On Multimodal Emotion Recognition for Human-Chatbot Interaction in the Wild

Nikola Kovačević
ETH Zurich
Zurich, Switzerland
nikola.kovacevic@inf.ethz.ch

Christian Holz
ETH Zurich
Zurich, Switzerland
christian.holz@inf.ethz.ch

Markus Gross
ETH Zurich
Zurich, Switzerland
grossm@inf.ethz.ch

Rafael Wampfler
ETH Zurich
Zurich, Switzerland
rafael.wampfler@inf.ethz.ch

ABSTRACT

The field of natural language generation is swiftly evolving, giving rise to powerful conversational characters for use in different applications such as entertainment, education, and healthcare. A central aspect of these applications is providing personalized interactions, driven by the ability of the characters to recognize and adapt to user emotions. Current emotion recognition models primarily rely on datasets collected from actors or in controlled laboratory settings focusing on human-human interactions, which hinders their adaptability to real-world applications for conversational agents. In this work, we unveil the complexity of human-chatbot emotion recognition in the wild. We collected a multimodal dataset consisting of text, audio, and video recordings from 99 participants while they conversed with a GPT-3-based chatbot over three weeks. Using different transformer-based multimodal emotion recognition networks, we provide evidence for a strong domain gap between human-human interaction and human-chatbot interaction that is attributed to the subjective nature of self-reported emotion labels, the reduced activation and expressivity of the face, and the inherent subtlety of emotions in such settings, emphasizing the challenges of recognizing user emotions in real-world contexts. We show how personalizing our model to the user increases the model performance by up to 38% (user emotions) and up to 41% (perceived chatbot emotions), highlighting the potential of personalization for overcoming the observed domain gap.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; **Natural language interfaces**; *User models*; *User studies.*

KEYWORDS

multimodal emotion recognition; conversational agents; chatbots; personalization; human-chatbot interaction

ACM Reference Format:
Nikola Kovačević, Christian Holz, Markus Gross, and Rafael Wampfler. 2024. On Multimodal Emotion Recognition for Human-Chatbot Interaction in the Wild. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 4–8, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3678957.3685759

1 INTRODUCTION

Emotions are intricate facets of human experience, influencing our perceptions, decisions, and interactions in profound ways [32]. Humans express emotions through a variety of signals such as text and speech, through vocal or facial cues, or through biological signals. Accurately recognizing these emotions is pivotal in understanding the sentiments and responses of individuals [31], which enables multiple use cases across various domains, e.g., for improving a business's product offerings by gauging public sentiment [5], offering vital insights into patients' mental health [27, 68], enhancing the effectiveness of e-learning systems [30], and improving automotive safety [67].

In contrast to static emotion recognition (SER), emotion recognition in conversation (ERC) models the conversational dynamics of the involved parties over time. Apart from human conversation, conversations with virtual agents are becoming more ubiquitous in a large number of real-world scenarios such as personal assistance [3], customer service [26], and video games [44], and are about to enter more challenging applications in education [1, 9] and health care [37, 54]. Attributed to the rapid and transformative evolution of natural language generation, conversational agents are transcending the realm of generic information conveyance and becoming wholesome companions that engage users in interactive, meaningful conversations.

ERC plays a pivotal role within the domain of these virtual conversational agents [49] as it enables them to perceive and adapt to users' emotional states, enhancing the quality of interactions and offering a more personalized and engaging experience [18, 24, 47], which makes such agents indispensable tools for a wide array of practical and experimental applications.

Efforts to address ERC have brought forward an abundant collection of emotion recognition datasets [12, 13, 40, 48] and corresponding emotion recognition models that cover a wide variety of neural architectures entailing both unimodal and multimodal approaches [6]. Thereby, state-of-the-art models focus on

ICMI '24, November 4–8, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11…$15.00
https://doi.org/10.1145/3678957.3685759

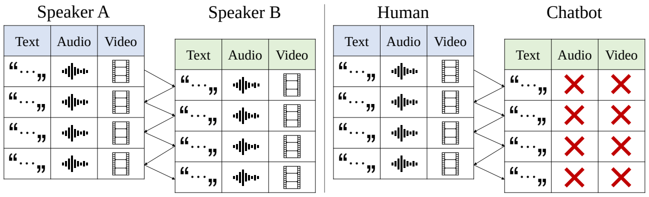


Figure 1: Comparison of Interaction Paradigms. In this work, we focus on text-based chatbot interactions (right), unlike previous research that assumed all modalities (left).

capturing conversational dynamics over time by modeling inter-speaker dependencies and exploiting advanced modality fusion strategies [16, 39, 42].

However, the current landscape of ERC models is primarily rooted in datasets sourced from actors or meticulously controlled lab environments entailing predominantly categorical emotion labels from external annotators. In contrast, real-world human-chatbot interactions follow a different conversational paradigm where emotions are not confined to scripted laboratory settings or acted facial and verbal expressions and where only the chatbot's text is available (see Figure 1). Thus, it is not clear whether previous works extend to the intricacies of real-world scenarios comprising conversational agents, which demands a re-evaluation of the challenges surrounding human-chatbot ERC to put recent advancements in the field into perspective.

In this work, we unveil the complexity of human-chatbot ERC in the wild by applying existing ERC architectures on a custom multimodal human-chatbot ERC dataset collected from 99 participants over three weeks. Our dataset consists of over 350 hours of human-chatbot interactions, capturing 8,003 self-reported emotion labels assessing the user's emotions and the perceived chatbot's emotions at regular intervals during the conversation. Using transformer-based embeddings for text, audio, and video, and speaker-specific state modeling, we build an end-to-end ERC pipeline that infers the emotional states of the user and the chatbot in terms of valence, arousal, and dominance (VAD) on three classes (low, medium, high).

Our experiments reveal a domain gap between human-human ERC and human-chatbot ERC that is mainly rooted in the subjective nature of the emotion labels, the reduced activation and expressivity of the face, and the subtlety of emotions in such settings. By incorporating user-specific personalization, our model's performance improved significantly by up to 41%, highlighting the effectiveness and potential of personalization for real-world applications.

This work is the first to consider ERC in the wild for real-world chatbots in a systematic way. Our results provide evidence for an inherent domain gap between human-human and human-chatbot interactions, questioning the applicability of existing human ERC models for this task and revealing a new research direction that calls for more nuanced models to focus on the complexities of human-chatbot ERC.

Our findings underscore the importance of refining user-centered computing for real-world applications, emphasizing the necessity for continuous research. Efforts to bridge the gap between human emotions and machine understanding are essential to ensure that future models can effectively address the varied and subjective emotional experiences of users, as demonstrated by our exploration of ERC in human-chatbot interactions.

1.1 Contributions

Our contributions are threefold:

- By collecting human-chatbot interaction data from 99 participants in the wild, we show that there is a domain gap between human-human ERC and human-chatbot ERC attributed to the subjectivity of user emotions, reduced facial activity, and inherent subtlety of emotions.
- We demonstrate the potential of personalization by increasing the model performance by up to 41% through user-specific calibration and fine-tuning.
- Our work is the first that systematically investigates ERC for real-world conversational agents, shedding light on open challenges in the field.

2 RELATED WORK

2.1 Emotion Recognition

Emotions can be recognized from a variety of modalities. In text, emotion is revealed through specific wordings and higher-level semantics of the content [7, 60]. Thereby, transformer-based encoder networks such as BERT [19] have proven to be an effective way of extracting semantically rich latent representations that can be used to infer the emotional loading of the text [39]. When dealing with audio data, vocal cues such as pitch, power, and chroma features were found to be indicative of human emotion [61]. Thereby, the audio signal is often encoded as a spectrogram from which audio features are extracted. For video data, facial expressions are often used to assess emotion [35]. Using convolutional neural networks, facial features representative of the current facial expression are extracted hierarchically, capturing local and global features for recognizing the current emotion.

In conversational settings, the emotional state can change over time, requiring emotion recognition models to handle sequences of features to capture the conversational dynamics. To this end, various datasets were collected and corresponding recognition models were proposed for tackling emotion recognition in conversations comprising different emotion taxonomies and assessment methods.

2.2 Emotion Taxonomy & Assessment

Emotion taxonomies employ dimensional or categorical models for describing emotions. The valence-arousal-dominance (VAD) model categorizes emotions based on three primary dimensions: valence (positive or negative emotion), arousal (intensity or energy level of the emotion), and dominance (perceived control or power of an emotion) [52]. This dimensional approach allows for a fine-grained representation of emotions and is typically assessed using pictorial schemes such as the Self-Assessment Manikin (SAM) [11]. Thereby, the emotional state is rated on a 9-point Likert scale by selecting the most appropriate visual representation from a set of graphical manikins for each VAD dimension.

On the other hand, categorical taxonomies describe emotions as a set of basic emotions, often in terms of Ekman's set of six basic emotions: anger, disgust, fear, happiness, sadness, and surprise [21, 22]. During assessment, the felt emotions are typically selected from a predefined list of emotions, optionally including an intensity scale per emotion. Although considered more intuitive than dimensional emotion models, basic emotions often only entail binary labels for each emotion and fall short of distinguishing positive emotions, comprising a coarse-grained and imbalanced representation of emotions. In our work, we focus on VAD given its fine-grained representation of emotions.

2.3 Datasets

Various ERC datasets have been proposed. DailyDialog [40] contains text-based multi-turn dialogues and corresponding basic emotion labels based on everyday human-human conversations. MELD (Multimodal EmotionLines Dataset) [48] contains a comprehensive collection of text, audio, and video data, encompassing the six basic emotions from multi-speaker movie dialogues. Further, IEMOCAP (Interactive Emotional Dyadic Motion Capture) [13] consists of text, audio, and video data from scripted and improvised human dialogues with both basic emotion labels and VAD labels.

Since these datasets are mainly based on scripted or acted dialogues from movie snippets or collected in laboratory settings, they might not accurately represent the nuances and variability in real-world conversations. Furthermore, the annotations were mainly obtained through multiple external raters, which was found to be easier to predict than intrinsic emotions [36, 43] and neglects discrepancies between the felt emotion of the speaker and the expressed emotion as perceived by the raters [12]. Finally, it is not clear whether the emotion dynamics of human-human conversation contained in most datasets extend to human-chatbot conversations. Given that there are structural differences in the way humans interact with other humans compared to interacting with virtual agents [17, 51], the same might apply to expressing emotions. We address these limitations by collecting authentic, unscripted conversations between humans and chatbots through multiple modalities in the wild.

2.4 Models

Both unimodal and multimodal models exist for ERC. Typically, unimodal ERC models entail text-based architectures while multimodal architectures complement the text modality by adding audio and video data.

2.4.1 Text-based Models. Text-based ERC models entail a wide variety of approaches, ranging from rule-based approaches comprising grammatical and logical rules to learning-based methods exploiting recent advancements in deep learning [4]. Chatterjee et al. [15] investigated emotion recognition from text-based human-chatbot interactions. They found that recurrent neural networks achieve decent performance in recognizing discrete emotion classes from latent text representations based on the past three conversational turns. Ghosal et al. [25] found that explicitly modeling both intra- and inter-speaker dependencies can further increase prediction accuracy. They proposed graph convolutional networks (GCN) to capture conversational dynamics over time. The approach was further refined later on, reaching state-of-the-art performance on different benchmarks [63]. Alternatively to GCNs, Li et al. [39] used hierarchical transformers to imbue the model with context- and speaker sensitivity. Using BERT [19] for utterance-level features and an additional high-level transformer for capturing global dynamics, they outperformed various state-of-the-art models on various datasets including MELD and IEMOCAP. To tackle data scarcity in ERC, Hazarika et al. [28] investigated transfer learning using pre-trained multi-turn dialogue models for emotion classifiers of conversations, achieving additional robustness.

2.4.2 Multimodal Models. Multimodal models operate on features extracted from multiple modalities that are then fused using different strategies. For example, Siriwardhana et al. [58] used pre-trained encoders (RoBERTa [41] for text, Wav2vec [56] for audio, and FAb-Net [64] for video). They combined the encoded modalities using pairwise inter-modality attention to obtain fused feature representations. Chudasama et al. [16] followed a similar approach but refined the feature extraction modules using a novel triplet network. In contrast, Xing et al. [65] fused inter-modality features using convolutional neural networks. In addition, multimodal methods also address context- and speaker sensitivity over time. For example, Majumder et al. [42] used bidirectional gated recurrent units (Bi-GRU) to model each speaker's state and the global state separately. Xing et al. [65] extended this idea by proposing an adapted dynamic memory network (A-DMN) for effectively fusing intra- and inter-speaker dependencies. In addition, speaker-specific embeddings are often added to the feature representations, which supports the model in distinguishing between speakers.

While these models work well for recognizing discrete emotions from human-human interactions, it is not clear how they perform for human-chatbot conversations where the audio and video modalities for the chatbot are missing (see Figure 1). Furthermore, the proposed methods rely on extracting conversational dynamics inherent to human-human conversations whereas the interaction dynamics between a human and a chatbot can differ [17, 51]. In this work, for predicting chatbot emotions we rely only on modalities available in human-chatbot conversations (i.e., text for the chatbot, and text, audio, and video for the user) for predicting user emotions and perceived chatbot emotions. We show that there is a substantial domain gap between the two tasks due to the inherent discrepancies in the interaction paradigm and we demonstrate how user-specific personalization can mitigate this problem.

3 DATA COLLECTION

We collected text, audio, and webcam data from 99 participants in the wild while they interacted with a GPT-3-based chatbot over three weeks. During the interactions with the chatbot, the participants filled in 8,003 self-reports indicating their own emotional state and the perceived emotional state of the chatbot in terms of valence, arousal, and dominance. The experiment was approved by the ethics board of ETH Zurich (application 2022-N-65).

3.1 Participants

We recruited 108 English-speaking participants (56 female, 52 male) between the ages 18 and 52 (mean = 25.1 years, standard deviation SD = 4.6 years) via our university's recruiting platform. The participants were required to actively engage in interactions with the chatbot on at least 10 different days over three weeks (average of 11 days, SD = 2 days). We incentivized participation through gamification similar to previous works [38, 59, 62]. The participants were compensated based on the number of completed self-reports (CHF 60 for at least 24 self-reports, or CHF 110 for at least 48 self-reports). Further, one participant was awarded CHF 1,000 in a lottery draw at the end of the study. The chance of winning could be increased by reaching performance-related levels (bronze = 30 self-reports, silver = 80, gold = 150, platinum = 250), based on which lottery tickets were awarded (bronze = 1 ticket, silver = 5, gold = 10, platinum = 20).

3.2 Apparatus

We used a web-based data collection framework from Kovačević et al. [38] consisting of a dashboard conveying participation statistics and a chat page for speech-based chatbot interactions (see Appendix B for screenshots). The framework offers three GPT-based chatbots with different genders, occupations, hobbies, origins, and emotional states. During interactions, the participants' text, audio, and video input were recorded. To ensure valid video data, a face detection model (SSD MobileNet V1 from `face-api.js`) periodically assessed the visibility of the face in selected frames. For further implementation details of the framework, we refer to the original paper by Kovačević et al. [38].

3.3 Procedure

Upon first login, participants filled out a pre-study questionnaire about their chatbot experience (48% reported having experience) and consented to data recording. They then underwent a tutorial on web page usage, starting with a hardware check for microphone and webcam functionality, followed by a sample conversation to introduce the system, with the option to revisit the tutorial anytime. To ensure conversational variety, participants interacted with different chatbot personae in different types of conversations (see Appendix B for details). Face detection was used every two seconds to validate the video input, alerting the user after three failed attempts and pausing the conversation. The camera could also be disabled manually anytime, which paused the conversation. Self-reports assessing the participants' and chatbots' emotional states were prompted in regular intervals (see Appendix B for details). Using the SAM [11], valence, arousal, and dominance were assessed on a 9-point Likert scale, with an option for flagging neutral emotions on a binary scale for later calibration (see Section 3.4). A daily limit of 10 conversations per participant was set to prevent misuse. Participants could opt out from the survey anytime. The study concluded with a demographic and feedback questionnaire.

3.4 Data Preprocessing

In total, we collected 2,734 conversations and 9,292 self-reports from 108 participants. To clean the dataset, we defined the following exclusion criteria: (1) constant VAD ratings, (2) too quick self-report completion (less than 10 seconds), and (3) functional issues (e.g., aborted conversation due to no detected face). After cleaning, 1,725 conversations and 8,003 self-reports (86% of the initial self-reports) from 99 participants remained. To align participants' subjective interpretation of the VAD scale, we standardized the ratings of each VAD dimension per participant using the mean VAD ratings of neutrally flagged emotions as a center and the standard deviation for separating the scale into three levels (low, neutral, high) with class boundaries at ±1 SD (see Appendix A for details).

Table 1: Average ratings for valence, arousal, and dominance per speaker (user or chatbot) and chatbot persona (Sarah, Vincent, Albert) in the range $[1, 9]$. The standard deviation is given in brackets.

	Dimension	All	Sarah	Vincent	Albert
User	Valence	5.53 (1.55)	5.49 (1.58)	5.46 (1.50)	5.62 (1.55)
	Arousal	4.49 (1.98)	4.55 (1.97)	4.40 (1.92)	4.50 (2.03)
	Dominance	5.45 (1.60)	5.42 (1.66)	5.42 (1.54)	5.50 (1.59)
Chatbot	Valence	5.48 (1.73)	5.52 (1.78)	5.35 (1.70)	5.53 (1.69)
	Arousal	4.88 (1.97)	5.06 (1.97)	4.79 (1.95)	4.78 (1.96)
	Dominance	5.19 (1.59)	5.31 (1.66)	5.07 (1.53)	5.16 (1.55)

3.5 Data Validation

The participants engaged for 4 hours and 18 minutes on our web page on average (SD = 2 hours 48 minutes). They were most active around noon and after 6 p.m., which coincides with common working hours and leisure time. The selection of chatbots was balanced (37.1% *Sarah*, 35.8% *Albert*, and 27.1% *Vincent*). A conversation lasted on average 12.7 minutes including self-reports (SD = 10.2 minutes) with 4.7 self-reports on average (SD = 2.6 self-reports). Despite regular interruptions for self-reports, over 80% of the participants indicated in the post-study questionnaire that they perceived the interruptions to have had little or no effect on the conversational flow.

Most conversations were rated as neutral or slightly positive (valence between 5 and 7). The same applies to dominance, though more conversations are concentrated on the neutral level (level 5), possibly stemming from participants struggling with assessing dominance correctly. For arousal, we observe higher variability in the ratings not coinciding with the other two dimensions (see Appendix A for details).

There was no noticeable difference in mean VAD ratings across speakers and chatbot personae (see Table 1). We found significant positive correlations between users' and chatbots' valence ($r = 0.59, p \ll 0.01$) and arousal ($r = 0.59, p \ll 0.01$), suggesting that users' emotions often coincide with the perceived chatbot emotions. For dominance, we found a significant negative correlation ($r = -0.18, p \ll 0.01$), which indicates that often either the user felt in control of the situation while perceiving the chatbot as being controlled or vice versa.

4 METHOD

We built two transformer-based multimodal emotion classification networks for predicting self-reported user and chatbot emotions in terms of valence, arousal, and dominance on three classes (low, medium high) as depicted in Figure 2. Following the human-chatbot interaction paradigm (see Figure1), the models should neglect audio and video for the chatbot side, focusing on efficiency during both preprocessing and inference to facilitate real-time interactive applications. This entails that the model can be deployed and fine-tuned on consumer hardware.

Table 2: Split statistics for the train-validation-test split.

Split	Users	Conversations	Utterances	Self-Reports
Train	79	1,364	25,478	6,383
Validation	10	166	3,305	755
Test	10	185	3,568	865
Total	99	1,725	32,350	8,003

4.1 Feature Extraction

In line with previous work in ERC, we used pre-trained encoder networks to embed the input modalities into meaningful latent feature representations. Such networks have the advantage of being trained on large-scale datasets in an unsupervised fashion to construct salient semantic representations, which allows for transfer learning to various downstream tasks such as emotion recognition [28].

4.1.1 Text Features. We used RoBERTa [41], a robust and optimized version of BERT [19], to encode the text data. RoBERTa can embed text of up to 512 tokens into a 768-dimensional vector space. Since punctuation characters can be of semantic value to the sentence, we separated such characters from other words using whitespaces to avoid removal by the tokenizer. Each utterance was tokenized and embedded separately. The token limit was never reached for any of the utterances.

4.1.2 Audio Features. We used DistilHuBERT [14], a distilled version of the HuBERT [29] model for audio feature encoding that uses convolutional blocks and multiple transformer encoder layers to construct latent audio features from audio files of variable length. It operates on windows of 20 milliseconds from audio signals sampled at 16 kHz, yielding a sequence of embeddings per utterance whereby the length of the sequence is proportional to the length of the audio signal. Since the audio samples were recorded using different participant microphones, we re-sampled all audio files to 16 kHz.

4.1.3 Video Features. We used image-based feature extraction on frames from the video stream. To this end, we used an EfficientNet [55] that was pre-trained on the AffectNet dataset [46], a large-scale collection of facial expressions and corresponding emotion labels. EfficientNet consists of multiple convolutional blocks applied on an RGB image of size 226-by-226 and produces a one-dimensional feature embedding of size 1,280. To cope with different resolutions and lighting conditions, we first cropped the face region using a bounding box obtained from a lightweight face detection model (SSD MobileNet v1 from `face-api.js`). We then resized the frames to 226-by-226 pixels and normalized them using the mean and standard deviation from the AffectNet database. While there are face detection modules that can run in real-time (e.g., MTCNN from `face-api.js`), they are less accurate than bigger models. To obey our efficiency constraints, we processed six frames per second because an empirical experiment showed that the described feature extraction pipeline for the video features can take up to 0.16 seconds per frame.

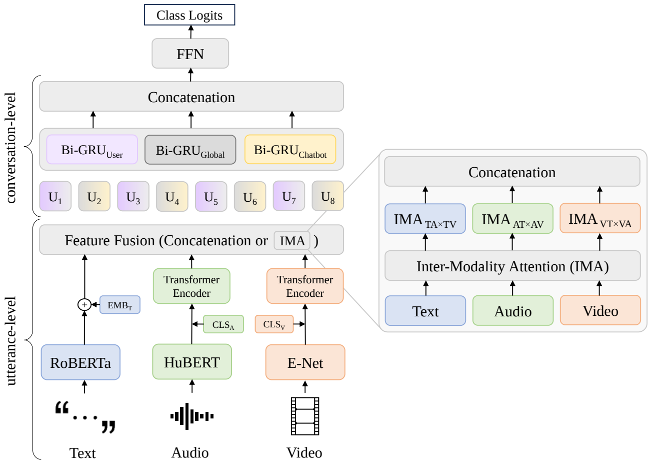


Figure 2: Overview of the multimodal network architecture inspired by Majumder et al. [42] and Siriwardhana et al. [58]. The unimodal model is obtained by ignoring two input branches. Note that in the unimodal case for audio and video, *Bi-GRU$_{User}$* and *Bi-GRU$_{Chatbot}$* are not used.

4.2 Unimodal Model

We built a separate unimodal model for each modality. By using the feature extraction modules described in Section 4.1 and a context window of W conversational turns (i.e., $2W$ utterances), the resulting embeddings are of size $[2W, 768]$ for text, $[2W, S_A, 768]$ for audio, and $[2W, S_V, 1280]$ for video, where $S_A = \{S^i_A, i \in [1 \dots 2W]\}$ and $S_V = \{S^i_V, i \in [1 \dots 2W]\}$ denote the sequence lengths of the embedded audio and video data of each utterance, respectively. For audio, the sequence length corresponds to the number of 20ms windows. For video, the sequence length corresponds to the number of extracted frames. To unify the dimensions of the embeddings per utterance, we reduced the audio and video embeddings along the second dimension to $[2W, 768]$ and $[2W, 1280]$ respectively by using a transformer encoder and a prepended modality-specific *CLS* token to be used as a global representation of that sequence. Furthermore, we added embeddings of speaker-specific identifiers (0 for the user, 1 for the chatbot) to the encoded utterances to support the model in distinguishing between speakers in the text branch.

While a big context window W is favorable for capturing long-range conversational dynamics, we fixed the context window at $W = 4$ (i.e., 8 utterances) given that the number of self-reports following a context window of size W drops significantly for $W > 4$ (see Appendix A for details). Inspired by DialogueRNN [42], we modeled the speaker states and the global state in the text modality separately using three bidirectional gated recurrent units (Bi-GRUs) in parallel to capture intra-speaker and inter-speaker dynamics explicitly. Thereby, we feed the entire context window to the global Bi-GRU while the speaker-specific Bi-GRUs are fed the corresponding speaker’s utterances only (see Figure 2). For the audio and video modalities, we used a single Bi-GRU because there is no audio and video data from the chatbot. We call the former model *MultiGRU* and the latter model *SingleGRU*. The output from the Bi-GRUs is fed through a classification head consisting of a feed-forward network.

Table 3: Ablation study of the prediction performance in terms of the macro F_1 score for valence, arousal, dominance, and the mean (VAD). The highest values for the unimodal and multimodal models are highlighted in bold.

		Modalities			Macro F_1 (User Labels)				Macro F_1 (Chatbot Labels)			
Model		Text	Audio	Video	VAD	Valence	Arousal	Dominance	VAD	Valence	Arousal	Dominance
	Baseline	–	–	–	24.00	20.40	23.64	27.96	22.18	21.44	17.34	27.76
Unimodal	MultiGRU	✓	–	–	**36.68**	**46.54**	**31.29**	32.21	**41.86**	**52.36**	**36.41**	**36.79**
	SingleGRU	–	✓	–	29.62	28.67	27.95	**32.25**	–	–	–	–
	SingleGRU	–	–	✓	28.41	29.38	26.25	29.60	–	–	–	–
Multimodal	Multimodal-Concat	✓	✓	–	**41.72**	**49.13**	38.26	**37.79**	**41.86**	**52.42**	**36.91**	**36.26**
	Multimodal-IMA	✓	✓	–	40.91	48.66	**39.10**	34.98	40.87	51.35	36.26	35.01
	Multimodal-Concat	✓	–	✓	39.80	47.90	35.03	36.47	40.66	51.60	35.14	35.23
	Multimodal-IMA	✓	–	✓	37.00	46.36	32.24	32.39	40.25	49.74	36.54	34.47
	Multimodal-Concat	✓	✓	✓	39.85	47.05	36.21	36.29	40.28	51.01	35.90	33.94
	Multimodal-IMA	✓	✓	✓	38.76	45.91	33.82	36.55	39.53	49.52	35.52	33.55

4.3 Multimodal Model

The multimodal models use all available modalities simultaneously. We explored two fusion strategies for combining the modality-specific features. The first model (*Multimodal-Concat*) concatenates the modality-specific features of the unimodal models. The second model (*Multimodal-IMA*) fuses the features using inter-modality attention (IMA) similar to Siriwardhana et al. [58]. Thereby, each pair of modalities X and Y is fed through a self-attention block where modality X is used as the query vector, and modality Y is used as the key and value vectors, enhancing the features from modality X with information from modality Y, denoted as $M_{XY} = IMA(X, Y)$. Next, the Hadamard product (i.e., component-wise multiplication) is applied to all pairs that share the same modality used for the query (i.e., $M_X = M_{XY} \otimes M_{XZ}$). Finally, the outputs of each IMA block are concatenated.

5 RESULTS

We split our dataset into training (80%), validation (10%), and test (10%) sets by enforcing disjoint but proportionate sets of users and comparable VAD class distributions across splits (see Table 2). More details on the label distribution can be found in Appendix A.

All models were trained on a consumer GPU with 24 GB VRAM (Nvidia RTX 3090). We evaluated our model on the macro F_1 score that weighs the F_1 scores of each class equally.

5.1 Unimodal Models

The performance of the unimodal models is listed in the top part of Table 3. As a naïve baseline, we used a model that always predicts the majority class. All unimodal models outperformed the baseline model. The performance difference was highest for the text modality (+12.61 macro F_1 for user emotions, +19.68 macro F_1 for chatbot emotions). Predicting valence showed the highest performance both for user emotions and chatbot emotions. In contrast, the audio and video modalities provided only marginal improvements over the baseline. However, the improvement was higher for user emotions than for chatbot emotions (up to +5.62 macro F_1 for the user and up to +3.64 macro F_1 for the chatbot).

5.2 Multimodal Models

For training the multimodal models, we initialized the text, audio, and video branches with the best weights from the corresponding unimodal models. The performance metrics for the multimodal models can be found in the bottom part of Table 3.

Multimodal-Concat. For predicting the user emotions, separately adding audio (up to +5.06 macro F_1) and video (up to +3.12 macro F_1) to text improved the performance across all dimensions. The combination of text and audio resulted in the highest performance (41.72 macro F_1). However, combining all three modalities did not provide any performance gain. On the contrary, adding the video modality slightly degraded the performance. For predicting the perceived chatbot emotions, adding video or audio did not substantially improve performance. The best-performing model based on text and audio achieves similar performance as its text-based counterpart (41.86 macro F_1). Adding video further degrades performance (−2.85 macro F_1 for dominance).

Multimodal-IMA. The IMA-based model shows very similar results to *Multimodal-Concat*. Again, text was most indicative, followed by audio, while video did not increase performance. However, the IMA-based models surpassed the simpler *Multimodal-Concat* model only slightly in terms of arousal for user emotions (+0.84 macro F_1) and performed slightly worse on average (up to −2.8 macro F_1 for text and audio).

Contrary to the unimodal setting, the best-performing multimodal models performed almost identically across different speaker labels, which stems from the lack of available data for the chatbot side.

5.3 Action Unit Analysis

To investigate the low performance of the video modality, we compare the facial activations in our dataset with facial activations in the MELD dataset [49] and the IEMOCAP dataset [13]. Since facial action units are strongly linked to human emotions [23, 66], we extracted action unit activations using the OpenFace toolkit and analyzed the difference in activation between consecutive frames to measure variability in the facial expressions (see Figure 3). Since all three datasets are conversational, there is no discrepancy for speaking-related action units (i.e., units 25 and 26). However, we

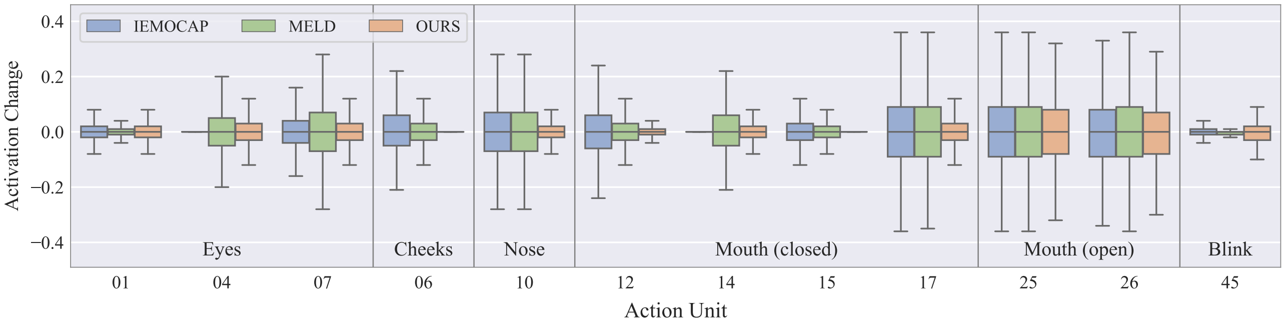


Figure 3: Variability of action unit activation across datasets obtained from computing the distribution of pairwise activation changes between consecutive frames over all utterances, grouped by activation units and datasets.

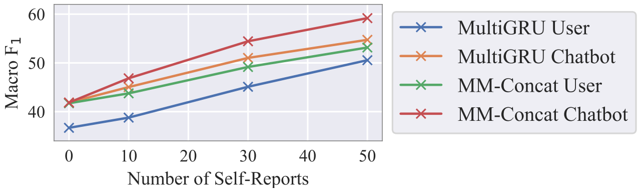


Figure 4: The performance gain in terms of macro F_1 VAD when personalizing the best unimodal models (MultiGRU using text) and the best multimodal models (Multimodal-Concat using text and audio) for different numbers of self-reports.

observe that the distribution is much more diverse for the acted datasets from MELD [49] and IEMOCAP [13], especially in the cheeks, nose, and closed mouth region. While there is more variation in our dataset in terms of blinking, this can be attributed to the tiring effect of the computer screen during the interaction.

5.4 Personalization

We leveraged user-specific personalization to further improve model performance. To this end, we fine-tuned the best unimodal and multimodal models for each user in the test set separately by using a portion of the user's data for fine-tuning. For a user with N self-reports, we fine-tuned each model on n self-reports of a user and evaluated on the remaining $N - n$ self-reports. Figure 4 shows the performance increase for $n = [0, 10, 30, 50]$. We notice a substantial linear performance increase of up to 41%.

5.5 Runtime

With a context window of four utterances, it takes 0.01s (±0.001s SD) for the text branch, 0.16s (±1.45s SD) for the audio branch, 0.34s (±1.07s SD) for the video branch, and 0.01s (±0.003s SD) for the Bi-GRU and the classification head. This sums up to 0.52s (±1.83s SD) to process on an Nvidia RTX 3090 without parallelizing the branches. Considering that it takes several seconds to speak each utterance, our models are suitable for real-time ERC. They can help a conversational agent generate the next utterance based on emotion predictions or monitor the user's emotional state over time.

6 DISCUSSION

6.1 Feature Fusion

Although IMA performed well in previous work [16, 58], our IMA-based models were inferior to our simpler *Multimodal-Concat* model. Given the human-chatbot interaction paradigm, the missing audio and video modalities for the chatbot make it challenging for the chatbot emotion recognition models to extract inter-speaker conversational dynamics. Instead, the model focuses on the text modality, as shown by the performance discrepancy between the text-based model and other unimodal models (see Table 3). While in the IMA blocks the saliency of the textual features is altered by the audio and video modalities when applying the Hadamard product, the *Multimodal-Concat* model can attain salient textual features more easily. Hence, given the low predictive value of audio and video, the IMA blocks fail to effectively cross-combine the features. We conclude that IMA can suffer from missing modalities which should be taken into account in the future.

6.2 Chatbot and User Emotions

There is a noticeable performance discrepancy between user emotion and chatbot emotion recognition for the text-based model. User emotions can be more difficult to predict because they are intrinsic and not always conveyed externally (i.e., the emotion information is not always included in the modalities) [36, 43]. On the other hand, the chatbot emotion labels are collected from external annotators (i.e., the users) rating the same entity (i.e., the chatbot) based on the same amount of information that is seen by the prediction model (i.e., text only), leading to a better generalizable model. Thus, the true subjective emotions are more difficult to predict compared to externally obtained labels, which aligns with previous findings [36, 43]. This remains true for the multimodal models as well since the models predicting user emotion labels are consistently inferior to the unimodal text-based chatbot emotion predictor despite using all three modalities, highlighting the severe impact of subjective labels on the model performance.

6.3 Modality Ablation

The text modality was most indicative of human emotion, followed by audio, which is in line with previous results [16, 58]. Furthermore, we observe that the audio and video modalities were more predictive of user emotion than of the perceived chatbot emotion, which follows from the lack of available audio and video data for the chatbot. Nevertheless, the performance for predicting chatbot emotions is above random, which can be attributed to correlations between the user and chatbot emotion labels. In contrast, text data is available for both the user and chatbot, allowing the model to focus on the specific speaker and include data from the interlocutor as context, which works well for both the user labels and the chatbot labels.

6.4 Action Unit Analysis

The video modality's low performance can be explained with the low variability in facial activation during interaction, which could stem from participants using facial expressions less frequently to convey their emotion compared to human-human interaction since they assume it to have no impact on the conversation. For example, the action units 6, 12, 14, and 15 are generally associated with happiness, sadness, disgust, and contempt [66] but were notably less active in our setting. As a consequence, the video feature extractor would need to capture micro-expressions indicative of emotions, which would require disentangling the emotion information from the spoken content and the individual style of the person, an aspect that could be solved with user-specific personalization.

6.5 Personalization

User-specific personalization showed to be effective for robust and consistent performance boosts of up to 41% independently of the model or the emotion labels used. Given self-report intervals of 90 seconds, personalizing the models would require between 15 minutes (10 self-reports) and less than 90 minutes (50 self-reports). The required time could be further reduced by employing more efficient ways of collecting the self-reports that can increase both the quality and the amount of collected data [50], which could result in even higher performance gains and higher willingness of users to engage in a personalization phase.

6.6 Implications and Potential Applications

Our findings indicate that current ERC architectures might not be fully effective for numerous practical applications as the training data fails to capture the complexities of human-chatbot interactions. Nonetheless, this observed domain gap can be mitigated by user-specific personalization, presenting new research opportunities and implications for virtual conversational agents. Specifically, recognizing the perceived chatbot emotions can enable developers to predict and tailor how specific user groups will interact with these agents, facilitating pre-release testing and adjustments.

Consumer products. In educational gaming, personalized ERC can be useful for AI-powered non-player characters (NPC) [57]. NPCs can modify their actions based on the emotions developers wish to trigger. In addition, negative emotions can denote dissatisfaction from which the system can learn and evolve, ultimately enhancing user experience [24]. In e-learning systems, accurate emotion prediction can track students' emotional states, which can increase the learning gain [8]. In the realm of conversational agents, such a system could be powered using AI-based digital characters that act as personal teachers catering to the student's specific needs, for example by calling for a break when detecting frustration.

Health care. In elderly homes, companionship and social interactions are integral to the mental well-being of the residents. Chatbot-driven digital characters can be useful as cheap and always available companions. By personalizing ERC models to residents, the chatbots could use the recognized emotion for tailoring an adequate response, cheering residents up when they feel sad, engaging them in interesting conversations when they are bored or feel lonely [2], or ultimately, trigger an intervention by a professional if strong negative feelings are detected [34].

Ethical Considerations. Despite numerous applications, ERC models that can reliably assess the user's reaction to the next utterance bear the risk of manipulating user emotions, which can negatively influence the user's mental health and well-being [33]. Ensuring informed consent, transparency, privacy, and data security is paramount to mitigating these risks.

6.7 Limitations and Future Work

While our data collection comprises university students, the conversational dynamics between a chatbot and people from different social strata could differ. In the future, we will investigate if our method extends to users from different cohorts. Furthermore, the performance gain from personalization could also stem from discrepancies between the training data of encoder models and human-chatbot conversation styles, which warrants further analysis, e.g., through specialized encoder models. Moreover, while the performance gain is substantial, the willingness of end users to engage in a personalization phase requires testing. For applications where a user-specific model is infeasible, user clustering [53] or partial personalization [45] can be used. Alternatively, a warm-up phase dedicated to collecting predictive labels for boosting personalization could cut the time required for reaching good performance [10].

7 CONCLUSION

In this work, we investigated multimodal emotion recognition for human-chatbot interactions in the wild. Based on a collected dataset comprising text, audio, and video data from 99 participants interacting with GPT-based chatbots over three weeks, we implemented different unimodal and multimodal emotion recognition models that predict the user emotion and the perceived chatbot emotion in terms of valence, arousal, and dominance on three classes (low, medium, high). Our results revealed a domain gap between human-human ERC and human-chatbot ERC rooted in the subjective nature of the labels, the low activation of the face during interaction, and the subtlety of emotions in human-chatbot interaction. Through user-specific personalization, we could improve the performance by up to 38% (user emotions) and 41% (perceived chatbot emotions), which outlines the potential of personalization for real-world applications. Moreover, the performance gap between user and chatbot emotion recognition models indicates that emotions are easier to predict from the reader's than the speaker's perspective, highlighting the challenges in recognizing user emotions in real contexts. By exploring the complexity of human-chatbot ERC and identifying a domain gap to human-human ERC, our work is a key step towards equipping future conversational agents with enhanced emotion sensitivity, a trait that is crucial for such models to transcend to diverse real-world scenarios.

ACKNOWLEDGMENTS

The authors thank all the participants for their time and effort. This work was supported by an ETH Zurich Research Grant under Grant No.: ETH-10 22-1.

REFERENCES

[1] Rania Abdelghani, Yen-Hsiang Wang, Xingdi Yuan, Tong Wang, Pauline Lucas, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2023. GPT-3-Driven Pedagogical Agents to Train Children's Curious Question-Asking Skills. *International Journal of Artificial Intelligence in Education* (jun 2023). https://doi.org/10.1007/s40593-023-00340-7

[2] Hojjat Abdollahi, Mohammad H. Mahoor, Rohola Zandie, Jarid Siewierski, and Sara H. Qualls. 2023. Artificial Emotional Intelligence in Socially Assistive Robots for Older Adults: A Pilot Study. *IEEE Transactions on Affective Computing* 14, 3 (July 2023), 2020–2032. https://doi.org/10.1109/taffc.2022.3143803

[3] Utku Günay Acer, Marc van den Broeck, Chulhong Min, Mallesham Dasari, and Fahim Kawsar. 2022. The City as a Personal Assistant. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (jul 2022), 1–31. https://doi.org/10.1145/3534573

[4] Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports* 2, 7 (may 2020). https://doi.org/10.1002/eng2.12189

[5] Israel Edem Agbehadji and Abosede Ijabadeniyi. 2020. Approach to Sentiment Analysis and Business Communication on Social Media. In *Bio-inspired Algorithms for Data Streaming and Visualization, Big Data Management, and Fog Computing*. Springer Singapore, 169–193. https://doi.org/10.1007/978-981-15-6695-0_9

[6] Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. 2023. A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications* 17 (feb 2023), 200171. https://doi.org/10.1016/j.iswa.2022.200171

[7] Saima Aman and Stan Szpakowicz. [n. d.]. Identifying Expressions of Emotion in Text. In *Text, Speech and Dialogue*. Springer Berlin Heidelberg, 196–205. https://doi.org/10.1007/978-3-540-74628-7_27

[8] Kiavash Bahreini, Rob Nadolski, and Wim Westera. 2014. Towards multimodal emotion recognition in e-learning environments. *Interactive Learning Environments* 24, 3 (May 2014), 590–605. https://doi.org/10.1080/10494820.2014.908927

[9] David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *SSRN Electronic Journal* (2023), 22 pages. https://doi.org/10.2139/ssrn.4337484

[10] Nikola Banovic and John Krumm. 2018. Warming Up to Cold Start Personalization. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (jan 2018), 1–13. https://doi.org/10.1145/3161175

[11] Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (mar 1994), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

[12] Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, 578–585. https://aclanthology.org/E17-2092

[13] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42, 4 (nov 2008), 335–359. https://doi.org/10.1007/s10579-008-9076-6

[14] Heng-Jui Chang, Shu wen Yang, and Hung yi Lee. 2022. Distilhubert: Speech Representation Learning by Layer-Wise Distillation of Hidden-Unit Bert. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. https://doi.org/10.1109/icassp43922.2022.9747490

[15] Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics. https://doi.org/10.18653/v1/s19-2005

[16] Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE. https://doi.org/10.1109/cvprw56347.2022.00511

[17] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3290605.3300705

[18] Bertrand David, Rene Chalon, Bingxue Zhang, and Chuantao Yin. 2019. Design of a Collaborative Learning Environment integrating Emotions and Virtual Assistants (Chatbots). In *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE. https://doi.org/10.1109/cscwd.2019.8791893

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/ARXIV.1810.04805

[20] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1l73iRqKm

[21] Paul Ekman. 1992. Are there basic emotions? *Psychological Review* 99, 3 (1992), 550–553. https://doi.org/10.1037/0033-295x.99.3.550

[22] Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion* 6, 3-4 (may 1992), 169–200. https://doi.org/10.1080/02699939208411068

[23] Paul Ekman and Wallace V. Friesen. 1978. Facial Action Coding System. https://doi.org/10.1037/t27734-000

[24] Jamie Fraser, Ioannis Papaioannou, and Oliver Lemon. 2018. Spoken Conversational AI in Video Games. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. ACM. https://doi.org/10.1145/3267851.3267896

[25] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/d19-1015

[26] Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2017. Towards Designing Cooperative and Social Conversational Agents for Customer Service.. In *ICIS*. 1–13.

[27] Muhammad Anas Hasnul, Nor Azlina Ab. Aziz, Salem Alelyani, Mohamed Mohana, and Azlan Abd. Aziz. 2021. Electrocardiogram-Based Emotion Recognition Systems and Their Applications in Healthcare—A Review. *Sensors* 21, 15 (jul 2021), 5015. https://doi.org/10.3390/s21155015

[28] Devamanyu Hazarika, Soujanya Poria, Roger Zimmermann, and Rada Mihalcea. 2021. Conversational transfer learning for emotion recognition. *Information Fusion* 65 (jan 2021), 1–12. https://doi.org/10.1016/j.inffus.2020.06.005

[29] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460. https://doi.org/10.1109/taslp.2021.3122291

[30] Maryam Imani and Gholam Ali Montazer. 2019. A survey of emotion recognition methods with emphasis on E-Learning environments. *Journal of Network and Computer Applications* 147 (dec 2019), 102423. https://doi.org/10.1016/j.jnca.2019.102423

[31] Jacob Israelashvili and Agneta Fischer. 2022. Recognition of Emotion from Verbal and Nonverbal Expressions and Its Relation to Effective Communication: A Preliminary Evidence of a Positive Link. *Journal of Intelligence* 11, 1 (dec 2022), 6. https://doi.org/10.3390/jintelligence11010006

[32] Carroll E. Izard. 2009. Emotion Theory and Research: Highlights, Unanswered Questions, and Emerging Issues. *Annual Review of Psychology* 60, 1 (jan 2009), 1–25. https://doi.org/10.1146/annurev.psych.60.110707.163539

[33] Amelia Katirai. 2023. Ethical considerations in emotion recognition technologies: a review of the literature. *AI and Ethics* (June 2023), 1–22. https://doi.org/10.1007/s43681-023-00307-3

[34] Salik Khanal, Arsénio Reis, João Barroso, and Vitor Filipe. 2018. *Using Emotion Recognition in Intelligent Interface Design for Elderly Care*. Springer International Publishing, 240–247. https://doi.org/10.1007/978-3-319-77712-2_23

[35] Byoung Ko. 2018. A Brief Review of Facial Emotion Recognition Based on Visual Information. *Sensors* 18, 2 (jan 2018), 401. https://doi.org/10.3390/s18020401

[36] Kazunori Komatani, Ryu Takeda, and Shogo Okada. 2023. Analyzing Differences in Subjective Annotations by Participants and Third-party Annotators in Multimodal Dialogue Corpus. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Prague, Czechia, 104–113. https://doi.org/10.18653/v1/2023.sigdial-1.9

[37] Diane M. Korngiebel and Sean D. Mooney. 2021. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *npj Digital Medicine* 4, 1 (jun 2021). https://doi.org/10.1038/s41746-021-00464-x

[38] Nikola Kovačević, Christian Holz, Markus Gross, and Rafael Wampfler. 2024. The Personality Dimensions GPT-3 Expresses During Human-Chatbot Interactions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 61 (may 2024), 36 pages. https://doi.org/10.1145/3659626

[39] Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020. HiTrans: A Transformer-Based Context- and Speaker-Sensitive Model for Emotion Detection in Conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.370

[40] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Greg Kondrak and Taro Watanabe (Eds.). Asian Federation of Natural Language Processing, Taipei, Taiwan, 986–995. https://aclanthology.org/I17-1099

[41] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/ARXIV.1907.11692

[42] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (jul 2019), 6818–6825. https://doi.org/10.1609/aaai.v33i01.33016818

[43] Lucien Maman, Gualtiero Volpe, and Giovanna Varni. 2022. Training Computational Models of Group Processes without Groundtruth: the Self- vs External Assessment's Dilemma. In *Companion Publication of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22 Companion)*. Association for Computing Machinery, New York, NY, USA, 18–23. https://doi.org/10.1145/3536220.3563687

[44] Samuel Mascarenhas, Manuel Guimaraes, Rui Prada, Joao Dias, Pedro A. Santos, Kam Star, Ben Hirsh, Ellis Spice, and Rob Kommeren. 2018. A Virtual Agent Toolkit for Serious Games Developers. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE. https://doi.org/10.1109/cig.2018.8490399

[45] Lakmal Meegahapola, William Droz, Peter Kun, Amalia de Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, George Gaskell, Altangerel Chagnaa, Amarsanaa Ganbold, Tsolmon Zundui, Carlo Caprini, Daniele Miorandi, Alethia Hume, Jose Luis Zarza, Luca Cernuzzi, Ivano Bison, Marcelo Rodas Britez, Matteo Busso, Ronald Chenu-Abente, Can Günel, Fausto Giunchiglia, Laura Schelenz, and Daniel Gatica-Perez. 2022. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (dec 2022), 1–32. https://doi.org/10.1145/3569483

[46] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. 2019. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing* 10, 1 (jan 2019), 18–31. https://doi.org/10.1109/taffc.2017.2740923

[47] Kyo-Joong Oh, Dongkun Lee, Byungsoo Ko, and Ho-Jin Choi. 2017. A Chatbot for Psychiatric Counseling in Mental Healthcare Service Based on Emotional Dialogue Analysis and Sentence Generation. In *2017 18th IEEE International Conference on Mobile Data Management (MDM)*. IEEE. https://doi.org/10.1109/mdm.2017.64

[48] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1050

[49] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. *IEEE Access* 7 (2019), 100943–100953. https://doi.org/10.1109/access.2019.2929050

[50] M. Prajwal, Ayush Raj, Sougata Sen, Snehanshu Saha, and Surjya Ghosh. 2023. Towards Efficient Emotion Self-report Collection Using Human-AI Collaboration. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 2 (jun 2023), 1–23. https://doi.org/10.1145/3596269

[51] Amon Rapp, Lorenzo Curti, and Arianna Boldi. 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies* 151 (jul 2021), 102630. https://doi.org/10.1016/j.ijhcs.2021.102630

[52] James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality* 11, 3 (sep 1977), 273–294. https://doi.org/10.1016/0092-6566(77)90037-x

[53] Koustuv Saha, Ted Grover, Stephen M. Mattingly, Vedant Das swain, Pranshu Gupta, Gonzalo J. Martinez, Pablo Robles-Granda, Gloria Mark, Aaron Striegel, and Munmun De Choudhury. 2021. Person-Centered Predictions of Psychological Constructs with Social Media Contextualized by Multimodal Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (mar 2021), 1–32. https://doi.org/10.1145/3448117

[54] Malik Sallam. 2023. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* 11, 6 (mar 2023), 887. https://doi.org/10.3390/healthcare11060887

[55] Andrey V. Savchenko. 2022. Video-Based Frame-Level Facial Analysis of Affective Behavior on Mobile Devices Using EfficientNets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2359–2366. https://arxiv.org/abs/2103.17107

[56] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised Pre-training for Speech Recognition. https://doi.org/10.48550/ARXIV.1904.05862

[57] Claudia Schrader, Julia Brich, Julian Frommel, Valentin Riemer, and Katja Rogers. 2017. *Rising to the Challenge: An Emotion-Driven Approach Toward Adaptive Serious Games*. Springer International Publishing, 3–28. https://doi.org/10.1007/978-3-319-51645-5_1

[58] Shamane Siriwardhana, Tharindu Kaluarachchi, Mark Billinghurst, and Suranga Nanayakkara. 2020. Multimodal Emotion Recognition With Transformer-Based Self Supervised Feature Fusion. *IEEE Access* 8 (2020), 176274–176285. https://doi.org/10.1109/access.2020.3026823

[59] Mirjam Stieger, Marcia Nißen, Dominik Rüegger, Tobias Kowatsch, Christoph Flückiger, and Mathias Allemand. 2018. PEACH, a smartphone- and conversational agent-based coaching intervention for intentional personality change: study protocol of a randomized, wait-list controlled trial. *BMC Psychology* 6, 1 (sep 2018). https://doi.org/10.1186/s40359-018-0257-9

[60] Carlo Strapparava and Rada Mihalcea. 2010. Annotating and Identifying Emotions in Text. In *Intelligent Information Access*. Springer Berlin Heidelberg, 21–38. https://doi.org/10.1007/978-3-642-14000-6_2

[61] Monorama Swain, Aurobinda Routray, and P. Kabisatpathy. 2018. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology* 21, 1 (jan 2018), 93–120. https://doi.org/10.1007/s10772-018-9491-z

[62] Rafael Wampfler, Severin Klingler, Barbara Solenthaler, Victor R. Schinazi, Markus Gross, and Christian Holz. 2022. Affective State Prediction from Smartphone Touch and Sensor Data in the Wild. In *SIGCHI Conference on Human Factors in Computing Systems* (New Orleans, Louisiana) *(CHI '22)*. ACM, New York, NY, USA.

[63] Binqiang Wang, Gang Dong, Yaqian Zhao, Rengang Li, Qichun Cao, Kekun Hu, and Dongdong Jiang. 2023. Hierarchically stacked graph convolution for emotion recognition in conversation. *Knowledge-Based Systems* 263 (mar 2023), 110285. https://doi.org/10.1016/j.knosys.2023.110285

[64] Olivia Wiles, A. Sophia Koepke, and Andrew Zisserman. 2018. Self-supervised learning of a facial attribute embedding from video. https://doi.org/10.48550/ARXIV.1808.06882

[65] Songlong Xing, Sijie Mai, and Haifeng Hu. 2022. Adapted Dynamic Memory Network for Emotion Recognition in Conversation. *IEEE Transactions on Affective Computing* 13, 3 (jul 2022), 1426–1439. https://doi.org/10.1109/taffc.2020.3005660

[66] U. Zarins. 2019. *Anatomy of Facial Expression*. Anatomy Next, Incorporated. 136–196 pages. https://books.google.ch/books?id=8UV5zQEACAAJ

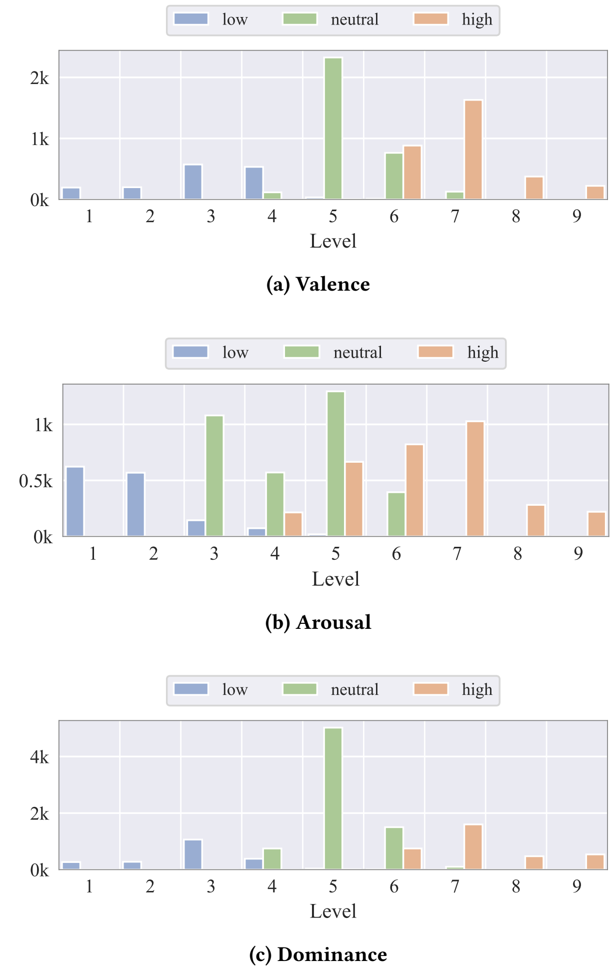
[67] Sebastian Zepf, Javier Hernandez, Alexander Schmitt, Wolfgang Minker, and Rosalind W. Picard. 2020. Driver Emotion Recognition for Intelligent Vehicles. *Comput. Surveys* 53, 3 (jul 2020), 1–30. https://doi.org/10.1145/3388790

[68] Zijian Zhou, Muhammad Adeel Asghar, Daniyal Nazir, Kamran Siddique, Mohammad Shorfuzzaman, and Raja Majid Mehmood. 2022. An AI-empowered affect recognition model for healthcare and emotional well-being using physiological signals. *Cluster Computing* 26, 2 (nov 2022), 1253–1266. https://doi.org/10.1007/s10586-022-03705-0

APPENDICES

A EMOTION LABEL DISTRIBUTION

Contrary to the naturally polarized dimensions of valence (positive vs. negative) and dominance (being controlled vs. being in control), the scale for arousal is not polarized and bears the risk of participants subjectively interpreting the scale differently. To mitigate this problem, we normalized the collected emotion self-reports per participant, allowing for a direct comparison of labels across participants. For each participant and each VAD dimension, we calculate the mean and standard deviation over all self-reports where "neutral emotion" was selected. We then standardize each self-report, centering the per-user self-reports around the level that each user subjectively perceived as neutral, naturally dividing the scale into three classes (low, neutral, high) with class boundaries at ±1 SD.



(a) Valence

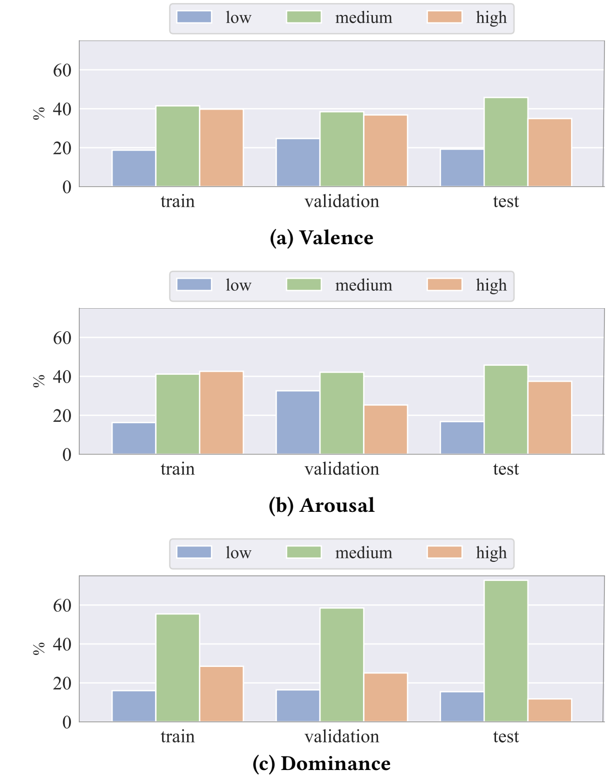
(b) Arousal

(c) Dominance

Figure A.1: Unnormalized label distribution for (a) valence, (b) arousal, and (c) dominance on a 9-point Likert scale. colored based on their class correspondence after user-specific normalization using the self-reported neutral emotion as the center for standardization.

Table A.1: The number and proportion of self-reports N that are preceded by at least W conversational turns.

Window Size W	1	2	3	4	5	6	7	8
# Self-Reports N	8,003	8,003	8,003	7,923	7,258	6,938	6,514	5,754
Proportion	100%	100%	100%	99%	90.7%	86.7%	81.4%	71.9%



(a) Valence

(b) Arousal

(c) Dominance

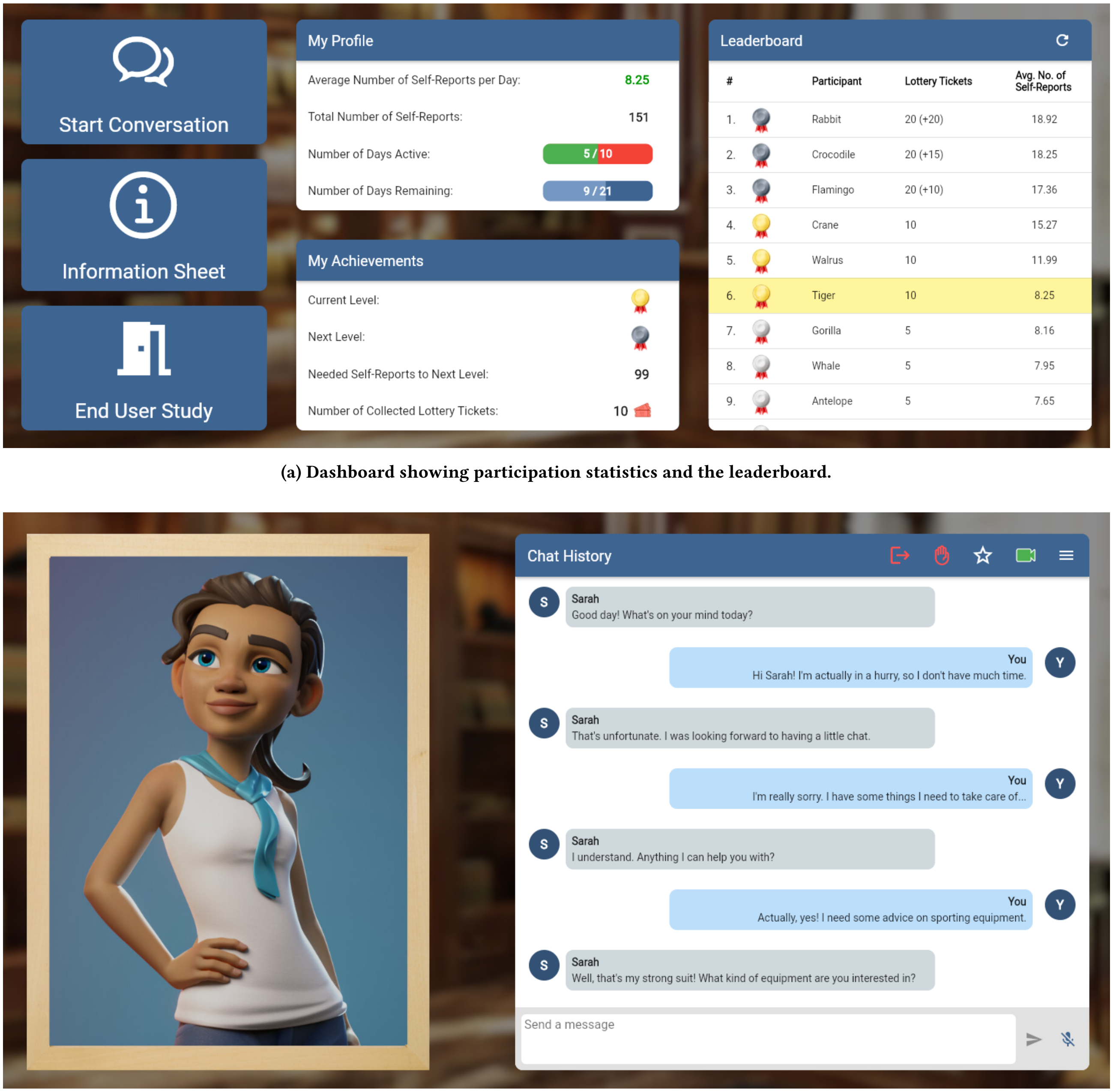
Figure A.2: The class distributions for the training, validation, and test splits for (a) valence, (b) arousal, and (c) dominance.

Figure A.1 shows the distribution of the VAD ratings on the unnormalized 9-point Likert scale, colored by their class correspondence after user-centered label normalization. As can be seen, the neutral class for valence was indeed skewed towards 6. However, for most users, 5 (the middle level) was considered neutral, which aligns with the definition of the scale. The same applies to dominance. However, for arousal, we notice that participants indeed interpreted the scale differently, with neutral levels ranging from 3 to 6, indicating that the intensity of an emotion bears high subjectivity, which highlights the importance of label normalization for the labels to become comparable.

The class distribution across train, validation, and test splits after user-specific label normalization is depicted in Figure A.2.

B DATA COLLECTION FRAMEWORK

The data collection framework is based on the implementation from Kovačević et al. [38]. It was implemented with Flutter and consisted of a dashboard conveying participation statistics and a chat page for chatbot interactions. Screenshots of the framework



My Profile

Average Number of Self-Reports per Day:	8.25
Total Number of Self-Reports:	151
Number of Days Active:	5 / 10
Number of Days Remaining:	9 / 21

My Achievements

Current Level:	
Next Level:	
Needed Self-Reports to Next Level:	99
Number of Collected Lottery Tickets:	10

Leaderboard

#	Participant	Lottery Tickets	Avg. No. of Self-Reports
1.	Rabbit	20 (+20)	18.92
2.	Crocodile	20 (+15)	18.25
3.	Flamingo	20 (+10)	17.36
4.	Crane	10	15.27
5.	Walrus	10	11.99
6.	Tiger	10	8.25
7.	Gorilla	5	8.16
8.	Whale	5	7.95
9.	Antelope	5	7.65

(a) Dashboard showing participation statistics and the leaderboard.

Chat History

S Sarah
Good day! What's on your mind today?

You Y
Hi Sarah! I'm actually in a hurry, so I don't have much time.

S Sarah
That's unfortunate. I was looking forward to having a little chat.

You Y
I'm really sorry. I have some things I need to take care of...

S Sarah
I understand. Anything I can help you with?

You Y
Actually, yes! I need some advice on sporting equipment.

S Sarah
Well, that's my strong suit! What kind of equipment are you interested in?

(b) Chat page showing an ongoing conversation with Sarah that has been generated using the prompt elements from previous work [38].

Figure B.1: Screenshots from the data collection showing the dashboard (a) and the chat page (b).

are depicted in Figure B.1. Participants chose from three chatbot personae for each new conversation, whereby the same chatbot could not be selected consecutively to ensure variety. Conversations commenced in one of four ways to add diversity: 1) the chatbot suggested a topic from a predefined list [20], 2) the user was asked to suggest a topic, 3) the chatbot chose a conversation starter from the DailyDialog dataset [40] with either a random emotional tone or 4) a tone matching the chatbot's prompted emotion. Interactions occurred via speech using Google's speech-to-text for transcription, with an option for participants to review and edit transcriptions before sending. Self-reports assessing the participants' and chatbots' emotional states were prompted every 90 seconds, indicated by a blinking star on the top right, and could be deferred by up to 30 seconds. The video recording was paused during self-reports. Conversations automatically ended after either 50 turns, two minutes of inactivity, or manually via the end button.