = -1 + (i + 0.5) \cdot 0.2, \quad \text{Valence} = -1 + (j + 0.5) \cdot 0.2$$

The initial baseline, provided as a continuous point (a_0, v_0), is mapped to the nearest grid cell using:

$$i = \max(0, \min(9, \lfloor (a_0 + 1)/0.2 \rfloor))$$

$$j = \max(0, \min(9, \lfloor (v_0 + 1)/0.2 \rfloor))$$

Action Space:

The agent can take one of five actions to adjust the baseline position:

- Move Left: Decrease valence by one grid cell ($j \leftarrow \max(0, j - 1)$)

- Move Right: Increase valence by one grid cell ($j \leftarrow \min(9, j+1)$)
- Move Up: Increase arousal by one grid cell ($i \leftarrow \min(9, i+1)$)
- Move Down: Decrease arousal by one grid cell ($i \leftarrow \max(0, i-1)$)
- Stay: Keep the current position unchanged

Reward Function

The reward function incorporates both emotional data and direct feedback, with direct feedback given higher priority due to its accuracy. At each time step $t+1$, after transitioning to state s_{t+1}, the reward r_{t+1} is computed based on the available data D_{t+1}:

- **Direct Feedback:** If direct feedback $b_{t+1} = (a_b, v_b)$ is available (e.g., every N steps):

$$r_{t+1} = -\text{distance}(\text{center}(s_{t+1}), b_{t+1})$$

- **Emotional Data:** If emotional data $e_{t+1} = (a_e, v_e)$ is available:

$$r_{t+1} = -\text{distance}(\text{center}(s_{t+1}), e_{t+1})$$

- **No Data:** If neither is available:

$$r_{t+1} = 0$$

Where distance is calculated using Euclidean distance:

$$\text{distance}(\text{center}(s), D) = \sqrt{(\text{center}(s)_a - a_d)^2 + (\text{center}(s)_v - v_d)^2}$$

with $\text{center}(s) = (-1 + (i+0.5) \cdot 0.2, -1 + (j+0.5) \cdot 0.2)$ for state $s = (i, j)$, and $D = (a_d, v_d)$.

Learning Algorithm

We use Q-learning with eligibility traces to handle sparse rewards effectively. The algorithm proceeds as follows:

1. Initialization:

- Q-table $Q(s, a) = 0$ for all states s and actions a
- Eligibility traces $e(s, a) = 0$ for all states s and actions a
- Initial state s_0
- Parameters: $\alpha = 0.1$ (learning rate), $\gamma = 0.9$ (discount factor), $\lambda = 0.9$ (eligibility trace decay), $\epsilon = 0.3$ (initial exploration rate)

2. At each time step t:

1. Observe current state s_t
2. Choose action a_t using epsilon-greedy policy:
    - With probability ϵ, select a random action
    - Otherwise, select $a_t = \arg\max_a Q(s_t, a)$
3. Take action a_t, transition to s_{t+1}
4. Receive data D_{t+1}:
    - If $(t + 1) \mod N = 0$, $D_{t+1} = b_{t+1}$ (direct feedback)
    - Else if emotional data is available, $D_{t+1} = e_{t+1}$
    - Else, $D_{t+1} = \text{None}$
5. Compute reward r_{t+1} as defined in the reward function
6. Compute temporal difference error:

$$\delta_t = r_{t+1} + \gamma \cdot \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)$$

7. Update eligibility traces:

$$e(s_t, a_t) \leftarrow e(s_t, a_t) + 1$$

8. Update Q-values for all states and actions:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \cdot \delta_t \cdot e(s, a) \quad \text{for all } s, a$$

9. Decay eligibility traces:

$$e(s, a) \leftarrow \gamma \cdot \lambda \cdot e(s, a) \quad \text{for all } s, a$$

10. Update state: $s_t \leftarrow s_{t+1}$
11. Decay exploration rate: $\epsilon \leftarrow \max(\epsilon_{\min}, \epsilon \cdot \epsilon_{\text{decay}})$

Exploration vs Exploitation:

To balance exploring new baseline positions with leveraging learned knowledge:

- Initial exploration rate: $\epsilon = 0.3$
- Minimum exploration rate: $\epsilon_{\min} = 0.05$
- Decay rate: $\epsilon_{\text{decay}} = 0.999$

(Code listings for the RL algorithm are provided in Appendix A.)

5

Results and Evaluation

This chapter presents the evaluation and results for the experiments that were discussed in Section 4.1. It also addresses the research objectives mentioned in Section 3.4.2. Chapter begins with Section 5.1, which explains the process of selecting participants for the study. The following sections, from Section 5.2 to Section 5.6, present detailed evaluations of each phase of the methodology. These include selecting suitable emotion recognition models, implementing the personalized multimodal fusion approach, identifying the emotional baseline, integrating emotional context into LLM responses, and refining the baseline using reinforcement learning. Each phase is evaluated using both qualitative and quantitative feedback from participants to assess how well the system addresses the defined research objectives.

5.1 Participant Selection

For this research, we selected 10 participants using a quota sampling method combined with convenience sampling. The participants were handpicked, but not based on expert knowledge. The aim was to make sure there was a balanced representation in terms of gender, background, and age.

Out of the 10 participants, there were 6 males and 4 females. From a background point of view, 6 were from Computer Science Background while 4 were from non-technical backgrounds. Age-wise, 2 participants were from the 15–20 age group, 5 were in the 24–26 range, and 2 were aged between 50–60.

This method helped to analyze differences in emotional expression and recognition based on these demographic factors. All participants took part with informed consent. The consent form used for this purpose is available in the Appendix (see Appendix A).

5.2 Phase 1: Selecting Suitable Emotion Recognition Models

In Phase 1, we mainly focused on evaluating how well different models performed when detecting emotional states using facial and vocal data. The evaluation was based on both the accuracy of emotion classification and how close the predicted intensity was to the acted intensity levels.

5.2.1 Facial Expression Experiment

This subsection discusses the results from the facial expression experiment. The emotion categorization performance of the models can be seen in Figure 5.1, and the intensity identification results are shown in Figure 5.2. In this experiment, two models were used: the HUME image expression model and CAGE.

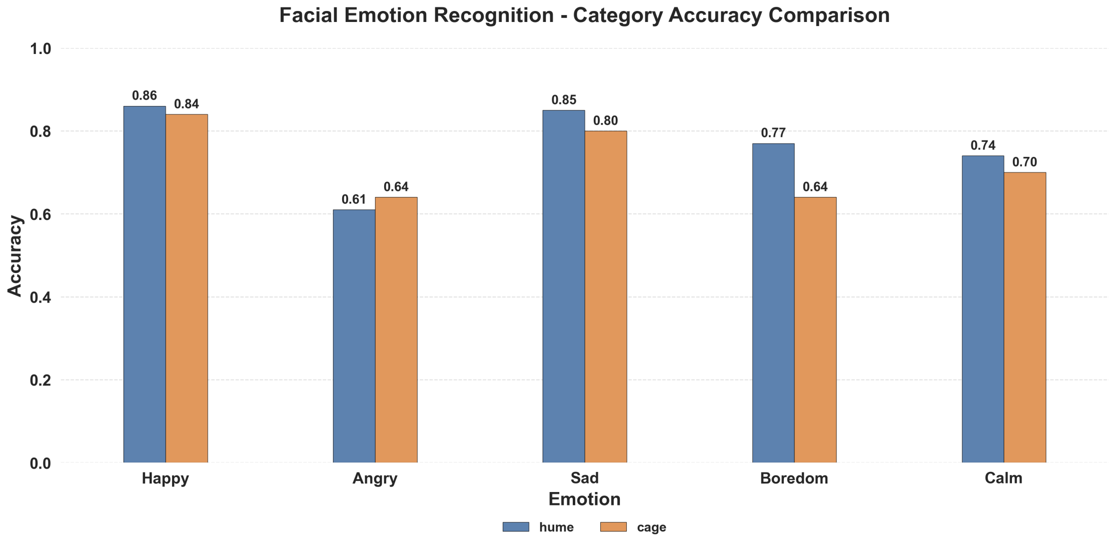


Figure 5.1: Emotion categorization results for facial expression experiment

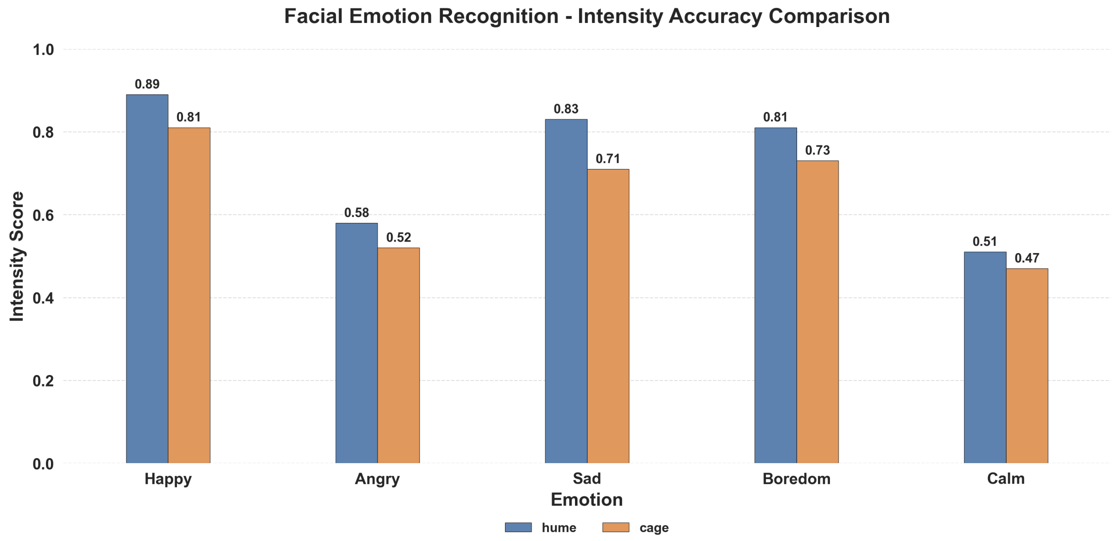


Figure 5.2: Emotion intensity identification results for facial expression experiment

Observation

This section presents a structured comparison of the performance of the CAGE and HUME models in emotion category recognition and intensity measurement, based on experimental results.

Emotion Category Recognition

Both models demonstrated high accuracy in detecting certain emotions, with varying performance across categories:

- **Happy**: HUME achieved 86% accuracy, slightly outperforming CAGE at 84%.
- **Sad**: HUME recorded 85% accuracy, compared to CAGE's 80%, showing a modest advantage.
- **Angry**: Both models struggled with anger recognition. CAGE performed slightly better at 64%, compared to HUME's 61%. This difficulty may stem from participants struggling to express anger naturally during the experiment.
- **Boredom**: HUME outperformed CAGE, with 77% accuracy compared to CAGE's 64%.
- **Calm**: HUME achieved 74% accuracy, while CAGE recorded 70%, indicating

a slight edge for HUME.

Intensity Measurement

HUME generally outperformed CAGE in measuring emotion intensity across all emotions:

- **Happy**: HUME achieved 89% accuracy in intensity prediction, compared to CAGE's 81%.
- **Sad**: HUME recorded 83% accuracy, significantly outperforming CAGE's 71%.
- **Angry**: HUME achieved 58% accuracy, compared to CAGE's 52%, showing a modest improvement.
- **Boredom**: HUME outperformed CAGE, with 81% accuracy compared to CAGE's 73%.
- **Calm**: HUME recorded 51% accuracy, compared to CAGE's 47%. Both models struggled, often confusing high calmness with low boredom.

Both models performed well in detecting happiness and sadness, with HUME showing a slight edge in category recognition for most emotions. Anger recognition remained challenging for both, likely due to unnatural expressions by participants. HUME consistently outperformed CAGE in intensity measurement across all emotions, with particularly strong performance for happiness and sadness. The confusion between high calmness and low boredom suggests potential limitations in distinguishing subtle emotional states.

5.2.2 Vocal Emotion Recognition Models Analysis

This subsection presents the analysis of vocal emotion recognition experiments. The comparison was done between the HUME audio expression model and Wave2Vec2 model for both emotion category recognition and intensity identification. The results are shown in Figure 5.3 and Figure 5.4.

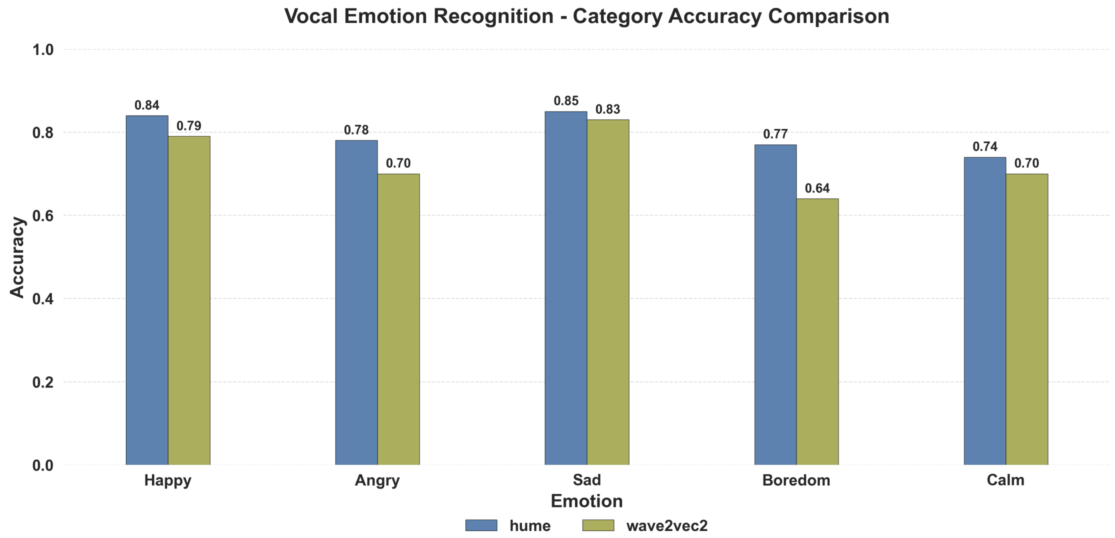


Figure 5.3: Vocal emotion categorization results: HUME vs Wave2Vec2

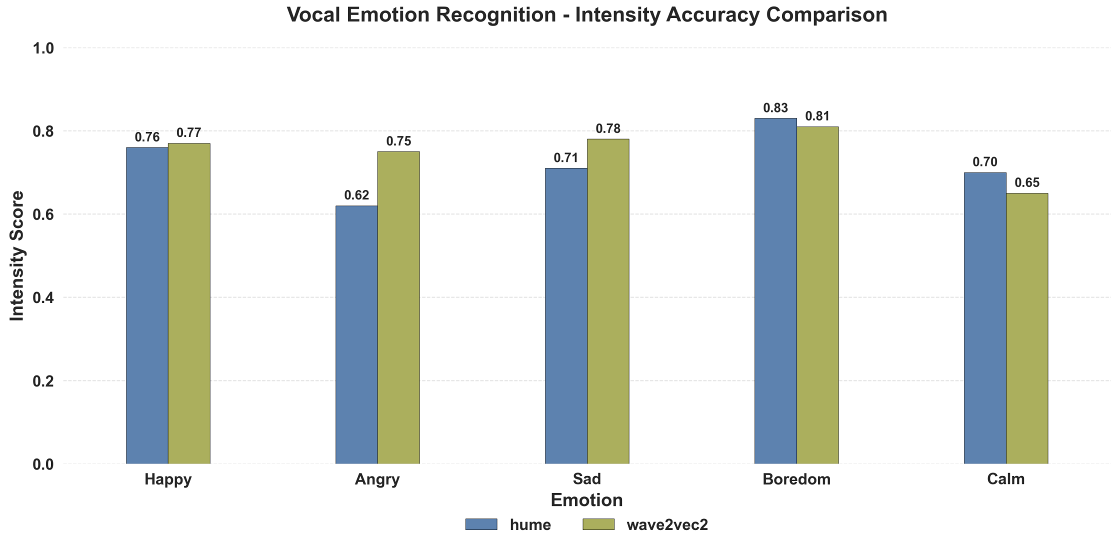


Figure 5.4: Vocal emotion intensity identification: HUME vs Wave2Vec2

Observation

This section presents a structured comparison of the performance of the HUME and Wave2Vec2 models in emotion category recognition and intensity measurement, based on experimental results for vocal data.

Emotion Category Recognition

Both models demonstrated varying performance across emotion categories:

- **Happy**: HUME achieved 84% accuracy, outperforming Wave2Vec2 at 79%.

- **Sad**: Both models recognized sadness well, with HUME scoring 85% and Wave2Vec2 83%, showing a slight advantage for HUME.
- **Angry**: HUME performed better in category identification with 78% accuracy, compared to Wave2Vec2's 70%.
- **Boredom**: HUME had stronger category recognition at 77%, compared to Wave2Vec2's 64%.
- **Calm**: Both models had similar results, with HUME scoring 74% and Wave2Vec2 70%.

Intensity Measurement

Performance in measuring emotion intensity varied, with each model showing strengths for specific emotions:

- **Happy**: Both models performed almost equally, with Wave2Vec2 achieving 77% accuracy and HUME 76%.
- **Sad**: Wave2Vec2 performed slightly better, with 78% accuracy compared to HUME's 71%.
- **Angry**: Wave2Vec2 showed a clear advantage, achieving 75% accuracy, compared to HUME's 62%.
- **Boredom**: Wave2Vec2 gave slightly better results, with 81% accuracy compared to HUME's 83%.
- **Calm**: HUME was slightly better, with 70% accuracy compared to Wave2Vec2's 65%.

Both models performed well in recognizing sadness, with HUME showing a slight edge in category recognition for most emotions, particularly happiness, anger, and boredom. Wave2Vec2 demonstrated strengths in intensity measurement, notably for anger and sadness, and performed comparably to HUME for happiness and boredom. The similar performance in calm category recognition suggests robustness in detecting subtler emotions, though intensity measurement differences indicate HUME's slight advantage for calmness. These results highlight complementary strengths, with HUME excelling in category recognition and Wave2Vec2 in specific intensity measurements.

Based on these results, and since we are planning to perform a multimodal analysis in the next phase, we chose to continue with the HUME model to maintain consistency between the facial and vocal emotion recognition results.

5.3 Phase 2: Personalized Multimodal Fusion

In this phase, we analyse the data collected from the experiments mentioned in Section 4.3, and apply decision-level fusion using weights that are calculated based on the MSE.

5.3.1 Personalized Weights

At the end of each task, we have facial and vocal predictions collected from the participant recordings. However, to properly evaluate how accurate those predictions are, we need a ground truth. Since emotion is a personal experience, we considered the self-reported emotional data provided by each participant as the ground truth. Participants reported which emotion they felt and how intense that emotion was. This information is used to calculate the personalized reliability of each modality for each person. The self-reported data is shown in the tables in Appendix A.

Then, Performed data cleaning by removing emotion labels that were outside the scope of our research. The remaining emotion labels were converted into arousal-valence values based on the mapping defined in Section 4.3.2. For comparison with self-reported values, we extracted the highest recorded emotion within each category from the system outputs. Participants were instructed to provide self-reported feedback by indicating the strongest emotion they felt during each task.

Subsequently, using the MSE-based weighting method described in Section 4.3.3, modality-specific errors for each participant were calculated. These were then used to generate personalized fusion weights, as shown in Listing 4.1. The resulting modality weights and corresponding MSE values are presented in Appendix A.

Analysis

A visual representation of the data presented in Appendix A is shown in Figure 5.7, and the average fusion weights per emotion are illustrated in Figure 5.5. Additionally, the distribution of the fusion weights across participants is displayed using a box plot in Figure 5.6.

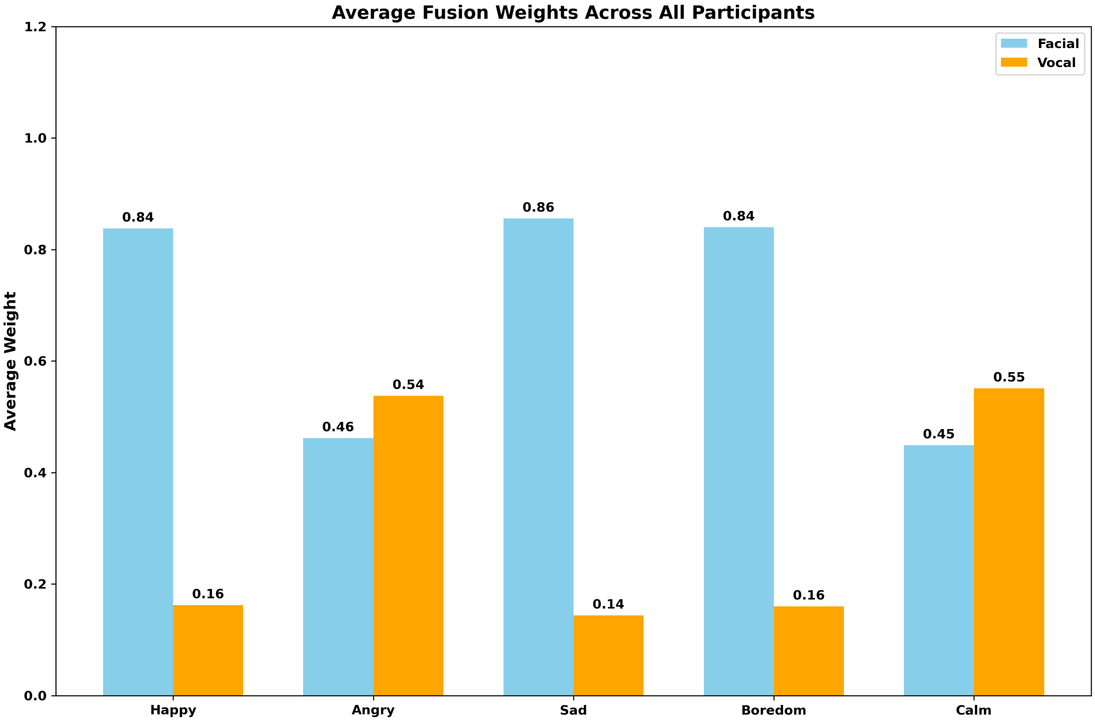


Figure 5.5: Average fusion weights per emotion category

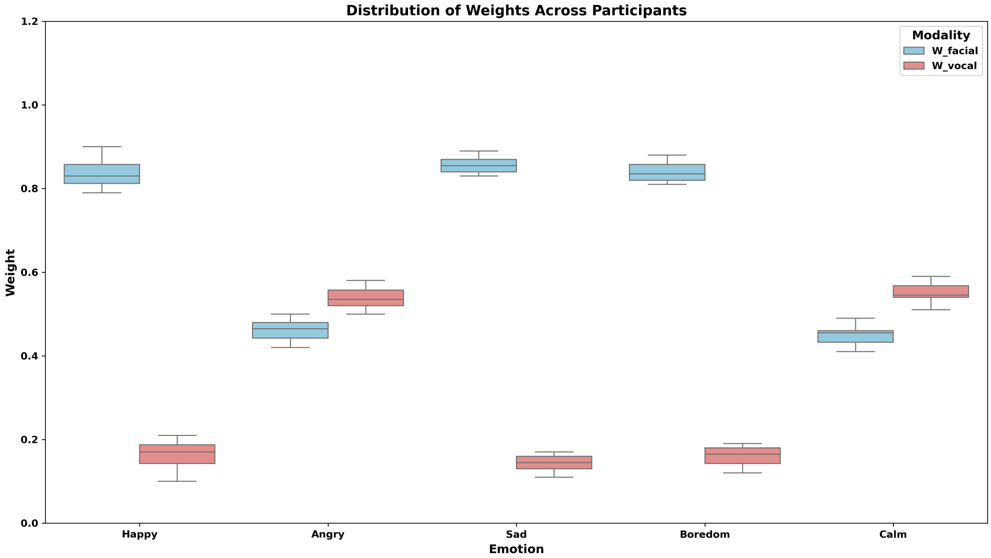


Figure 5.6: Weight distribution boxplot across all participants

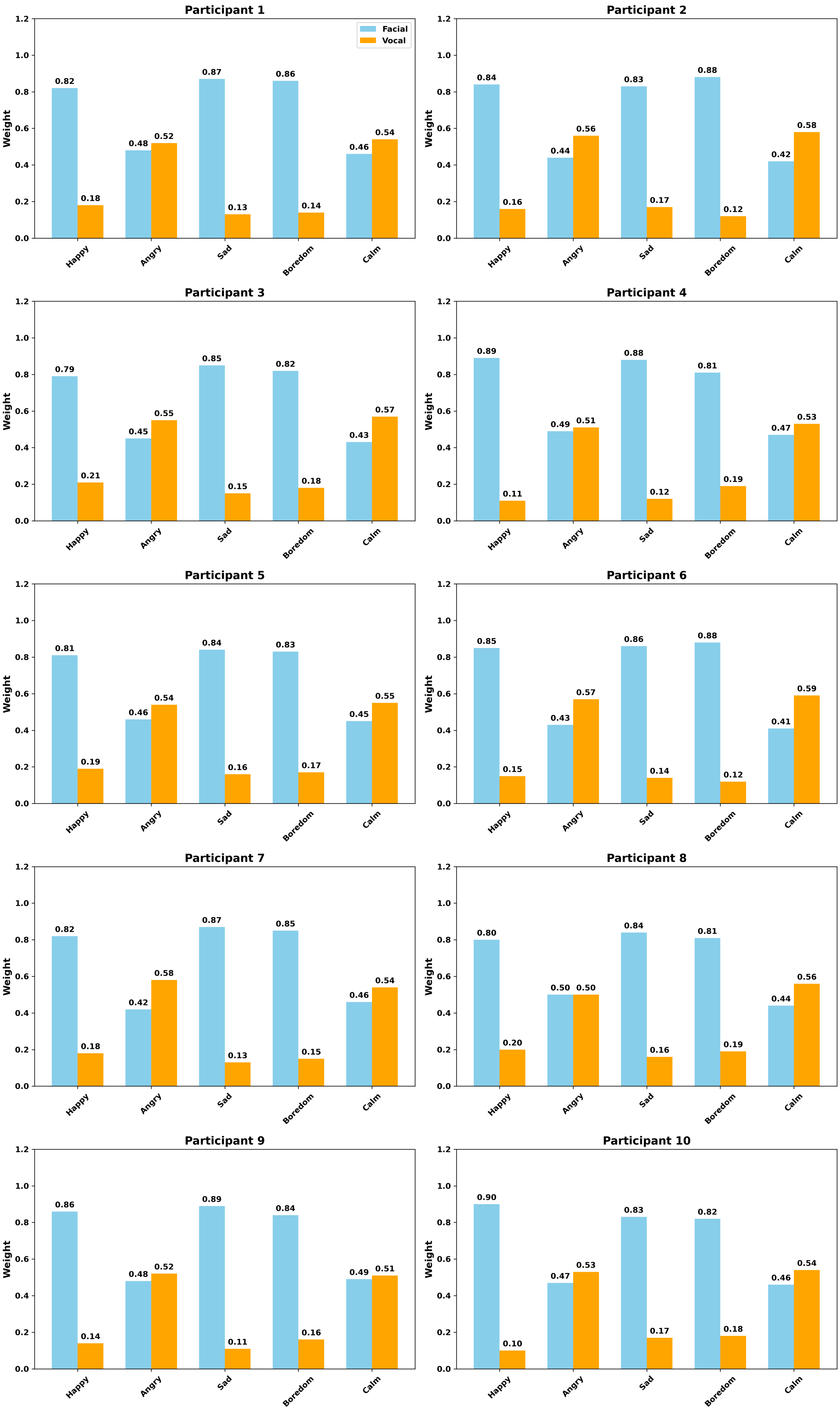


Figure 5.7: Fusion weight visualization for each participant

The average fusion weights calculated for each emotion category are summarized below:

Emotion	$\mathbf{W_{facial}}$	$\mathbf{W_{vocal}}$
Angry	0.46	0.54
Boredom	0.84	0.16
Calm	0.45	0.55
Happy	0.84	0.16
Sad	0.86	0.14

From these weights, the dominant modality for each emotion was determined as follows:

Emotion	**Dominant Modality**
Angry	Vocal
Boredom	Facial
Calm	Vocal
Happy	Facial
Sad	Facial

Key Insights from Fusion Weight Analysis

The analysis of fusion weights yields several interesting observations. Firstly, certain emotions such as Happy, Sad, and Boredom are predominantly recognized through facial expressions, with facial weights averaging over 0.84. On the other hand, Angry and Calm emotions appear to depend more on vocal cues, although the weights remain relatively balanced, indicating contributions from both modalities.

In terms of distribution, Happy and Sad emotions show the strongest modality preference, where the system clearly leans toward facial data. Conversely, the weights for Angry and Calm are more evenly distributed, suggesting that both modalities are essential for recognizing these emotions.

Furthermore, individual variations across participants show higher consistency in fusion weights for emotions like Happy, Sad, and Boredom. This implies that these emotions might have more universal expression patterns. However, for Angry and

Calm, there is greater variation, pointing to the possibility that the expression of these emotions can differ significantly between individuals.

5.3.2 Multimodal Fusion Analysis

In this phase, we performed multimodal fusion by combining both facial and vocal emotional predictions assigned to each participant. This was done using a weighted fusion formula for both emotion category (A) and intensity (V), where weights were assigned per emotion based on the reliability of each modality:

$$A_{fused}(e) = W_{facial}(e) \times A_{facial} + W_{vocal}(e) \times A_{vocal}$$

$$V_{fused}(e) = W_{facial}(e) \times V_{facial} + W_{vocal}(e) \times V_{vocal}$$

Using this method, we created a new dataset with fused emotional values. These fused values were then analyzed and compared with the ground truth using different statistical methods.

Evaluation Methods

To measure how well the fusion worked, several statistical techniques were applied.

First, we used Euclidean Distance to calculate how close the predicted emotion coordinates were to the ground truth. The formula used is:

$$\text{euclidean\_distance}(x_1, y_1, x_2, y_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Next, we applied a paired t-test to check if there was any significant improvement in the fused results compared to using facial or vocal data alone. The formula used was:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

where $\bar{d}$ is the mean of the differences between paired observations, s_d is the standard deviation of those differences, and n is the number of samples.

To measure improvement, we calculated the percentage gain of the fused method over the individual modalities. The improvement formula was:

$$\text{improvement} = \left(\frac{\text{base\_distance} - \text{fused\_distance}}{\text{base\_distance}}\right) \times 100$$

In the implementation, it was done using:

```
facial_vs_fused = ((metrics_df[’facial_distance’] - metrics_df[’fused_distance’]
                   metrics_df[’facial_distance’]) * 100


vocal_vs_fused = ((metrics_df[’vocal_distance’] - metrics_df[’fused_distance’])
                  metrics_df[’vocal_distance’]) * 100
```

Emotion recognition accuracy was also used as a key metric. It was calculated by dividing the number of correctly predicted emotions by the total number of samples:

$$\text{accuracy} = \frac{\text{number\_of\_matches}}{\text{total\_number\_of\_samples}}$$

Standard statistical functions were also used for summarizing results, such as mean and standard deviation:

$$\bar{x} = \frac{\sum x_i}{n}, \quad s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

These methods helped us understand how much the fusion improved the emotion recognition task and how reliable the predictions were compared to the original single-modality models.

Analysis

The fused method consistently shows the lowest average Euclidean distance across all emotions, which means it aligns better with the ground truth compared to using facial or vocal data alone. As shown in Figure 5.8, this improvement is seen across almost all emotion categories.

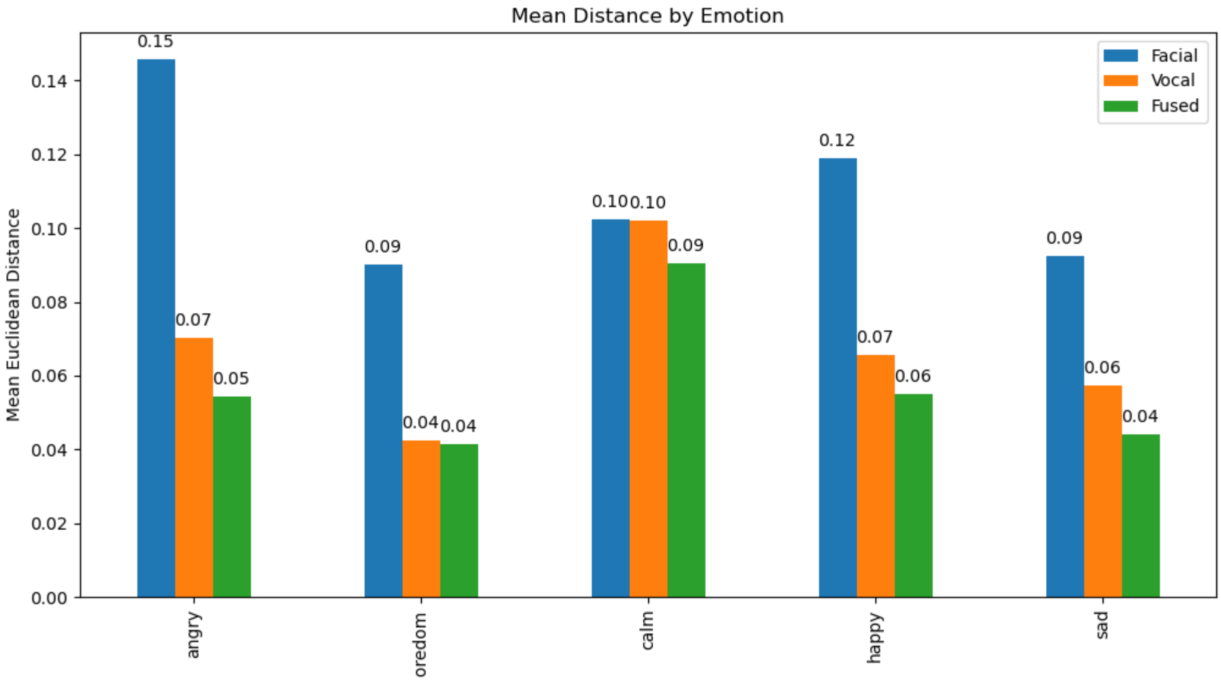


Figure 5.8: Performance comparison by Euclidean distance for each emotion

Table 5.1 presents the mean Euclidean distances and their standard deviations for each method:

Table 5.1: Mean Euclidean Distances for Each Method

Method	**Mean Distance ± Std. Dev.**
Facial	0.1112 ± 0.0773
Vocal	0.0681 ± 0.0472
Fused	0.0577 ± 0.0481

The fused method's mean distance of 0.0577 is lower than that of the vocal method (0.0681) and substantially lower than the facial method (0.1112). A lower Euclidean distance signifies predictions that are closer to the ground truth, demonstrating that the fused approach enhances accuracy by leveraging the strengths of both modalities.

To assess the statistical significance of these improvements, paired t-tests were conducted. The results are shown in Table 5.2:

Table 5.2: T-test Results for Fused Method Comparisons

Comparison	**T-Statistic**	**P-Value**	**Significant**
Facial vs Fused	9.5129	0.0000	Yes
Vocal vs Fused	3.6117	0.0004	Yes

Both p-values are below the 0.05 threshold, confirming that the fused method's improvements over the facial and vocal methods are statistically significant. The larger t-statistic for the facial vs. fused comparison (9.5129) compared to the vocal vs. fused comparison (3.6117) suggests a more pronounced enhancement over the facial method.

Figure 5.9 shows a visual representation of the fused method's performance improvements over the facial and vocal methods across emotions and intensity levels. While not explored in depth here, they offer insights into specific conditions where the fused approach excels or where individual modalities may retain advantages

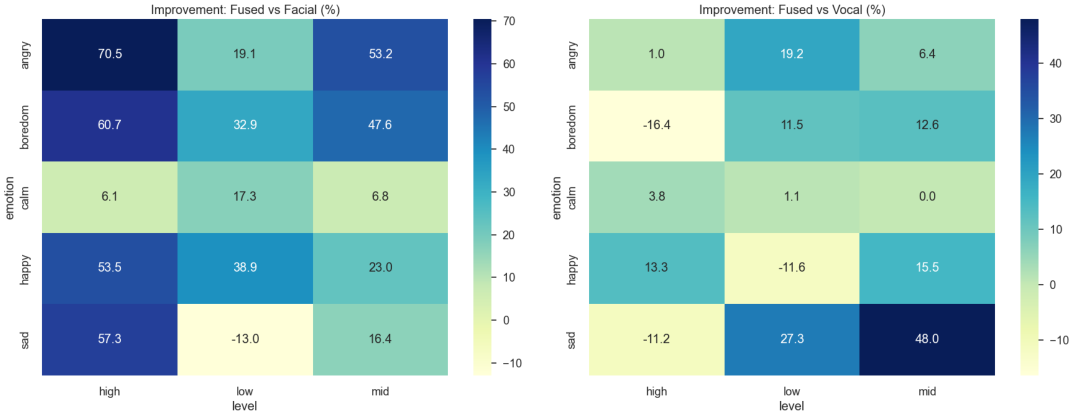


Figure 5.9: Improvement heatmaps: Fused vs Facial and Fused vs Vocal by Emotion and Intensity

The fused method yields an average improvement of 33.92% over the facial method and 6.52% over the vocal method. This disparity reflects the vocal method's stronger baseline performance compared to the facial method, leaving less room for improvement when fused with vocal data.

Emotion-wise improvements further highlight the fused method's efficacy across different emotional categories, as shown in Figure 5.10.

The fused method consistently improves over both individual methods for all emotions. Notable gains over the facial method are observed for "angry" (47.58%) and "boredom" (47.54%), with the smallest improvement for "calm" (10.18%). Over the vocal method, the largest improvement occurs for "sad" (16.27%), while "calm" shows the smallest gain (1.72%). These variations suggest that the fused method's benefits are emotion-specific, likely influenced by the relative strengths of facial and vocal cues for each emotion.

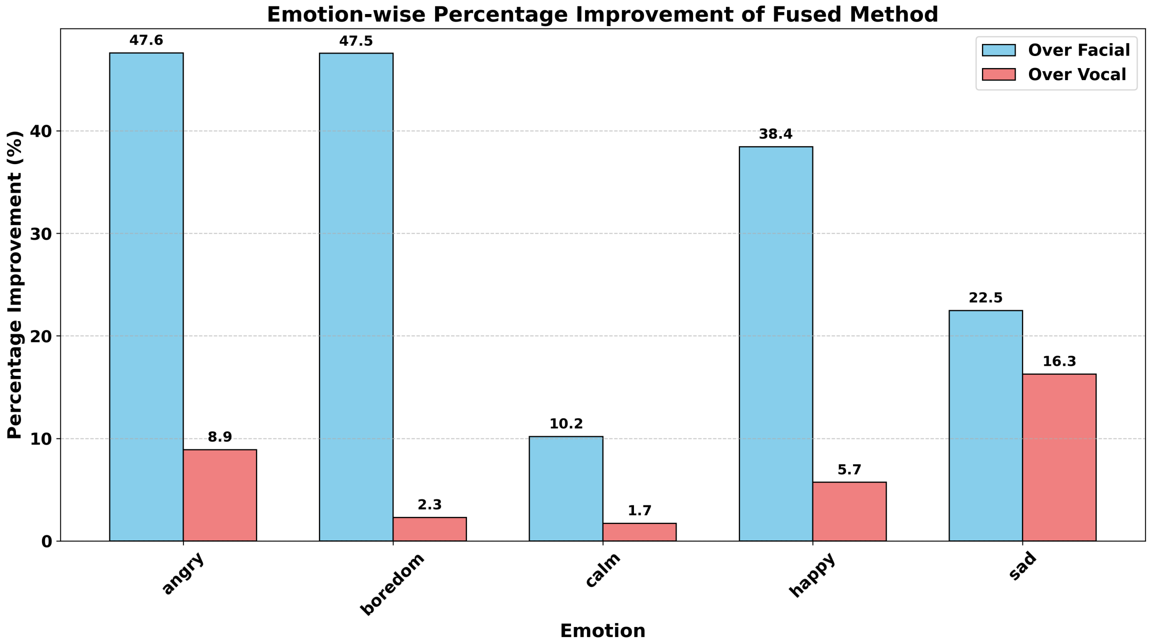


Figure 5.10: Statistical improvement (percentage) by emotion

5.4 Phase 3: Initial Baseline Identification

In this phase, we aimed to identify the initial emotional baseline of each participant by mapping their emotional data points onto the valence-arousal space. For this purpose, we used the fused emotional data collected in the previous phase. Each participant's emotional expressions were visualized on the two-dimensional plane, representing emotional valence and arousal. The full mapping of all participants' initial emotional data points is shown in Figure 5.11.

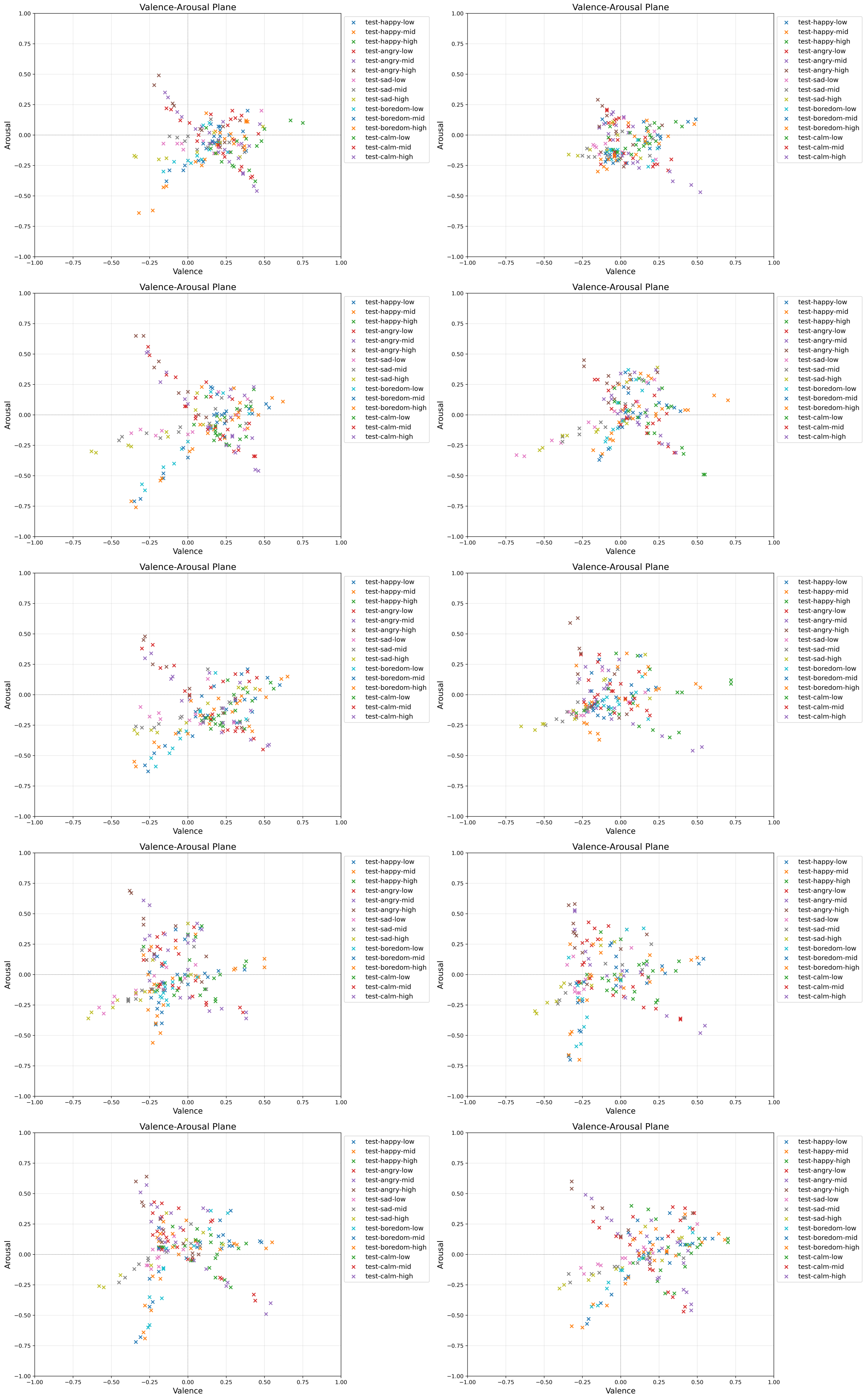


Figure 5.11: Initial Emotional data mapping for all participants

To estimate the baseline more accurately, we applied KDE over the mapped points. After identifying the baseline zones using KDE, we evaluated the correctness and reliability of the results through a participant questionnaire.

Baseline Identification via KDE

After mapping the emotional data points using KDE, we examined the resulting density distributions to find the regions with the highest concentration of data. For each participant, we generated 3D surface plots showing these density values, as presented in Figure 5.12. The peak of each surface represents the potential baseline. We selected a 0.1 × 0.1 area around the point of highest density. This square region was taken as the baseline zone. The exact valence and arousal values at the peak density, along with the top-left and bottom-right coordinates of the baseline square, are presented in Table 5.3.

Table 5.3: Participant baseline values estimated from KDE peak density

Participant	**Valence**	**Arousal**	**Max Density**	**Top-Left**	**Bottom-Right**
p1	0.20	-0.07	597.3728	(0.15, -0.02)	(0.25, -0.12)
p2	-0.04	-0.15	690.2971	(-0.09, -0.10)	(0.01, -0.20)
p3	0.21	-0.07	337.2014	(0.16, -0.02)	(0.26, -0.12)
p4	0.05	0.01	430.7903	(0.00, 0.06)	(0.10, -0.04)
p5	0.17	-0.18	366.9585	(0.12, -0.13)	(0.22, -0.23)
p6	-0.16	-0.09	489.9168	(-0.21, -0.04)	(-0.11, -0.14)
p7	-0.16	-0.10	345.6209	(-0.21, -0.05)	(-0.11, -0.15)
p8	-0.22	-0.06	262.2035	(-0.27, -0.01)	(-0.17, -0.11)
p9	-0.12	0.09	335.896	(-0.17, 0.14)	(-0.07, 0.04)
p10	0.19	0.01	318.8352	(0.14, 0.06)	(0.24, -0.04)

The identified baseline zones were also visualized on a 2D plane for better interpretability, as shown in Figure 5.13. This visual comparison allows us to see how each participant's baseline position varies in the valence-arousal space.

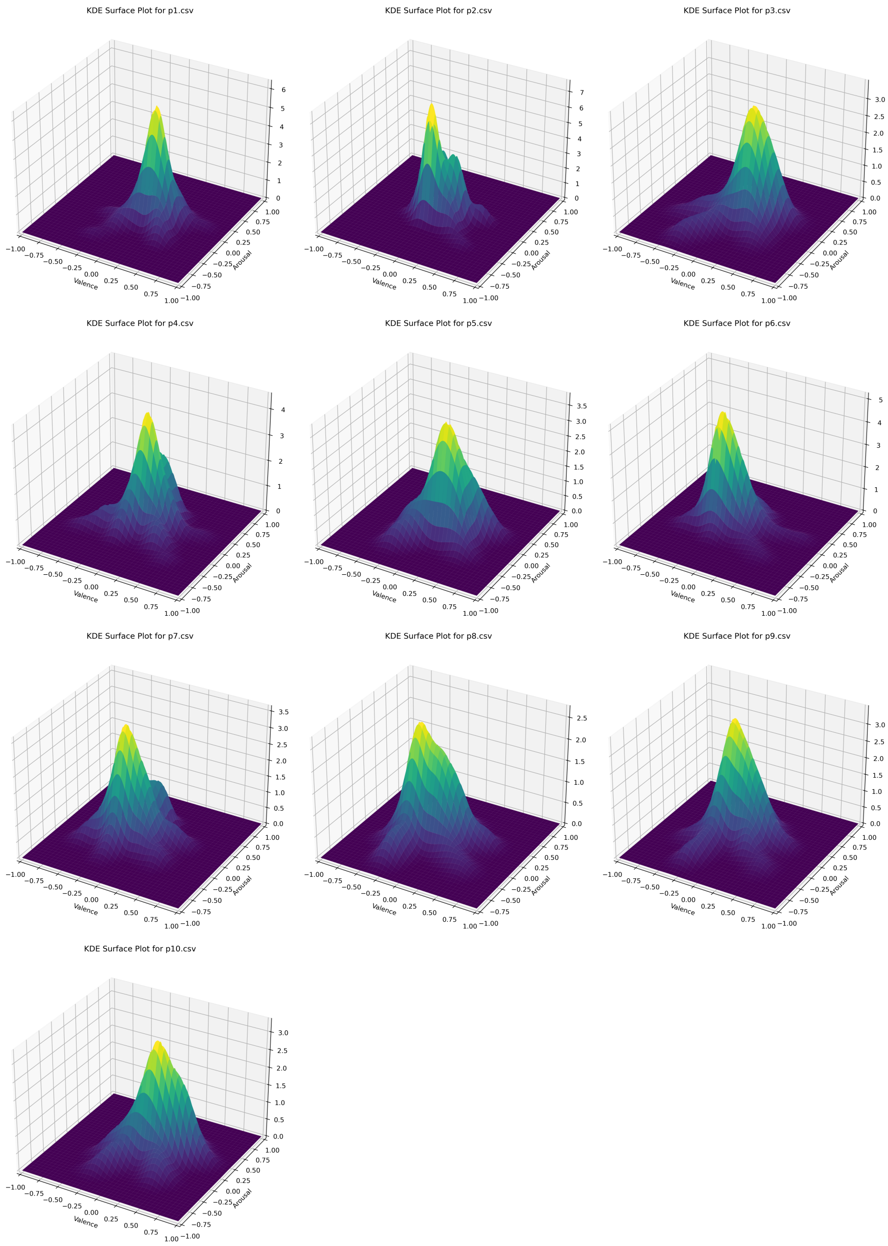


Figure 5.12: 3D surface plots of KDE-based emotional distributions for all participants

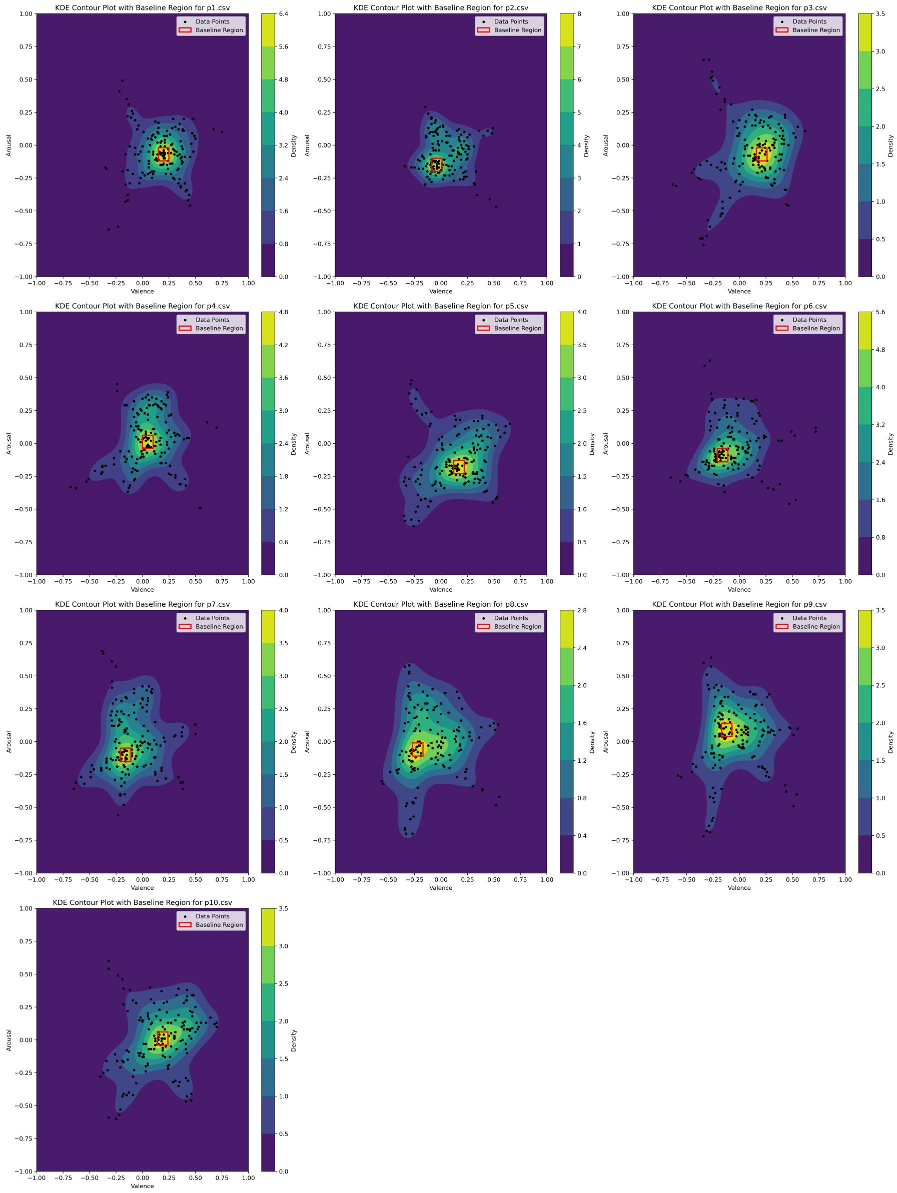


Figure 5.13: 2D mapping of KDE-estimated baseline zones for each participant

Evaluation

To evaluate the accuracy and relevance of the identified baseline values, we conducted a questionnaire with all participants. Each participant was asked to reflect on the emotional states represented in their baseline region and indicate how closely those states matched their typical emotional condition during the experiment sessions.

The full questionnaire and participant responses are provided in the Appendix (see Section A). Based on the collected responses, we calculated the frequency of agreement between the participants and the computed baseline zones. This frequency distribution is shown in Figure 5.14, which highlights the overall agreement levels across all participants.

In addition, we visualized the relationship between the identified baselines and the participants' self-reported emotional states using a valence-arousal scatter plot. This comparison, presented in Figure 5.15, helps illustrate how closely the KDE-based baseline aligns with the participants' own perception.

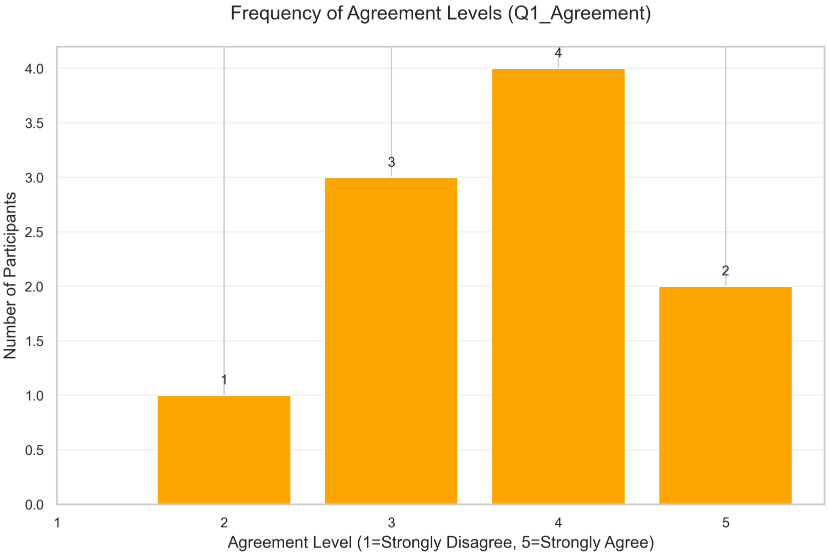


Figure 5.14: Participant agreement frequency with identified baseline values

The analysis of the questionnaire responses provides valuable insights into the accuracy and acceptance of the computed baseline values.

General Agreement Level: The mean agreement score for Question 1 was 3.70 out of 5, indicating that most participants tended to agree with their computed emotional baseline. Furthermore, 60% of participants explicitly expressed agreement, which supports the reliability of the baseline computation method for the majority.

Baseline Discrepancy: To measure how much the participants' proposed baselines differed from the computed values, we used the Euclidean distance. The mean distance was found to be 0.120 with a standard deviation of 0.122. This result shows that, on average, participants' self-identified baseline points differ from the computed values by around 0.12 units in the valence-arousal space. The similar-

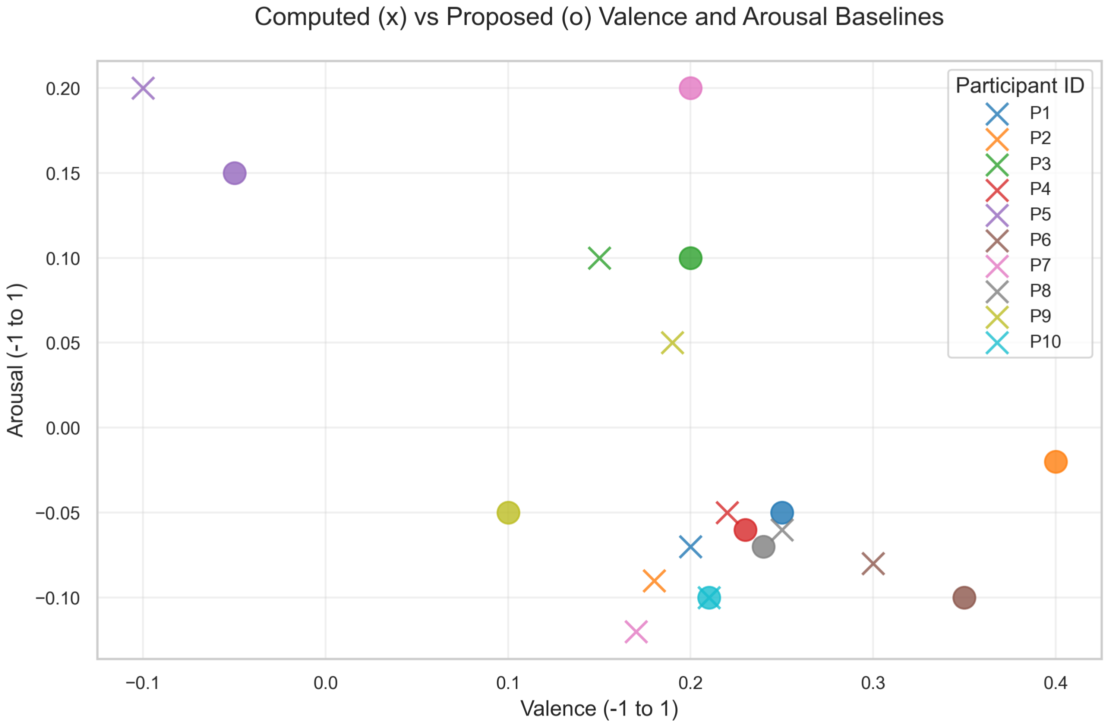


Figure 5.15: Scatter plot comparing identified and participant-proposed baseline coordinates

ity between the mean and standard deviation also indicates a consistent pattern in how much the computed and proposed baselines deviate from each other. Table 5.4 summarizes the computed distances for each participant.

Table 5.4: Participant Agreement and Baseline Discrepancy Summary

Participant ID	**Agreement (1-5)**	**Proposed (Valence, Arousal)**	**Distance**
P1	4	(0.25, -0.05)	0.054
P2	3	(0.40, -0.02)	0.230
P3	4	(0.20, 0.10)	0.305
P4	5	(0.23, -0.06)	0.014
P5	3	(-0.05, 0.15)	0.071
P6	4	(0.35, -0.10)	0.054
P7	2	(0.20, 0.20)	0.321
P8	4	(0.24, -0.07)	0.014
P9	3	(0.10, -0.05)	0.134
P10	5	(0.21, -0.10)	0.000

5.5 Phase 4: LLM Response Evaluation

In this phase, participants were asked to evaluate the responses generated by the language model based on the procedure outlined in Section 4.5. The responses were assessed using a Likert scale across four criteria: *Relevance*, *Emotional Alignment*, *Empathy*, and *Satisfaction*. Each participant rated responses to both a standard (controlled) query and an emotionally-enhanced query.

This evaluation allowed us to compare how the emotional enhancement influenced user perception of the generated responses. The comparison of average ratings between the two types of queries across all four evaluation dimensions is illustrated in Figure 5.16.

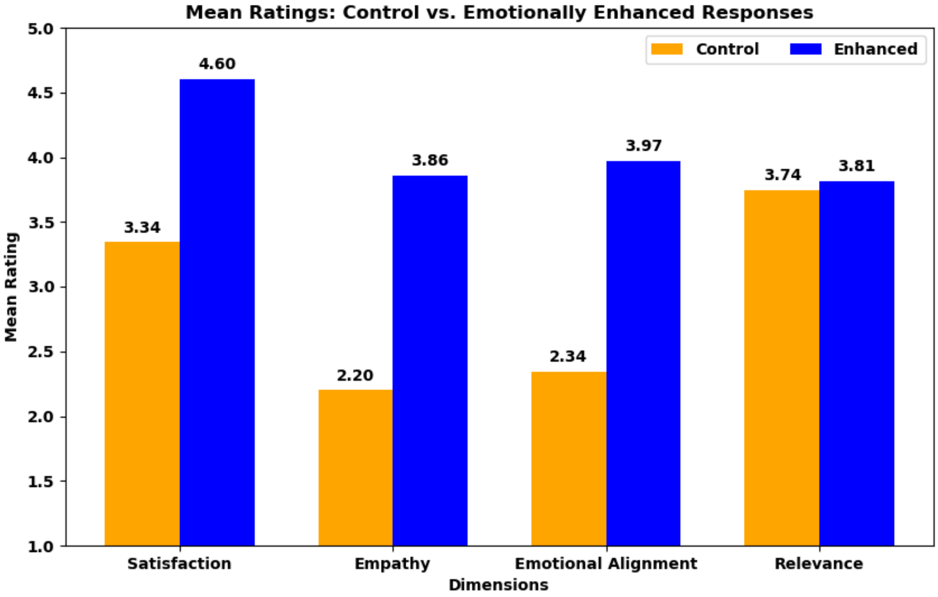


Figure 5.16: Comparison of average ratings for LLM responses

Overall Analysis Results

Dimension	**Mean Control**	**Mean Enhanced**	**Mean Difference**	**Improvement (%)**
Satisfaction	3.343	4.600	1.257	37.6
Empathy	2.200	3.857	1.657	75.3
Emotional Alignment	2.343	3.971	1.629	69.5
Relevance	3.743	3.814	0.071	1.9

Table 5.5: Comparison of participant evaluations for control vs emotionally enhanced responses

As seen in Table 5.5, there is a clear overall increase in user satisfaction, indicating a robust and consistent enhancement in user experience when emotional tailoring is applied. This suggests that emotionally enhanced responses are more engaging and fulfilling for users across various types of queries.

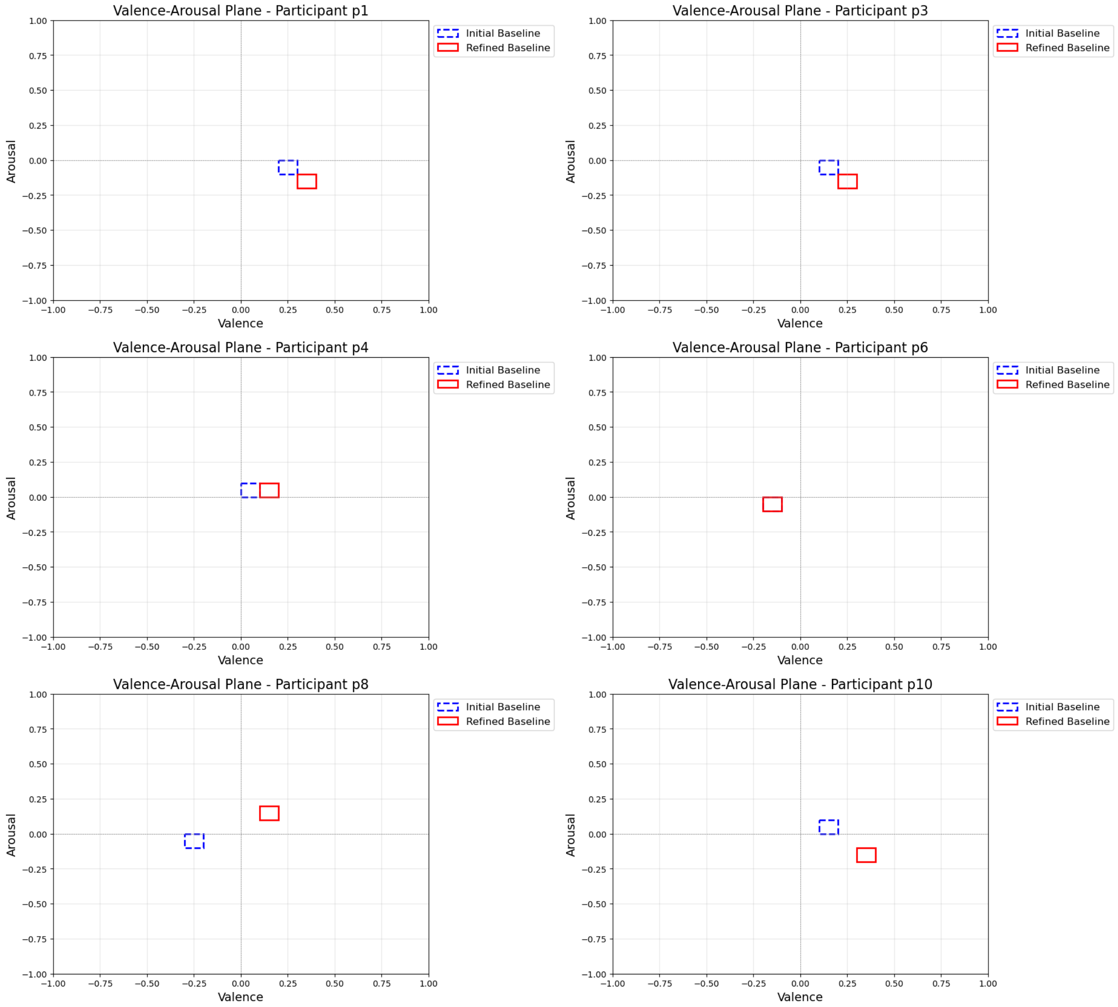
The most significant improvements are observed in *Empathy* (75.3%) and *Emotional Alignment* (69.5%), followed by *Satisfaction* (37.6%). These dimensions, being closely linked to subjective and emotional user experiences, benefit substantially from emotional enhancement.

On the other hand, the impact on *Relevance* is limited. With only a 1.9% improvement and a low proportion of participants reporting a positive change as relevance is more dependent on content correctness, which was already adequately addressed by the control responses.

5.6 Phase 5: Baseline Refinement

Following the initial baseline identification, we selected a subset of participants whose self-identified baselines aligned closely with the computed values. This selection was based on their agreement level. The selected participants included P1, P4, P6, P8, and P10. We conducted initial training using those datapoints available around baseline region.

Subsequently, data was collected during the refinement tasks described in Paragraph 4.6.2.Then observed about the baseline shofts. Initial and refine baseline regions are illustrated in Figure 5.17 for above participants.


Valence-Arousal Plane - Participant p1

Valence-Arousal Plane - Participant p3

Valence-Arousal Plane - Participant p4

Valence-Arousal Plane - Participant p6

Valence-Arousal Plane - Participant p8

Valence-Arousal Plane - Participant p10

Figure 5.17: Initial and refined baseline regions for selected participants.

The results of this section are detailed in Appendix A, where participant feedback on the refined baselines was collected. As shown in the table, the agreement scores and refined baseline coordinates were evaluated using a Likert scale. Out of the six participants, four (P1, P3, P4, and P6) rated the refined baselines with a score of 4 or higher, indicating agreement with the computed values. This reflects a 66.67% agreement rate, suggesting that the refined baseline identification method was successful for the majority of users.

Table 5.6: Participant Agreement on Refined Baselines

PID	**Refined Baseline (V, A)**	**Likert Score**
P1	[(0.3, -0.1), (0.4, -0.2)]	4
P3	[(0.2, -0.1), (0.3, -0.2)]	5
P4	[(0.1, 0.1), (0.2, 0.0)]	5
P6	[(-0.2, 0.0), (-0.1, -0.1)]	4
P8	[(0.1, 0.2), (0.2, 0.1)]	2
P10	[(0.3, -0.1), (0.4, -0.2)]	3

6

Conclusion

This chapter summarizes the key findings of the research by revisiting the research aim and questions, highlighting the contributions made, and discussing limitations and possible future directions. Section 6.1 provides a summary of how each research question was addressed through the different stages of the study. Section 6.2 outlines the main contributions of this work, including the personalized multimodal fusion approach and the novel baseline identification mechanism. Finally, Section 6.3 discusses the limitations of the current study and suggests several directions for future research to improve and expand the system further.

6.1 Conclusions about the Research Questions and Aim

The main aim of this research was to develop a personalized emotion recognition system using facial and vocal signals, and to integrate these emotional insights into LLM responses to make them more emotionally intelligent. The research was divided into five stages, and each stage focused on addressing specific research questions and objectives.

In the first stage, several pre-implemented facial and vocal emotion recognition models were compared to answer RQ 1.1 (What are suitable pre-implemented models that can be used to get a higher accuracy for emotion recognition?). Between HUME and CAGE for facial expression, HUME scored considerably higher overall.

In the case of audio models, both HUME Audio and Wave2Vec2 showed different strengths, but the performance gap between them was smaller compared to facial models. We selected the HUME model for both modalities to maintain consistency. This model selection helped to lay a strong foundation for building a multimodal framework.

In the second stage, we addressed RQ 1.2 (How the recognized emotion values from different modalities fused together in order to get more personalized arousal-valence value?) by fusing the facial and vocal predictions using decision-level fusion. Mean Squared Error was used to calculate the weights for fusion. The fused predictions showed lower average Euclidean distance compared to using either facial or vocal data alone, especially in cases where one modality was weaker. This means the fused model was more accurate and closely matched the ground truth emotions.

The third stage focused on identifying an initial emotional baseline to answer RQ 2.1 (What techniques are most suitable for establishing an initial emotional baseline and how can this baseline be dynamically adjusted over time to reflect changes in the user's emotional responses and self-reported feedback?). We used Kernel Density Estimation based on participant data. To verify accuracy, we collected feedback from participants through a questionnaire. The average agreement score was 3.70 out of 5, showing that most users agreed with the identified baseline values.

In the fourth stage, we evaluated how emotional input could change the output of a LLM model. We used GPT-4o-mini and found a noticeable increase in user satisfaction when emotional context was included with user queries. This confirms the value of emotionally aware responses and successfully answers RQ 3.1 (How does integrating personalized emotional state information with user queries affect the relevance and user satisfaction of responses from LLMs?).

Finally, in the fifth stage, we revisited the emotional baseline and used reinforcement learning to refine it over time. Among six participants, four rated the refined baseline with 4 or higher out of 5, giving a 66.67% agreement rate. This result shows that the method works for most users and improves personalization over time, again supporting RQ 2.1.

Overall, this research shows that using personalized emotional state recognition with facial and audio data can meaningfully improve how LLMs respond to users. The findings give a starting point for building emotionally aware AI systems and highlight areas for future improvements like larger datasets, longer-term user adaptation, and better real-time emotion tracking.

6.2 Research Contributions

This research provides several key contributions in the area of personalized emotion recognition and emotionally aware AI systems. The focus was on building a system that works differently for each user by understanding their unique emotional patterns through both facial and vocal signals.

- **Personalized Emotion Recognition using MSE-based Fusion**

  One of the main contributions is the implementation of a personalized multimodal emotion recognition method. This was done by combining facial and vocal predictions using a decision-level fusion approach. The weights were calculated based on the MSE between the predicted values and user self-reported emotions. This personalized approach was applied for identifying five target emotions: Happy, Angry, Sad, Boredom, and Calm.

- **Emotional Baseline Identification in Arousal-Valence Plane**

  Another key contribution is the method used to identify each user's emotional baseline in the arousal-valence space. KDE was used on the collected data to estimate a stable starting point for each participant's emotional state.

- **Baseline Refinement using Emoji-Based Feedback Mechanism**

  A novel method was introduced to refine the baseline over time using user feedback. Instead of traditional methods, participants gave feedback using simple emoji-based inputs. These responses were used in a reinforcement learning loop to adjust the emotional baseline.

These contributions together help to create a more personalized and emotionally intelligent interaction between users and AI systems. They also offer a foundation for future work in real-time and long-term emotion tracking systems for individuals.

6.3 Limitations and Future Work

This research focused on five basic emotions (Happy, Angry, Sad, Boredom, and Calm) which helped build the core of the system, but these do not fully represent the entire arousal-valence plane. Including a wider range of emotional categories like fear, surprise, or mixed emotions would make the system more complete and useful in real-life applications. Also, the use of only facial and vocal signals may not capture the full complexity of human emotions. Previous studies suggest that physiological signals such as EEG, ECG, and GSR can provide deeper emotional insights and should be explored in future versions. The fusion process in this research was based on a statistical MSE approach, which showed promising results, but more advanced fusion methods like deep learning-based techniques could offer better personalization and flexibility depending on the situation. In terms of emotional baseline estimation, machine learning models can be used to enhance the current KDE-based approach and allow the system to learn and adapt more effectively over time. Lastly, since the experiments were conducted with a small group of participants from similar backgrounds, future work should focus on involving more diverse users in terms of age, language, and culture to ensure the system performs well across different populations.

Bibliography

Barrett, L. F. et al. (2004), 'Journal of personality and social psychology', *Journal of Personality and Social Psychology* **87**(5), 684–697.

Bench, S. W. & Lench, H. C. (2013), 'On the function of boredom', *Behavioral Sciences* **3**(3), 459–472.

Bevilacqua, F., Engström, H. & Backlund, P. (2019), 'Game-calibrated and user-tailored remote detection of stress and boredom in games', *Sensors* **19**(13), 2877.

Blog, N. T. (2018), 'Artwork personalization at netflix', https://medium.com/netflix-techblog/artwork-personalization-c589f074ad76.

Brooks, J. A., Kim, L., Opara, M., Keltner, D., Fang, X., Monroy, M., Corona, R., Tzirakis, P., Baird, A., Metrick, J. et al. (2024), 'Deep learning reveals what facial expressions mean to people in different cultures', *Iscience* **27**(3).

Brooks, J. A., Tzirakis, P., Baird, A., Kim, L., Opara, M., Fang, X., Keltner, D., Monroy, M., Corona, R., Metrick, J. et al. (2023), 'Deep learning reveals what vocal bursts express in different cultures', *Nature Human Behaviour* **7**(2), 240–250.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S. & Narayanan, S. S. (2008), 'Iemocap: Interactive emotional dyadic motion capture database', *Language resources and evaluation* **42**, 335–359.

BuzzFeedVideo (2018), 'Can You Watch This Without Yawning?', https://www.youtube.com/watch?v=M3QYDtSbhrA. Accessed: 2025-01-09.

Calm (2020), 'Breathe bubble', https://www.youtube.com/watch?v=uxayUBd6T7M. Accessed: 2025-01-04.

Choi, Y. K., Lee, S. M. & Li, H. (2013), 'Audio and visual distractions and implicit brand memory: a study of video game players', *Journal of Advertising* **42**(2-3), 219–227.

Cooper, L. (2011), 'A study investigating the relaxation effects of the music track weightless by marconi union in consultation with lyz cooper'.

Cowen, A. S. & Keltner, D. (2021), 'Semantic space theory: A computational approach to emotion', *Trends in Cognitive Sciences* **25**(2), 124–136.

Davidson, R. J. (1998), 'Affective style and affective disorders: Perspectives from affective neuroscience', *Cognition & emotion* **12**(3), 307–330.

Edirisinghe, R. (2020), 'Reinforcement learning for personalized recommendations'.

Ekman, P. & Friesen, W. V. (1971), 'Constants across cultures in the face and emotion.', *Journal of personality and social psychology* **17**(2), 124.

Filmmaker, A. (2013), 'The World's Most BORING video...', https://www.youtube.com/watch?v=1VrYV0odeFY. Accessed: 2025-01-08.

Gelbrich, K., Hagel, J. & Orsingher, C. (2021), 'Emotional support from a digital assistant in technology-mediated services: Effects on customer satisfaction and behavioral persistence', *International Journal of Research in Marketing* **38**(1), 176–193.

Hanbazazh, A. & Reeve, C. (2021), 'Pop-up ads and behaviour patterns: A quantitative analysis involving perception of saudi users', *International Journal of Marketing Studies* **13**(4), 31.

Hume AI (2025*a*), 'Expression measurement'. Accessed: March 28, 2025.
URL: https://www.hume.ai/expression-measurement

Hume AI (2025*b*), 'Expression measurement - Hume API'. Accessed: March 28, 2025.
URL: https://dev.hume.ai/docs/expression-measurement/overview

Inkster, B., Sarda, S. & Subramanian, V. (2018), 'An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being:

real-world data evaluation mixed-methods study', *JMIR mHealth and uHealth* **6**(11), e12106.

Jalal, M. A., Moore, R. K. & Hain, T. (2019), Spatio-temporal context modelling for speech emotion classification, *in* '2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)', IEEE, pp. 853–859.

Jones, C., Scholes, L., Johnson, D., Katsikitis, M. & Carras, M. C. (2014), 'Gaming well: links between videogames and flourishing mental health', *Frontiers in psychology* **5**, 76833.

Just For Laughs Gags (2011), 'Pink Elephant Prank', https://www.youtube.com/watch?v=ZwJfXgTO7J4. Accessed: 2025-01-04.

JustMusicTV (2015), 'Marconi union - weightless (official video)', https://www.youtube.com/watch?v=UfcAVejslrU. Accessed: 2024-12-28.

Kantharia, K. J. & P., G. I. (2015), Facial behavior recognition using soft computing techniques: A survey, *in* 'Proceedings of the International Conference on Advances in Computer Engineering and Applications', IEEE, pp. 30–34.

Kim, Y., Lee, J., Kim, S., Park, J. & Kim, J. (2024), Understanding users' dissatisfaction with chatgpt responses: Types, resolving tactics, and the effect of knowledge level, *in* 'Proceedings of the 29th International Conference on Intelligent User Interfaces', pp. 385–404.

Kollias, D. & Zafeiriou, S. (2018), 'Aff-wild2: Extending the aff-wild database for affect recognition', *arXiv preprint arXiv:1811.07770* .

Kosti, R., Alvarez, J. M., Recasens, A. & Lapedriza, A. (2019), 'Context based emotion recognition using emotic dataset', *IEEE transactions on pattern analysis and machine intelligence* **42**(11), 2755–2766.

Kuijsters, A., Redi, J., De Ruyter, B. & Heynderickx, I. (2016), 'Inducing sadness and anxiousness through visual media: Measurement techniques and persistence', *Frontiers in psychology* **7**, 1141.

Kutsuzawa, G., Umemura, H., Eto, K. & Kobayashi, Y. (2022), 'Classification of

74 facial emoji's emotional states on the valence-arousal axes', *Scientific Reports* **12**(1), 398.

Lim, N. (2016), 'Cultural differences in emotion: differences in emotional arousal level between the east and the west', *Integrative Medicine Research* **5**(2), 105–109.

Liu, S., See, K. C., Ngiam, K. Y., Celi, L. A., Sun, X. & Feng, M. (2022), 'Deep reinforcement learning for personalized treatment recommendation', *Statistics in Medicine* **41**(20), 4034–4056.

Livingstone, S. R. & Russo, F. A. (2018), 'The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english', *PloS one* **13**(5), e0196391.

Lootens, L. (2013), 'The Champ (1979) - Final', https://www.youtube.com/watch?v=b5qwTeCj4jc. Accessed: 2023-10-05.

Lotfian, R. & Busso, C. (2017), 'Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings', *IEEE Transactions on Affective Computing* **10**(4), 471–483.

Maffei, A. & Angrilli, A. (2019), 'E-movie-experimental movies for induction of emotions in neuroscience: An innovative film database with normative data and sex differences', *Plos one* **14**(10), e0223124.

Mariacher, N., Schl"ogl, S. & Monz, A. (2021), Investigating perceptions of social intelligence in simulated human-chatbot interactions, *in* 'Progresses in Artificial Intelligence and Neural Systems', Springer, pp. 513–529.

Markey, A., Chin, A., Vanepps, E. M. & Loewenstein, G. (2014), 'Identifying a reliable boredom induction', *Perceptual and Motor Skills* **119**(1), 237–253.

Martinez-Lucas, L., Abdelwahab, M. & Busso, C. (2020), 'The msp-conversation corpus', *Interspeech 2020* .

Mehendale, N. (2020), 'Facial emotion recognition using convolutional neural networks (ferc)', *SN Applied Sciences* **2**.

Moise, I., Gašević, D. & Hauge, J. B. (2020), 'Reinforcement learning for personalization: A systematic literature review', *Data Science* **3**(2), 107–147.

Mollahosseini, A., Hasani, B. & Mahoor, M. H. (2017), 'Affectnet: A database for facial expression, valence, and arousal computing in the wild', *IEEE Transactions on Affective Computing* **10**(1), 18–31.

Mouse, M. (2007), 'Little Motel', `https://www.youtube.com/watch?v=zqQTODR3kR8`.

Nojavanasghari, B., Baltrušaitis, T., Hughes, C. E. & Morency, L.-P. (2016), Emoreact: a multimodal approach and dataset for recognizing emotional responses in children, *in* 'Proceedings of the 18th acm international conference on multimodal interaction', pp. 137–144.

Norscia, I., Zanoli, A., Gamba, M. & Palagi, E. (2020), 'Auditory contagious yawning is highest between friends and family members: Support to the emotional bias hypothesis', *Frontiers in Psychology* **11**, 442.

Oberlander, L. A. M. & Klinger, R. (2018), An analysis of annotated corpora for emotion classification in text, *in* 'Proceedings of the 27th international conference on computational linguistics', pp. 2104–2119.

Paltoglou, G. & Thelwall, M. (2012), 'Seeing stars of valence and arousal in blog posts', *IEEE Transactions on Affective Computing* **4**(1), 116–123.

Picard, R. W. (2000), 'Affective computing', *Journal of Advanced Composition* .

Plutchik, R. (1980), A general psychoevolutionary theory of emotion, *in* 'Theories of emotion', Elsevier, pp. 3–33.

Poria, S., Cambria, E., Bajpai, R. & Hussain, A. (2017), 'A review of affective computing: From unimodal analysis to multimodal fusion', *Information Fusion* **37**, 98–125.

Posner, J., Russell, J. A. & Peterson, B. S. (2005), 'The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology', *Development and psychopathology* **17**(3), 715–734.

Professor, J. (2024), 'Relaxing Music and Rain — Stress Relief Calm Sleep 1 Hour', https://www.youtube.com/watch?v=PjUZbgZfMOo. Accessed: 2023-10-05.

Reddit users (n.d.), 'Daily song discussion 124: Little motel', https://www.reddit.com/r/ModestMouse/comments/137ax8j/daily_song_discussion_124_little_motel/. Accessed: 2025-04-17.

Ribeiro, F. S., Santos, F. H., Albuquerque, P. B. & Oliveira-Silva, P. (2019), 'Emotional induction through music: Measuring cardiac and electrodermal responses of emotional states and their persistence', *Frontiers in psychology* **10**, 451.

Russell, J. A. (1980), 'A circumplex model of affect.', *Journal of personality and social psychology* **39**(6), 1161.

Sabour, S., Liu, S., Zhang, Z., Liu, J. M., Zhou, J., Sunaryo, A. S., Li, J., Lee, T., Mihalcea, R. & Huang, M. (2024), 'Emobench: Evaluating the emotional intelligence of large language models', *arXiv preprint arXiv:2402.12071* .

Salama AbdELminaam, D., Almansori, A. M., Taha, M. & Badr, E. (2020), 'A deep facial recognition system using computational intelligent algorithms', *Plos one* **15**(12), e0242269.

Savchenko, A. V. (2023), Emotieffnets for facial processing in video-based valence-arousal prediction, expression classification and action unit detection, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 5716–5724.

Savchenko, A. V. (2024), 'Hsemotion team at the 7th abaw challenge: Multi-task learning and compound facial expression recognition', *arXiv preprint arXiv:2407.13184* .

Schuller, B. W., Batliner, A., Amiriparian, S., Barnhill, A., Gerczuk, M., Triantafyllopoulos, A., Baird, A. E., Tzirakis, P., Gagne, C., Cowen, A. S. et al. (2023), The acm multimedia 2023 computational paralinguistics challenge: Emotion share & requests, *in* 'Proceedings of the 31st ACM International Conference on Multimedia', pp. 9635–9639.

Siedlecka, E. & Denson, T. F. (2019), 'Experimental methods for inducing basic emotions: A qualitative review', *Emotion Review* **11**(1), 87–97.

Smidt, K. E. & Suvak, M. K. (2015), 'A brief, but nuanced, review of emotional granularity and emotion differentiation research', *Current Opinion in Psychology* **3**, 48–51.

soundsforyou (2021), 'BUSY TRAFFIC NOISES, CAR HORNS (SOUND EFFECTS) HD', https://www.youtube.com/watch?v=d0k1JFAAMCo. Accessed: 2025-1-05.

Speer, M. E. & Delgado, M. R. (2017), 'Reminiscing about positive memories buffers acute stress responses', *Nature human behaviour* **1**(5), 0093.

Today, P. (2020), 'This song can induce more relaxation than a massage', Accessed: 2025-04-17. https://www.psychologytoday.com/us/blog/the-reality-of-gen-z/202007/this-song-can-induce-more-relaxation-than-a-massage.

Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G. & Pantic, M. (2021), 'Estimation of continuous valence and arousal levels from faces in naturalistic conditions', *Nature Machine Intelligence* .
URL: https://www.nature.com/articles/s42256-020-00280-0

Tzirakis, P., Baird, A., Brooks, J., Gagne, C., Kim, L., Opara, M., Gregory, C., Metrick, J., Boseck, G., Tiruvadi, V. et al. (2023), Large-scale nonverbal vocalization detection using transformers, *in* 'ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 1–5.

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F. & Schuller, B. W. (2023), 'Dawn of the transformer era in speech emotion recognition: Closing the valence gap', *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp. 1–13.

Wagner, N., Mätzler, F., Vossberg, S. R., Schneider, H., Pavlitska, S. & Zöllner, J. M. (2024), 'Cage: Circumplex affect guided expression inference'.

Wang, H., Fu, Y., Liao, Q. & Huang, D. (2019), 'Personalized project recommendations: using reinforcement learning', *EURASIP Journal on Wireless Communications and Networking* **2019**(1), 1–11.

Wang, M. & Chen, Z. (2022), 'Laugh before you study: does watching funny videos before study facilitate learning?', *International Journal of Environmental Research and Public Health* **19**(8), 4434.

Wang, X., Li, X., Yin, Z., Wu, Y. & Liu, J. (2023), 'Emotional intelligence of large language models', *Journal of Pacific Rim Psychology* **17**, 18344909231213958.

Zhang, S., Zhang, S., Huang, T. & Gao, W. (2017), 'Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching', *IEEE transactions on multimedia* **20**(6), 1576–1590.

A

Appendix

Self-reported Emotions and Intensities

This appendix provides the details of the self-reported emotions and intensities used in the emotion elicting task under phase2.

Table A.1: Self-reported values for **Happy** emotion

Participant ID	**Test 1 (Intensity)**	**Test 2 (Intensity)**	**Test 3 (Intensity)**
1	Happy (0.3)	Happy (0.5)	Happy (0.8)
2	Happy (0.6)	Happy (0.7)	Happy (0.5)
3	Happy (0.6)	Happy (0.7)	Happy (0.5)
4	Happy (0.4)	Happy (0.8)	Happy (0.3)
5	Happy (0.6)	Happy (0.7)	Happy (0.5)
6	Happy (0.2)	Happy (0.5)	Happy (0.7)
7	Happy (0.5)	Happy (0.7)	Happy (0.5)
8	Happy (0.7)	Happy (0.7)	Happy (0.5)
9	Happy (0.5)	Happy (0.7)	Happy (0.8)
10	Happy (0.6)	Happy (0.7)	Happy (0.8)

Table A.2: Self-reported values for **Angry** emotion

Participant ID	**Test 1 (Intensity)**	**Test 2 (Intensity)**	**Test 3 (Intensity)**
1	Angry (0.4)	Angry (0.6)	Angry (0.6)
2	Angry (0.3)	Angry (0.2)	Angry (0.3)
3	Angry (0.6)	Angry (0.6)	Angry (0.8)
4	Angry (0.5)	Angry (0.3)	Angry (0.8)
5	Angry (0.6)	Angry (0.6)	Angry (0.8)
6	Angry (0.1)	Angry (0.2)	Angry (0.8)
7	Angry (0.4)	Angry (0.7)	Angry (0.9)
8	Angry (0.5)	Angry (0.7)	Angry (0.8)
9	Angry (0.3)	Angry (0.8)	Angry (0.9)
10	Angry (0.4)	Angry (0.7)	Angry (0.8)

Table A.3: Self-reported values for **Sad** emotion

Participant ID	**Test 1 (Intensity)**	**Test 2 (Intensity)**	**Test 3 (Intensity)**
1	Sad (0.1)	Surprised (0.2)	Sad (0.5)
2	Sad (0.2)	Sad (0.3)	Confusion (0.3)
3	Sad (0.3)	Confused / Surprise (0.8)	Sad (0.8)
4	Sad (0.9)	Confused (0.4)	Sad (0.6)
5	Sad (0.3)	Confused / Surprise (0.8)	Sad (0.8)
6	Confused (0.4)	Sad (0.5)	Sad (0.7)
7	Sad (0.7)	Sad (0.5)	Sad (0.8)
8	Sad (0.3)	Sad (0.5)	Sad (0.7)
9	Surprised (0.3)	Sad (0.5)	Sad (0.8)
10	Sad (0.4)	Confused / Surprise (0.8)	Sad (0.9)

Table A.4: Self-reported values for **Boredom** emotion

Participant ID	**Test 1 (Intensity)**	**Test 2 (Intensity)**	**Test 3 (Intensity)**
1	Boredom (0.3)	Boredom (0.6)	Boredom (0.8)
2	Boredom (0.1)	Boredom (0.2)	Boredom (0.4)
3	Boredom (0.7)	Boredom (0.9)	Boredom (1)
4	Boredom (0.4)	Boredom (0.6)	Boredom (0.5)
5	Boredom (0.7)	Boredom (0.9)	Boredom (1)
6	Neutral (0.1)	Neutral (0.2)	Boredom (0.5)
7	Boredom (0.3)	Boredom (0.5)	Boredom (0.7)
8	Boredom (0.7)	Boredom (0.9)	Boredom (0.9)
9	Boredom (0.5)	Boredom (0.7)	Boredom (0.7)
10	Boredom (0.6)	Boredom (0.8)	Boredom (0.9)

Table A.5: Self-reported values for **Calm** emotion

Participant ID	**Test 1 (Intensity)**	**Test 2 (Intensity)**	**Test 3 (Intensity)**
1	Calm (0.5)	Calm (0.4)	Calm (0.8)
2	Calm (0.1)	Calm (0.3)	Calm (0.7)
3	Calm (0.4)	Calm (0.7)	Calm (0.8)
4	Calm (0.8)	Calm (0.6)	Calm (0.5)
5	Calm (0.4)	Calm (0.7)	Calm (0.8)
6	Calm (0.5)	Calm (0.2)	Calm (0.7)
7	Calm (0.3)	Calm (0.5)	Calm (0.6)
8	Calm (0.4)	Calm (0.6)	Calm (0.8)
9	Calm (0.4)	Focused (0.6)	Calm (0.8)
10	Calm (0.4)	Calm (0.7)	Calm (0.7)

Modality Weights and MSE for Participants

This section provides the Mean Squared Error (MSE) values for each participant, along with the corresponding modality weights for the facial and vocal modalities.

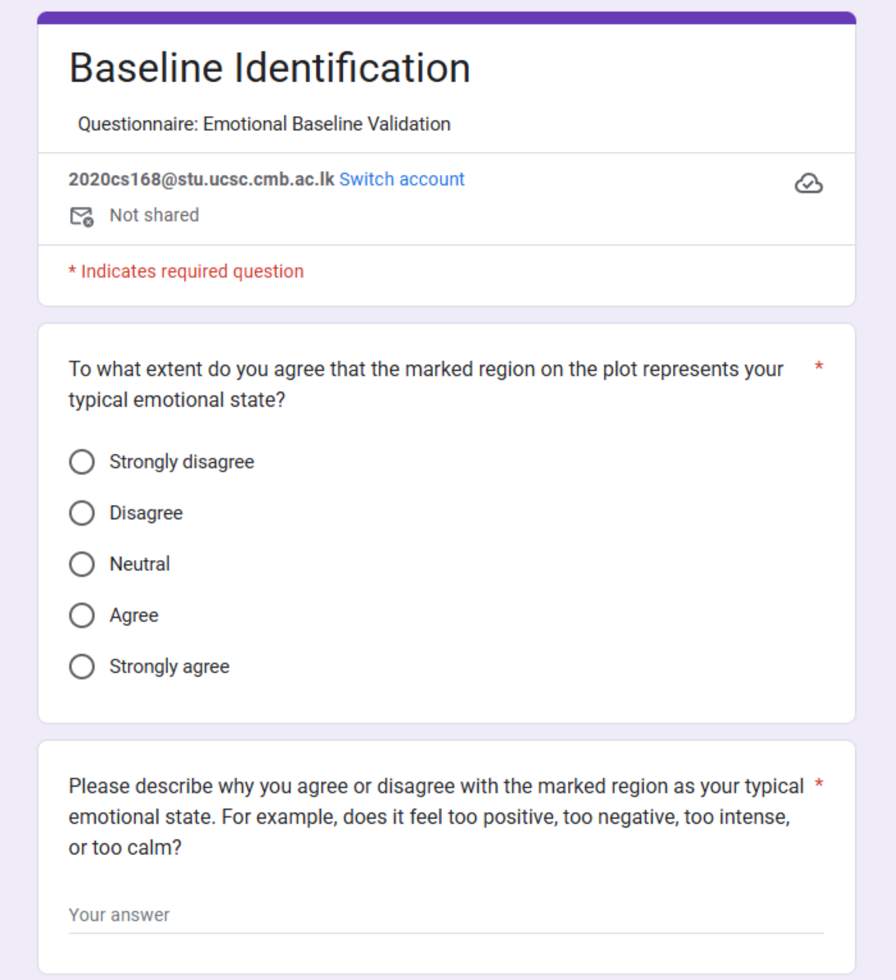
Table A.6: Modality Weights and MSE for Participants (Left: 1–5, Right: 6–10)

Participants 1–5						**Participants 6–10**					
P.	**Emo**	**W_f**	**W_v**	**MSE_f**	**MSE_v**	**P.**	**Emo**	**W_f**	**W_v**	**MSE_f**	**MSE_v**
1	Hpy	0.82	0.18	0.00118	0.00145	6	Hpy	0.85	0.15	0.00122	0.00124
1	Ang	0.48	0.52	0.03890	0.00636	6	Ang	0.43	0.57	0.00433	0.00025
1	Sad	0.87	0.13	0.00133	0.00375	6	Sad	0.86	0.14	0.00092	0.00119
1	Brd	0.86	0.14	0.00520	0.00294	6	Brd	0.88	0.12	0.00015	0.00010
1	Clm	0.46	0.54	0.00800	0.01214	6	Clm	0.41	0.59	0.00261	0.00084
2	Hpy	0.84	0.16	0.01389	0.00333	7	Hpy	0.82	0.18	0.01931	0.00395
2	Ang	0.44	0.56	0.00597	0.00815	7	Ang	0.42	0.58	0.01410	0.00064
2	Sad	0.83	0.17	0.00474	0.00050	7	Sad	0.87	0.13	0.00149	0.00105
2	Brd	0.88	0.12	0.00114	0.00336	7	Brd	0.85	0.15	0.00084	0.00035
2	Clm	0.42	0.58	0.00485	0.00169	7	Clm	0.46	0.54	0.00468	0.00248
3	Hpy	0.79	0.21	0.03200	0.00041	8	Hpy	0.80	0.20	0.02560	0.00447
3	Ang	0.45	0.55	0.02763	0.00305	8	Ang	0.50	0.50	0.00101	0.00332
3	Sad	0.85	0.15	0.03398	0.00301	8	Sad	0.84	0.16	0.00327	0.00142
3	Brd	0.82	0.18	0.01058	0.00035	8	Brd	0.81	0.19	0.00237	0.00017
3	Clm	0.43	0.57	0.01746	0.01366	8	Clm	0.44	0.56	0.01061	0.00403
4	Hpy	0.89	0.11	0.00078	0.00350	9	Hpy	0.86	0.14	0.00137	0.00281
4	Ang	0.49	0.51	0.02585	0.00849	9	Ang	0.48	0.52	0.00152	0.00053
4	Sad	0.88	0.12	0.00232	0.00666	9	Sad	0.89	0.11	0.00034	0.00092
4	Brd	0.81	0.19	0.00918	0.00343	9	Brd	0.84	0.16	0.00322	0.00036
4	Clm	0.47	0.53	0.00461	0.00924	9	Clm	0.49	0.51	0.00995	0.00903
5	Hpy	0.81	0.19	0.00792	0.00020	10	Hpy	0.90	0.10	0.00137	0.00700
5	Ang	0.46	0.54	0.01917	0.00193	10	Ang	0.47	0.53	0.01006	0.00221
5	Sad	0.84	0.16	0.01656	0.00070	10	Sad	0.83	0.17	0.01767	0.00769
5	Brd	0.83	0.17	0.00988	0.00017	10	Brd	0.82	0.18	0.00830	0.00255
5	Clm	0.45	0.55	0.00636	0.00473	10	Clm	0.46	0.54	0.00038	0.00912

Emotion Eliciting Tasks Application

The emotion eliciting tasks application used in this research is available at https://github.com/avishka-sathyanjana/amelia-client.git. This application was used to systematically elicit target emotions during phase 2 of the study and to record the participants' facial and vocal responses.

Appendix B - Questionnaire for Baseline Evaluation



Baseline Identification

Questionnaire: Emotional Baseline Validation

2020cs168@stu.ucsc.cmb.ac.lk Switch account

Not shared

* Indicates required question

To what extent do you agree that the marked region on the plot represents your typical emotional state? *

○ Strongly disagree

○ Disagree

○ Neutral

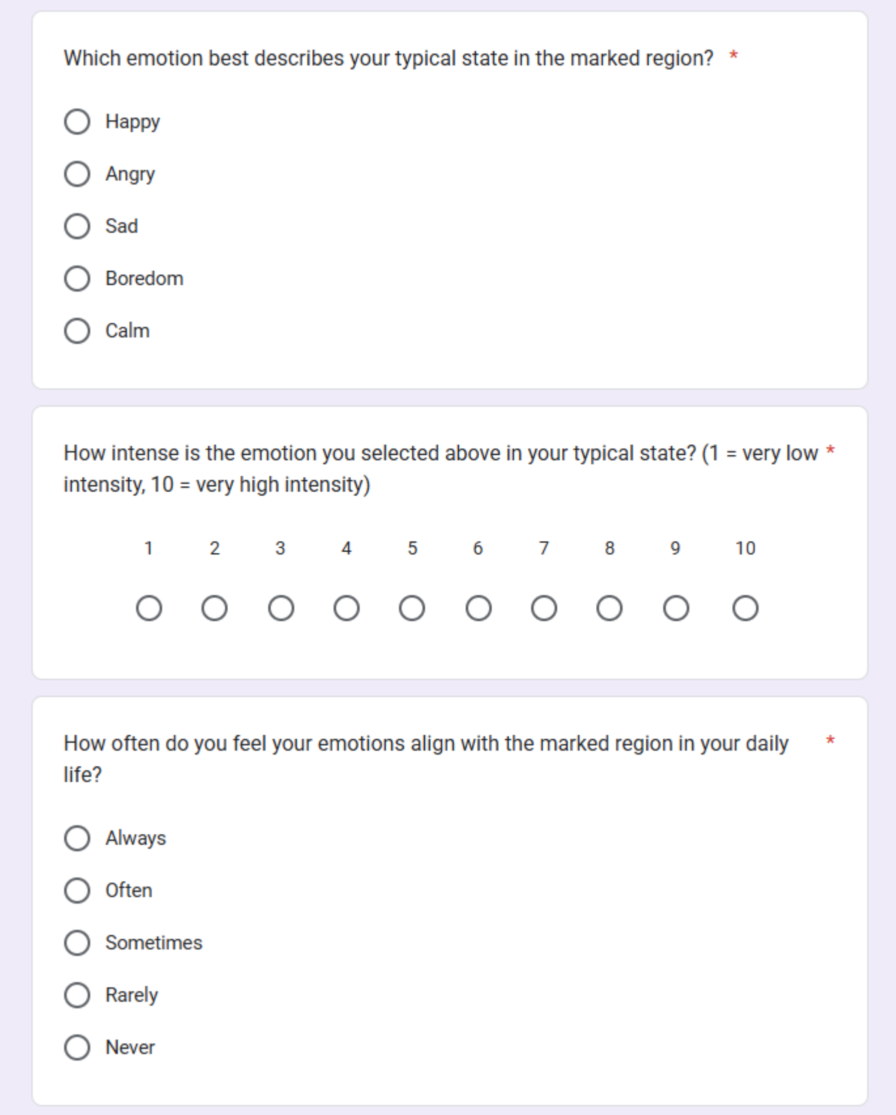
○ Agree

○ Strongly agree

Please describe why you agree or disagree with the marked region as your typical emotional state. For example, does it feel too positive, too negative, too intense, or too calm? *

Your answer

Figure A.1: Baseline evaluation Questionnaire Screenshot 1



Which emotion best describes your typical state in the marked region? *

○ Happy

○ Angry

○ Sad

○ Boredom

○ Calm

How intense is the emotion you selected above in your typical state? (1 = very low intensity, 10 = very high intensity) *

1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○

How often do you feel your emotions align with the marked region in your daily life? *

○ Always

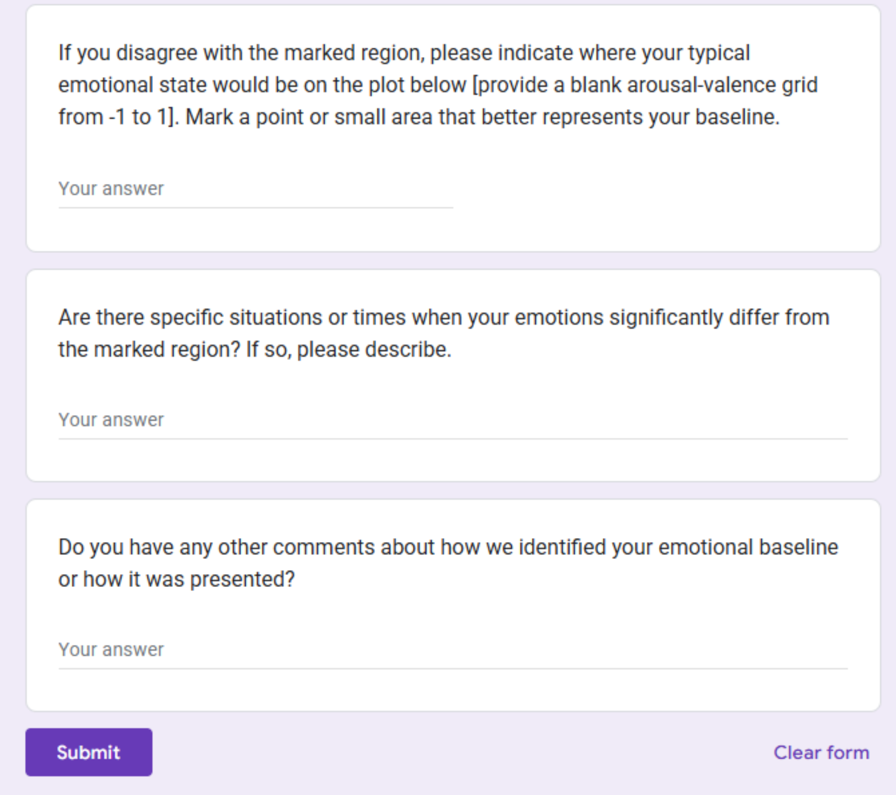
○ Often

○ Sometimes

○ Rarely

○ Never

Figure A.2: Baseline evaluation Questionnaire Screenshot 2



If you disagree with the marked region, please indicate where your typical emotional state would be on the plot below [provide a blank arousal-valence grid from -1 to 1]. Mark a point or small area that better represents your baseline.

Your answer

Are there specific situations or times when your emotions significantly differ from the marked region? If so, please describe.

Your answer

Do you have any other comments about how we identified your emotional baseline or how it was presented?

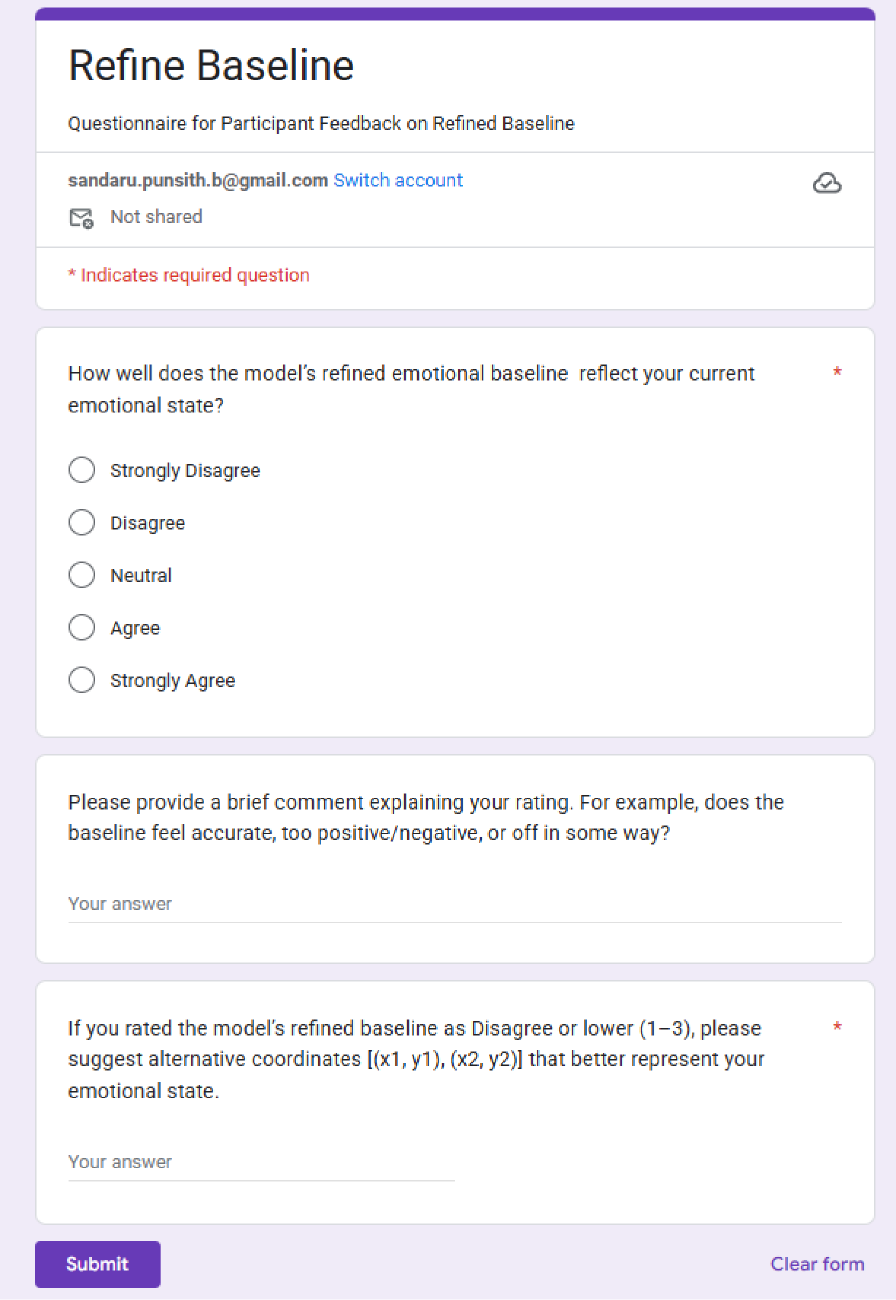
Your answer

Figure A.3: Baseline evaluation Questionnaire Screenshot 3

Participant Responses and Baseline Data

- Link to the response results: https://drive.google.com/file/d/1h2wEjiNGcKktWG46RgT view?usp=sharing
- Identified Baseline Data of the participants: https://drive.google.com/drive/folders/1zxJeYd1LYd66iCwQO1sZSns3szW29ZV4?usp=drive_link
- CSV file of fused emotional data: https://drive.google.com/file/d/1USmPup8HeVyokCh view?usp=drive_link

Appendix C - Questionnaire for Refine Baseline Evaluation



Refine Baseline

Questionnaire for Participant Feedback on Refined Baseline

sandaru.punsith.b@gmail.com Switch account

Not shared

* Indicates required question

How well does the model's refined emotional baseline reflect your current emotional state? *

○ Strongly Disagree

○ Disagree

○ Neutral

○ Agree

○ Strongly Agree

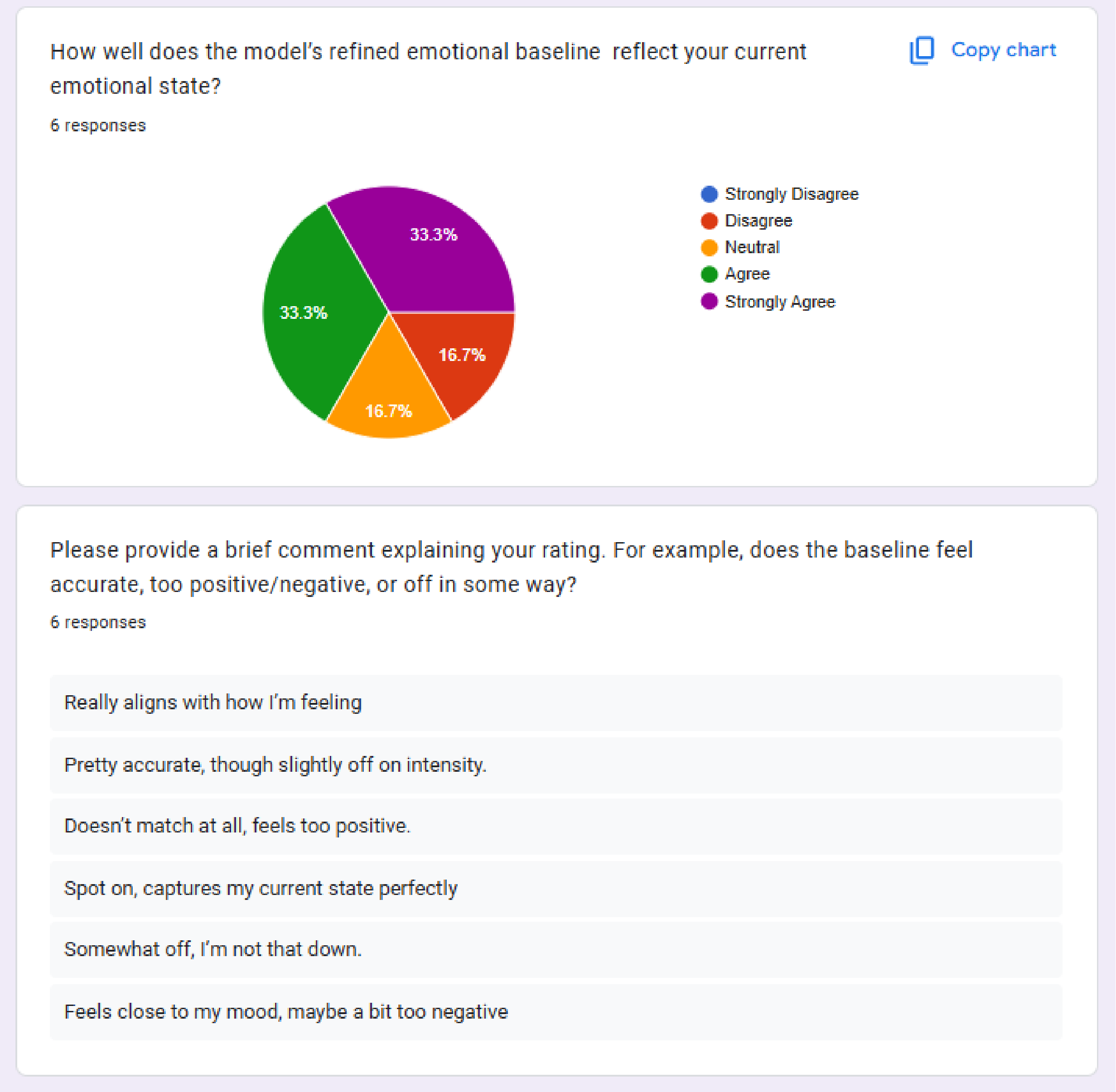
Please provide a brief comment explaining your rating. For example, does the baseline feel accurate, too positive/negative, or off in some way?

Your answer

If you rated the model's refined baseline as Disagree or lower (1–3), please suggest alternative coordinates [(x1, y1), (x2, y2)] that better represent your emotional state. *

Your answer

Figure A.4: Baseline Refinement Questionnaire Screenshot 1



How well does the model's refined emotional baseline reflect your current emotional state?

6 responses

Please provide a brief comment explaining your rating. For example, does the baseline feel accurate, too positive/negative, or off in some way?

6 responses

Really aligns with how I'm feeling

Pretty accurate, though slightly off on intensity.

Doesn't match at all, feels too positive.

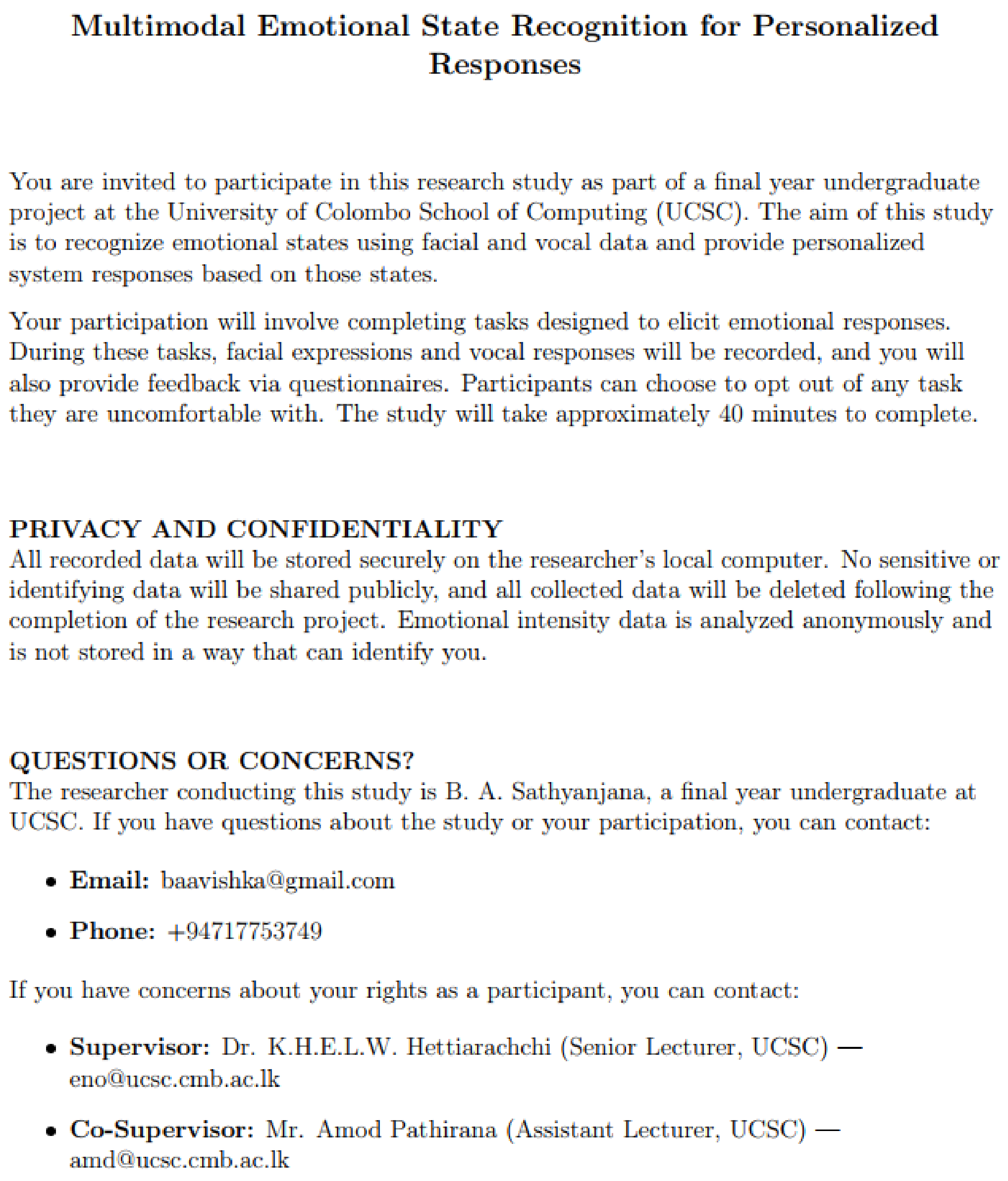
Spot on, captures my current state perfectly

Somewhat off, I'm not that down.

Feels close to my mood, maybe a bit too negative

Figure A.5: Results for the questionnaire

Appendix D - Informed Consent Form



Multimodal Emotional State Recognition for Personalized Responses

You are invited to participate in this research study as part of a final year undergraduate project at the University of Colombo School of Computing (UCSC). The aim of this study is to recognize emotional states using facial and vocal data and provide personalized system responses based on those states.

Your participation will involve completing tasks designed to elicit emotional responses. During these tasks, facial expressions and vocal responses will be recorded, and you will also provide feedback via questionnaires. Participants can choose to opt out of any task they are uncomfortable with. The study will take approximately 40 minutes to complete.

PRIVACY AND CONFIDENTIALITY

All recorded data will be stored securely on the researcher's local computer. No sensitive or identifying data will be shared publicly, and all collected data will be deleted following the completion of the research project. Emotional intensity data is analyzed anonymously and is not stored in a way that can identify you.

QUESTIONS OR CONCERNS?

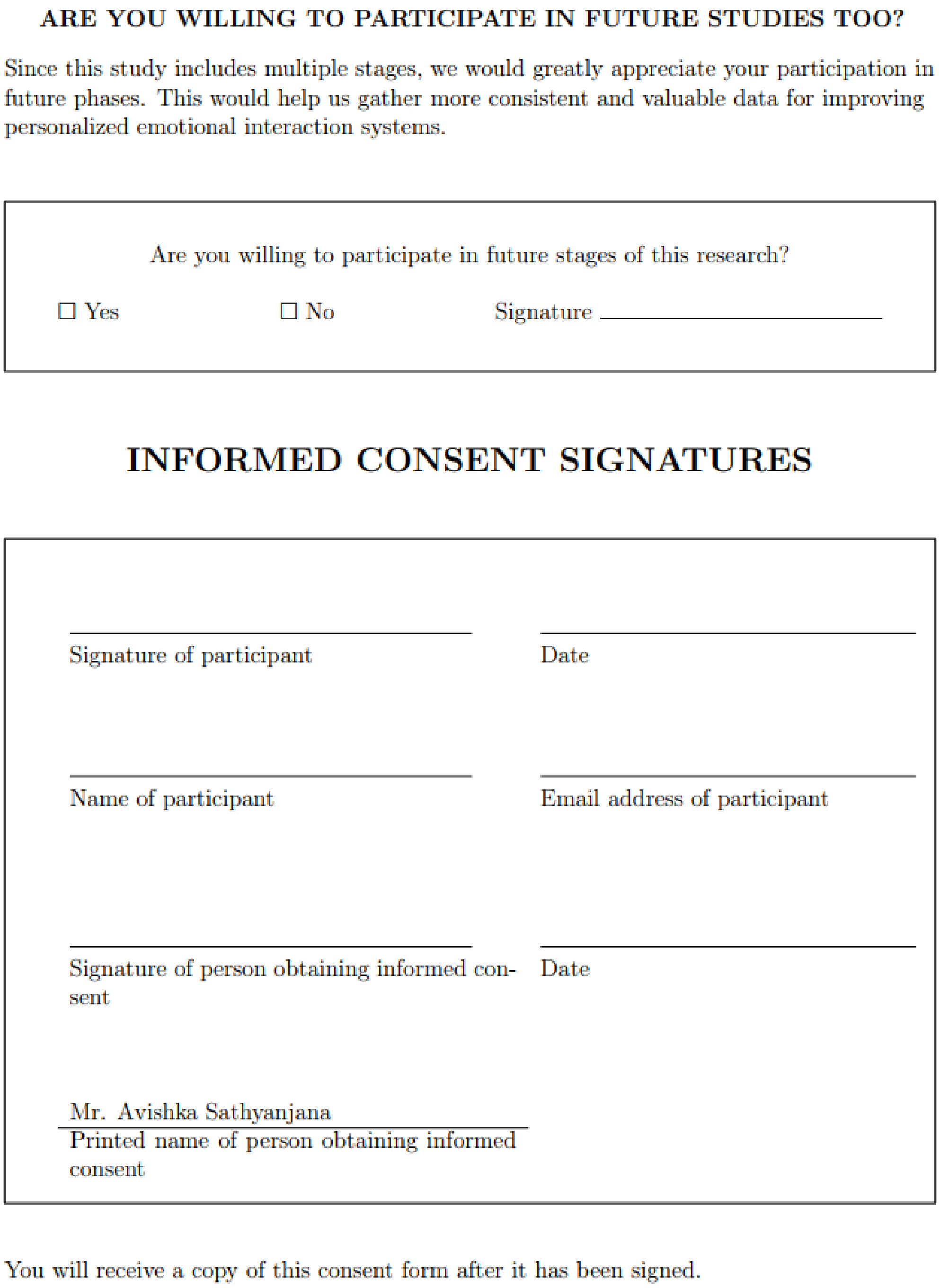
The researcher conducting this study is B. A. Sathyanjana, a final year undergraduate at UCSC. If you have questions about the study or your participation, you can contact:

- **Email:** baavishka@gmail.com
- **Phone:** +94717753749

If you have concerns about your rights as a participant, you can contact:

- **Supervisor:** Dr. K.H.E.L.W. Hettiarachchi (Senior Lecturer, UCSC) — eno@ucsc.cmb.ac.lk
- **Co-Supervisor:** Mr. Amod Pathirana (Assistant Lecturer, UCSC) — amd@ucsc.cmb.ac.lk

Figure A.6: Informed Consent Form Page 1



ARE YOU WILLING TO PARTICIPATE IN FUTURE STUDIES TOO?

Since this study includes multiple stages, we would greatly appreciate your participation in future phases. This would help us gather more consistent and valuable data for improving personalized emotional interaction systems.

Are you willing to participate in future stages of this research?

□ Yes □ No Signature ____________________

INFORMED CONSENT SIGNATURES

____________________ ____________________
Signature of participant Date

____________________ ____________________
Name of participant Email address of participant

____________________ ____________________
Signature of person obtaining informed consent Date

Mr. Avishka Sathyanjana
____
Printed name of person obtaining informed consent

You will receive a copy of this consent form after it has been signed.

Figure A.7: Informed Consent Form Page 2

Appendix E - Code listings

Reinforcement Learning Components

```python
# Model Parameters
grid_size = 10              # 10x10 grid for arousal-valence plane
num_states = grid_size * grid_size  # Total number of states (100)
num_actions = 5             # Actions: left, right, up, down, stay
alpha = 0.1                 # Learning rate
gamma = 0.9                 # Discount factor
lambda_trace = 0.9          # Eligibility trace decay rate
epsilon = 0.3               # Initial exploration probability
epsilon_min = 0.05          # Minimum exploration probability
epsilon_decay = 0.999       # Exploration decay rate
N = 100                     # Interval for direct feedback
initial_baseline_a = 0.0    # Initial arousal baseline
initial_baseline_v = 0.0    # Initial valence baseline
```

Listing A.1: Model Parameters

```python
def state_to_index(i, j, grid_size=10):
    """Convert grid coordinates (i, j) to a single state index."""
    return i * grid_size + j

def index_to_state(idx, grid_size=10):
    """Convert a state index to grid coordinates (i, j)."""
    i = idx // grid_size
    j = idx % grid_size
    return i, j

def continuous_to_state(a, v, grid_size=10):
    """Map continuous arousal (a) and valence (v) in [-1, 1]
```

```
        to a state index."""
    step = 2 / grid_size
    i = max(0, min(grid_size - 1, int(np.floor((a + 1) / step))))
    j = max(0, min(grid_size - 1, int(np.floor((v + 1) / step))))
    return state_to_index(i, j)

def get_center(state, grid_size=10):
    """Get the center coordinates (arousal, valence) of a state."""
    i, j = index_to_state(state, grid_size)
    step = 2 / grid_size
    center_a = -1 + (i + 0.5) * step
    center_v = -1 + (j + 0.5) * step
    return center_a, center_v
```

Listing A.2: State and Coordinate Conversion

```
def get_action(state, Q, epsilon, num_actions=5):
    """Select an action using epsilon-greedy policy."""
    if np.random.rand() < epsilon:
        return np.random.randint(num_actions)  # Explore: random action
    return np.argmax(Q[state])  # Exploit: best action

def get_next_state(state, action, grid_size=10):
    """Determine the next state based on the current state and action."""
    i, j = index_to_state(state, grid_size)
    if action == 0:    # Move left
        j = max(0, j - 1)
    elif action == 1:  # Move right
        j = min(grid_size - 1, j + 1)
    elif action == 2:  # Move up
        i = min(grid_size - 1, i + 1)
```

```python
    elif action == 3:  # Move down
        i = max(0, i - 1)
    # Action 4: Stay, no change
    return state_to_index(i, j)
```

Listing A.3: Action Selection and State Transition

```python
def get_reward(current_state, D, grid_size=10):
    """Calculate reward as negative distance to data point D, or 0 if None."""
    if D is None:
        return 0
    center_a, center_v = get_center(current_state, grid_size)
    a_d, v_d = D
    distance = np.sqrt((center_a - a_d) ** 2 + (center_v - v_d) ** 2)
    return -distance
```

Listing A.4: Reward Calculation

```python
def get_direct_feedback(t):
    """
    Simulate direct feedback from the user (e.g., via emoji system).
    Returns: tuple (arousal_b, valence_b)
    """
    # Placeholder: Replace with actual user input system
    return (0.1, 0.2)  # Example feedback

def has_emotional_data(t):
    """
    Check if emotional data is available at time t.
    Returns: bool
    """
    # Placeholder: 50% chance of data availability
    return np.random.rand() < 0.5

```

```python
def get_emotional_data(t):
    """
    Get emotional data when available (e.g., from calm states).
    Returns: tuple (arousal_e, valence_e)
    """
    # Placeholder: Replace with actual data collection
    return (0.05, 0.15)  # Example emotional data
```

Listing A.5: Placeholder Functions for Data Input

```python
def run_q_learning():
    """Main Q-learning loop with eligibility traces."""
    # Initialize Q-table and eligibility traces
    Q = np.zeros((num_states, num_actions))
    e = np.zeros((num_states, num_actions))

    # Set initial state from initial baseline
    current_state = continuous_to_state(initial_baseline_a, initial_baseline_v)

    # Initialize epsilon for exploration
    current_epsilon = epsilon

    # Training loop
    num_steps = 10000  # Number of training steps
    for t in range(num_steps):
        # Select action using epsilon-greedy
        action = get_action(current_state, Q, current_epsilon, num_actions)

        # Transition to next state
        next_state = get_next_state(current_state, action, grid_size)

        # Determine data D based on time step and
```

```
           availability
        D = None
        if (t + 1) % N == 0:
            D = get_direct_feedback(t + 1)  # Direct feedback every N steps
        elif has_emotional_data(t + 1):
            D = get_emotional_data(t + 1)   # Emotional data if available

        # Calculate reward
        reward = get_reward(next_state, D, grid_size)

        # Compute temporal difference error
        delta = reward + gamma * np.max(Q[next_state]) - Q[current_state, action]

        # Increment eligibility trace for current state-action pair
        e[current_state, action] += 1

        # Update Q-table and decay eligibility traces for all state-action pairs
        for s in range(num_states):
            for a in range(num_actions):
                Q[s, a] += alpha * delta * e[s, a]
                e[s, a] *= gamma * lambda_trace

        # Update current state
        current_state = next_state

        # Update exploration rate
        current_epsilon = max(epsilon_min, current_epsilon * epsilon_decay)

```

```
    # Return final estimated baseline for evaluation
    final_a, final_v = get_center(current_state)
    return final_a, final_v, Q
```

Listing A.6: Q-Learning with Eligibility Traces