Multimodal User Enjoyment Detection in Human-Robot Conversation: The Power of Large Language Models

Andre Pereira
KTH Royal Institute of Technology
Sweden
atap@kth.se

Lubos Marcinek
KTH Royal Institute of Technology
Sweden
lubosm@kth.se

Jura Miniota
KTH Royal Institute of Technology
Sweden
jura@kth.se

Sofia Thunberg
Linköping University
Sweden
thunberg.sofia@gmail.com

Erik Lagerstedt
School of Informatics, University of Skövde
Sweden
erik.lagerstedt@his.se

Joakim Gustafson
KTH Royal Institute of Technology
Sweden
jocke@speech.kth.se

Gabriel Skantze
KTH Royal Institute of Technology
Sweden
skantze@kth.se

Bahar Irfan
KTH Royal Institute of Technology
Sweden
birfan@kth.se

Abstract

Enjoyment is a crucial yet complex indicator of positive user experience in Human-Robot Interaction (HRI). While manual enjoyment annotation is feasible, developing reliable automatic detection methods remains a challenge. This paper investigates a multimodal approach to automatic enjoyment annotation for HRI conversations, leveraging large language models (LLMs), visual, audio, and temporal cues. Our findings demonstrate that both text-only and multimodal LLMs with carefully designed prompts can achieve performance comparable to human annotators in detecting user enjoyment. Furthermore, results reveal a stronger alignment between LLM-based annotations and user self-reports of enjoyment compared to human annotators. While multimodal supervised learning techniques did not improve all of our performance metrics, they could successfully replicate human annotators and highlighted the importance of visual and audio cues in detecting subtle shifts in enjoyment. This research demonstrates the potential of LLMs for real-time enjoyment detection, paving the way for adaptive companion robots that can dynamically enhance user experiences.

CCS Concepts

• **Computing methodologies** → **Model development and analysis**; *Supervised learning*; Natural language processing; • **Human-centered computing** → *Natural language interfaces.*

Keywords

User Enjoyment, Affect Recognition, Human-Robot Interaction, Large Language Models, Multimodal, Older Adults

ICMI '24, November 04–08, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3685729

ACM Reference Format:
Andre Pereira, Lubos Marcinek, Jura Miniota, Sofia Thunberg, Erik Lagerstedt, Joakim Gustafson, Gabriel Skantze, and Bahar Irfan. 2024. Multimodal User Enjoyment Detection in Human-Robot Conversation: The Power of Large Language Models. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 04–08, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3678957.3685729

1 Introduction

Despite the potential and proliferation of Large Language Models (LLMs) that enable conversational robots in various domains, including therapy [29], service [4], and care for older adults [20, 25], these works still exhibit some limitations that lead to unpleasant experiences, underscoring the importance of accurately detecting user enjoyment during conversations with robots. While current affect recognition systems offer a glimpse into user's enjoyment by detecting laughter and smiles [30], they fall short of capturing the multifaceted expression of enjoyment, especially in conversations.

In our prior work, we created the HRI-CUES scale [21] using manual annotation of user enjoyment in conversations with a robot that captures both overall enjoyment and its nuances at the turn (exchange) level. In this paper, we focus on methods for automated enjoyment detection in HRI by exploring the potential of LLMs to rate our scale. We investigate the following hypotheses:

- H1: Specific prompt design choices and techniques, such as self-reflection [22], enhance LLMs' ability to accurately interpret enjoyment levels from conversation content.
- H2: Multimodal LLMs, by incorporating video and audio, outperform text-only in interpreting enjoyment in conversations.
- H3: A supervised learning model that combines LLM outputs with temporal, visual, and audio cues improves predictions.
- H4: Participants' self-reported enjoyment metrics will not strongly correlate with automated enjoyment scores as indicated by our previous manual annotation study [21].

We test these hypotheses by comparing the performance of different models and prompting techniques against each other and our established ground truth, derived from multiple human annotators.

2 Background

2.1 What is Enjoyment?

A variety of theories offer different perspectives on enjoyment. One influential framework is the concept of flow [6]. This theory suggests enjoyment arises from engagement and mastery, not relaxation. Building on flow theory is the concept of true fun [40], requiring flow, playful spontaneity, and social connection. Absence of any component may still provide pleasure or some enjoyment, but not true fun. In contrast, other models emphasize enjoyment in less intense states [34], which aligns with the circumplex model of emotion [7, 39, 43]. The circumplex model positions arousal (low to high) and valence (negative to positive) as distinct axes. Some enjoyment theories [34] prioritize emotions such as contentment, even with lower arousal, while others emphasize highly aroused states like excitement. The 'happy' emotion often directly related to enjoyment reflects a state of high arousal and high positive valence. A holistic model of enjoyment should then consider the full spectrum of enjoyment experiences. Interactions like casual conversations may not be highly challenging, yet they contain aspects of both arousal and positive emotion. All of these factors become relevant when modeling enjoyment in conversations.

2.2 Assessing Enjoyment

The most common way to measure enjoyment is through self-reporting [18, 35, 36, 52]. While domain-specific measures exist, in many human-computer interaction (HCI) and HRI studies, enjoyment is a secondary focus. Specific domains, such as healthcare, measure enjoyment using questionnaires like the Quality of Life Enjoyment and Satisfaction Questionnaire (Q-LES-Q) [8] or the Physical Activity Enjoyment Scale (PACES) [24]. Enjoyment is often represented by metrics composed of multiple items [16, 27, 28, 38, 47] or more simple single self-reported items [5, 37]. However, there are limitations with self-reporting, such as participants confirming to researcher expectations [10, 19]. In addition, it is often desirable to estimate a user's feelings from an external perspective when self-reporting is not possible or when the goal is to automate behavior. Previous studies that take an external perspective on casual conversations have measured interaction quality or user satisfaction [42, 48, 50]. Systems that are developed to classify enjoyment from an external perspective use simplistic approaches that consider smiles, laughter, and frowns to indicate valence [30]. However, externally perceived signals are typically ambiguous and highly context-dependent [13, 15].

2.3 Automatic Enjoyment Recognition

With transcription becoming faster and more accurate [41], the transcription of a conversation is an important source of information about how enjoyable a conversation is. Previous work [45, 49] has shown the potential for automatic affect classification in text using machine learning or rule-based approaches. More recently, studies e.g., [31] demonstrate the value of LLMs for classifying valence in text. However, modern LLMs might not be as strong in classification tasks as fine-tuned classifier models, even though they outperform humans in explaining the reasons for their classifications [53]. Speech contains more than the words that are spoken, there is also prosody, voice quality, and other acoustic features that give meaning to the words. For instance, Sagha et al. used openSMILE [9] on four different speech corpora (including one focused on HRI) and trained a classifier to recognize the valence of utterances. That study concludes that classifiers can be improved by training them for the target age group, and their emotion recognizer correlates with emotional expressivity. Ma et al. introduces the ElderReact dataset, a video and audio dataset of elderly individuals reacting to short video clips. They show that both facial expressions and voice patterns contain important multimodal features for automatic emotion prediction using Naive Bayes, support vector machines (SVM), and XGBoost classifiers. However, they highlight limitations in the models' accuracy and the need to tailor training data specifically for the elderly population to avoid generalization issues. Despite these advances in automatic emotion recognition, particularly for older adults [33], there's a lack of systems specifically targeting enjoyment within conversations. This gap highlights the area this paper seeks to address. If user enjoyment can be detected autonomously in a conversation, it opens the possibility for interventions to improve the interaction where necessary.

3 Dataset

This work builds upon the data from our prior work on the participatory design development (two studies) of an autonomous companion robot for older adults that uses an LLM for conversations, as described in [20]. These conversations cover a wide range of topics (open-domain dialogue), from discussions on hobbies, activities, and politics to information inquiries. A robot interaction from a preliminary study with more interaction failures and two pilot interactions from a subsequent study (after robot improvement) were used to develop an enjoyment scale. Based on this scale, the annotators individually rated 25 robot interactions from the second study, with moderate to good alignment. This paper uses the user enjoyment scale, and the annotation data from 25 videos.

3.1 Robot and Setup

The Furhat robot was used with a neutral-looking face mask. The robot smiled, raised eyebrows, shifted gaze, and blinked during conversations to reinforce naturalness through non-verbal feedback based on silences in user input, albeit without context analysis. When the robot stopped listening to the user, the robot looked away until a response was generated to illuminate the thinking process. A USB microphone array (Seeed Studio) was used to obtain clear audio for Google Cloud Speech-to-Text speech recognition. GPT-3.5 (text-davinci-003, OpenAI) was used for dialogue generation, which was the most capable LLM at the time of the studies (March 2023). An empathetic persona was created by prompting the robot to ask open-ended and follow-up questions, and reflect on the conversation context. Amazon Polly was used for speech generation. A wizard interface was used to initiate interaction with the user with a pre-scripted phrase. However, the rest of the interaction was fully autonomous, based on the user's recognized speech and responses generated by the LLM. The robot ended the conversation after 7

minutes with a pre-scripted phrase to ensure user experiences were comparable. Details about the robot and setup can be found in [20].

3.2 Data Points

The data features 25 Swedish-speaking healthy older adults (12 men, 13 women) with a mean age of 74.6 (SD = 5.8) who talked individually to the robot for ∼7 minutes. Interaction duration was M = 7.4 min (SD = 1.5) with 12 to 29 turns per participant. Each turn lasted 5 to 61 seconds (M = 17.7, SD = 7.2). The total duration was 174 min, corresponding to 590 turns. The interactions were video-recorded by an external camera facing both the participant and the robot at a side angle, as well as through the robot camera to record the participant's face. Conversational turns (exchanges) were chosen as the basis of annotations, because they were mostly similar in duration for participants, as well as to facilitate the exploration of user enjoyment through algorithms that can be fed into LLMs to improve the interaction continuously. A turn ends/starts when the participant stops speaking, offering the potential for automated systems to assess and generate responses accordingly.

3.3 Annotators

While a combination of multidisciplinary backgrounds to fully understand and analyze the multimodal complexity of enjoyment, in addition to familiarity with the target population and culture, is challenging to find in a single annotator, a group of annotators can complement each other in such a task. In our prior work [21], three annotators with complementary and relevant backgrounds were selected to create a user enjoyment scale for HRI and validate it. Annotator 1, referred to as A1, specializes in user enjoyment, while A2 works on HRI with older adults and cognitive science, and A3 specializes in multi-modal HRI and cognitive science. The annotators had a mean age of 30 (SD = 2.94). They were native Swedish speakers and thoroughly familiar with Swedish culture, which was important in understanding the nuances and culture-specific idioms (e.g., 'beat around the bush') and proverbs (e.g., 'even a worm will turn') in conversations for enjoyment detection. Annotators rated enjoyment per turn, assessing both the robot's response and the participant's input, and provided an overall enjoyment assessment for the entire interaction.

4 HRI CUES Enjoyment Scale Prompting

The Human-Robot Interaction Conversational User Enjoyment Scale (HRI CUES) was tailored to assess enjoyment from an external perspective during open-domain dialogue with robots. The scale provides a structured framework for measuring user enjoyment, and its validity was established through annotations of the dataset presented in the previous section. In this section, we present the scale, the rater instructions, and half[1] of the examples created to represent the annotator's rating process. These were used (as presented) to create a prompt for LLMs that would align with the annotators. We italicized the parts that require multimodal (temporal, audio, and visual) cues for enjoyment detection, which were only included in the prompt for multimodal LLMs.

[1]The remaining examples are provided in the Supplementary Material and all examples can be viewed in this Youtube video: https://youtu.be/VmKvGM0pyec.

4.1 Scale

1 Very low enjoyment — Discomfort and/or frustration
2 Low enjoyment — Boredom or interaction failure
3 Neutral enjoyment — Politely keeping up the interaction
4 High enjoyment — Smooth and effortless interaction
5 Very high enjoyment — Immersion in the conversation and/or deeper connection with the robot

4.2 Rater Instructions

To rate the exchange higher on the user enjoyment scale (4 and 5), look for signs of enjoyment, *such as smirking, movement*, flow of conversation (the topic is moving forward), no strain or discomfort, asking questions [to the robot], smooth turn taking, *dynamic tonality (and dynamic phrasing of sentences)*, being playful, sharing personal experiences [to the robot], sharing an understanding (common ground), and anthropomorphizing [the robot].

To rate the exchange lower on the scale (1 and 2), look for signs such as low energy, *sighing*, tiredness, repeated questions [from the robot], *long breaths, restless movements (i.e., adaptors, such as moving in the chair from side to side or changing arm position), flat tonality*, silence, *awkward and negative facial expressions, flaring nostrils, disengagement cues (e.g., turning away from the robot)*, and topic closure (e.g., "Let's talk about something else").

Neutral enjoyment (3) refers to a lack of these cues, in which conversation content (and context) becomes more relevant, such as having small talk or continuing the conversation without having much interest in the topic.

In cases where the exchange has cues from multiple levels of the scale, use the dominant level in that interaction. This could be done by observing the intensity of the cues, the significance of the cues, or the interaction trajectory. On the other hand, when there are strong cues from two moderately or highly distinct levels (as opposed to subsequent levels), rate the exchange with a value in between. For instance, if the exchange contains discomfort (1) and the human is politely keeping the interaction (3), the exchange should be rated as 2, the mid-point between the two levels.

Each participant will have a different set of signals. The beginning of the interaction will determine the baseline behavior of the participant, *based on their rhythm and gestures, from which the person could deviate from during the interaction. This means that the same type of gesture (e.g., keeping one's arms crossed) can be interpreted differently between different participants. Instead put an emphasis on the change in behavior.* Separate content from context, that is, put attention on what is being said (conversation content, e.g., topic), but focus should be on the whole feeling of the exchange.

The interaction failure does not necessarily refer to a robot failure (e.g., incorrect response, speech recognition failure, turn-taking error, disengagement cue), since robot failures can lead to amusement, anthropomorphism, or empathy in the user, therefore, increasing user enjoyment. The interaction failure rather refers to the situation when either the user (e.g., interrupting the robot) or the robot made a failure that resulted in the conversation being disrupted, leading to low enjoyment in the user.

4.3 Examples

• **Participant ID: 1**, **Turn: 11**, **Exchange:** ROBOT: Oh, I just wasn't sure what you meant. Are you asking why I didn't wanna

go with you? HUMAN: No, I asked where you want to go. Do you have a favorite place?
Reasoning: Human sounds disappointed that the robot misunderstood it, repeats the previous question, *looks at the experimenter 3 times (gaze aversion), and sighs before posing their second question for clarification. Human has low energy.* These show signs of frustration. **Score:** 1

- **Participant ID:** 1, **Turn:** 8, **Exchange:** ROBOT: I am sorry, I didn't understand you. HUMAN: Okay. I will have a walk in nature reserve in..

Reasoning: Robot interrupts the human while speaking at the beginning of the turn, to which Human raises their head in a slow nod with annoyance. Human looks at the experimenter (gaze aversion), then repeats the sentence they were saying prior to being interrupted, in a rushed way. Human inhales to finish their sentence, which show signs of boredom. **Score:** 2

- **Participant ID:** 1, **Turn:** 18, **Exchange:** ROBOT: That sounds great! I love nature too. HUMAN: What kind of nature do you like?

Reasoning: Robot interrupts Human while speaking at the beginning of the exchange. Human stops talking and makes an awkward expression. Robot repeated phrasing ("That sounds X. I love Y too") for the fifth time. "I love nature too" has been repeated three times in the conversation. Human *nods, smiles, and* asks a question to move the conversation forward politely. **Score:** 3

- **Participant ID:** 1, **Turn:** 23, **Exchange:** ROBOT: Mount Everest is about 8848 meters high. HUMAN: Oh, good. Thanks. I didn't know it that exactly.

Reasoning: Interaction is smooth, there are no interruptions or failures. Human is interested in the conversation because the Robot understood the Human's question and responds correctly. *Human nods in affirmation and responds in high energy.* **Score:** 4

- **Participant ID:** 1, **Turn:** 22, **Exchange:** ROBOT: I can try. Human: How high is Mount Everest?

Reasoning: Robot interrupts the person while speaking. Human has fast turn-taking in response to the Robot's response, showing immersion in the conversation. *Human smiles, swings sideways in the chair, and leans forward, showing excitement, while asking the question.* **Score:** 5

5 User Enjoyment Detection

5.1 Human Annotator Baseline

5.1.1 Data properties and distribution. The dataset used in this analysis was not designed to provoke dis-enjoyment. On the contrary, it was designed to be the best version of what a open-domain conversational system could be at the moment. However, at the time of writing and the time of the data collection having a conversation with a robot solely by using real-time speech recognition systems and text-based LLMs still leads to interaction and conversational challenges [20]. As such, the properties of our data reflect overall successful interactions with some interaction failures and sparing moments of low enjoyment representative of human-robot conversations designed with LLMs. Figure 1 shows the distribution of the human annotator ratings data, which reveals that the majority of the exchanges were perceived to be neutral (3). During the pre-processing stage, approximately 7% of the words in our dataset, located at the ends of utterances (affecting roughly 20% of the utterances), were filtered out when compared to a perfect

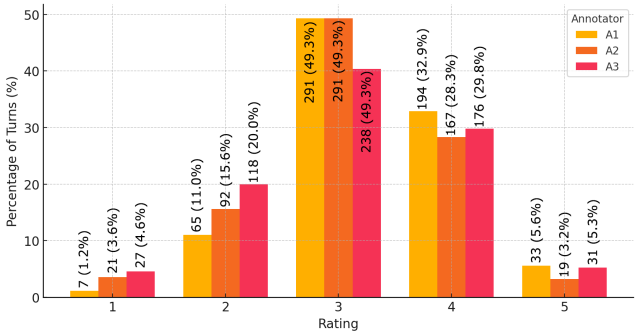

Figure 1: Distribution of ratings by annotator.

transcription. While this may introduce some noise, we believe it does not impact the study's outcomes given the comparative nature of our evaluation design. For replication purposes, we are sharing the code, dataset, and output for all models[2].

5.1.2 Performance and baseline metrics. The metrics we will use to compare our human annotators baseline with automated data will be Mean Absolute Error (MAE), Accuracy, unweighted macro F1-Score (F1) and Balanced Accuracy (BA). MAE reflects the average error across predictions. Accuracy provides a straightforward measure of overall correctness (the proportion of total predictions that were exactly right). F1 provides a balance between precision and recall, taking into account both false positives and false negatives. BA ensures each class is treated equally, important for balanced performance assessment across all rating categories. Table 1 Part 1, gives an overview of human annotator agreement, their performance and the performance of a majority baseline where each exchange is rated as the most prominent rating (3).

5.1.3 User self-reported correlations. At the end of each interaction, the participants provided 4 self-reported Likert scale (1 to 5) questionnaire items related to enjoyment:

- (Satisfaction) I was satisfied with my conversation with the robot.
- (Fun) It was fun talking to the robot.
- (Interesting) The conversation with the robot was interesting.
- (Strangeness, reversed) It felt strange talking to the robot.

The items had high reliability (Cronbach's alpha = 0.84). These self-reported scores were compared against annotators' overall enjoyment scores in two distinct manners. A simple rounded average of all the annotated exchanges for each participant and a final overall rating attributed by each annotator. Spearman correlations were used across the 4 Likert scale items. No significant correlation was found for any single annotator. There was only one moderate ($r = 0.42$) significant ($p = 0.04$) positive correlation by averaging the final overall rating from all annotators with strangeness.

5.2 Optimizing LLM Prompts for Annotation

To create an LLM based annotator we tested several prompting variations, starting each with: "Given the following scale and the current exchange between a robot and a human, rate the user enjoyment in the current exchange with an integer value (1 to 5)", referring to the scale in Sect. 4.1. Optional variations included:

(1) **(Instructions)** - Adding detailed instructions from Sect. 4.2.

[2]Dataset+Code: https://github.com/andre-pereira/ICMI2024LLMsEnjoymentDetection

Table 1: Performance metrics for annotator baselines and various model configurations. Lower MAE values are preferable, while higher values are better for other metrics. The ground truth consists of 1770 individual ratings (590 exchange annotations across 3 annotators). Table 1 lines 1-3 excluded self-ratings (1180 ratings). For each metric we consider model predictions and average results across the annotators.

Part 1 - Annotator Agreement Metrics				
Comparison	**MAE**	**Accuracy**	**F1**	**BA**
Annotator 1	0.672	0.450	0.331	0.323
Annotator 2	**0.614**	**0.469**	**0.370**	0.391
Annotator 3	0.629	0.456	0.364	**0.398**
Annotators Averaged	0.638	0.459	0.355	0.371
Majority Baseline (3s)	0.615	0.463	0.126	0.200
Part 2 - Prompt Configurations with GPT 3.5 turbo				
GPT3.5 Instructions	0.896	0.321	0.222	0.267
GPT3.5 No Instructions	0.846	0.340	0.212	0.254
GPT3.5 History	0.891	0.322	0.216	0.268
GPT3.5 No History	0.845	0.342	0.218	0.247
GPT3.5 Examples	0.861	0.333	0.215	0.261
GPT3.5 No Examples	0.895	0.322	0.220	0.260
GPT3.5 Scores	0.892	0.321	0.214	0.264
GPT3.5 No Scores	0.867	0.333	0.218	0.259
GPT3.5 Reasoning	0.905	0.317	0.226	0.280
GPT3.5 No Reasoning	0.835	0.345	0.207	0.238
GPT3.5 All features	0.933	0.292	**0.232**	**0.298**
GPT3.5 No features	**0.780**	**0.373**	0.213	0.223
Part 3 - SOTA text-only Model Performance				
GPT4 History+Reasoning	0.853	0.351	**0.276**	**0.315**
GPT4 All-Reasoning	**0.647**	**0.456**	0.242	0.247
GPT4 All features	0.730	0.407	0.269	0.280
GPT4 No features	0.651	0.434	0.242	0.244
Part 4 - Multimodal Gemini Pro 1.5 Performance				
Gemini Hist+Reas	0.786	0.373	0.283	**0.308**
Gemini All-Reas	**0.763**	**0.397**	0.243	0.254
Gemini All features	0.856	0.333	0.256	0.286
Gemini No features	0.807	0.364	0.244	0.275
Gemini Hist+Reas+2Vid	0.830	0.360	0.266	0.299
Gemini All+2Vid-Reas	0.859	0.357	0.277	0.300
Gemini All+2Vid	0.786	0.385	0.275	0.286
Gemini All+Vid	0.782	0.382	**0.294**	0.305
Gemini All+Vid-Reas	0.852	0.365	0.267	0.291
Part 5 - Multimodal Supervised Learning Performance				
XGBoost Temporal only	0.825	0.389	0.186	0.199
XGBoost Audio only	0.725	0.441	0.191	0.215
XGBoost Video only	0.758	0.435	0.186	0.212
XGBoost All Modalities	0.710	0.462	0.205	0.228
LSTM LLM only	**0.620**	**0.506**	**0.230**	**0.251**
LSTM All Modalities	0.646	0.476	0.207	0.232

(2) **(Examples)** - Adding 10 examples, 2 for each rating, to provide context and further align with annotator reasoning (Sect. 4.3)
(3) **(History)** - Adding the interaction history preceded with the text "The history of the dialog is as follows:" and followed by a newline and the text "[Exchange X]: [Robot]: ... [Person]: ..."for every previous turn in the conversation.
(4) **(Scores)** - Adding previous scores for each exchange. This was conditioned on having History.
(5) **(Reasoning)** - Require the model to reason (similar to the examples, when they are provided) prior to giving a score of enjoyment. This is akin to using self-reflection in LLMs [22], which has shown to reduce hallucination and improve performance across various tasks, such as formulating plans and strategies before generating responses.

To finalize the prompt, the following text that includes the actual exchange to be rated was added: "According to the scale, rate the following current exchange: [Robot]: ... [Person] ..." alongside a request for a particular format if reasoning was used : "Reply in the EXACT following format: [Reasoning] ... [Score] X" and a simpler "Reply with only one integer value between 1 and 5, without any additional text!" was added for the variations without reasoning.

To examine the influence of these prompt variations, we prompted GPT 3.5 Turbo version 0125 with all possible permutations in our entire dataset (590 exchanges). We used OpenAI's API with the default temperature parameters and generated 24 new sets of classifications for each exchange of the 25 participants. In total 14160 separate prompts were used in this initial investigation (see Table 1 Part 2). There was a clear trend that adding more information to the prompt made the first two central tendency metrics (MAE and Accuracy) worse and the more outlier-sensitive metrics (F1 and BA) better. This is easily perceived when looking at the simplest option (No features) that was simply the scale and the exchange vs the variation with all the features enabled where the balanced accuracy increases by 75 basis points.

Regarding correlations with user self-reported ratings, a strong positive significant ($r = 0.512, p = 0.009$) correlation was present between the model with all features and user satisfaction. The model with Examples+History ($r = 0.495, p = 0.011$), History+Reflection ($r = 0.424, p = 0.035$), No features ($r = 0.590, p = 0.002$), Instructions+History ($r = 0.561, p = 0.004$), and Instructions+History+Scores ($r = 0.531, p = 0.006$) also significantly correlated with the same item. Regarding the other self-reported items, there was only one model (Examples+History+Results) that moderately correlated with the fun ($r = 0.43, p = 0.012$) and interesting items ($r = 0.410, p = 0.042$). Curiously, none of the models correlated with the remaining item (strangeness), the only human annotators correlated with.

5.3 State-of-the-art LLM-based Annotator

Based on the results of the previous section, we selected 4 configurations for further testing: the top performing All features and No features, an intermediate model with History+Reflection and a model with all features except reflection. We use these configurations with the latest model from OpenAI at the time of writing (gpt-4-turbo-2024-04-09) to generate predictions for these 4 prompting configurations (see results in Table 1 Part 3).

Similar to GPT3.5, GPT4 models without reasoning outperform in MAE and Accuracy. The model with all features except reasoning achieved a similar level of performance to the human annotators, even surpassing some of the annotators within these metrics. The top performing configuration in the other two metrics was the model that only contained History and Reasoning. This configuration excluded the more detailed instructions and examples established by the annotators. This resulted in a poorer performance

in MAE and Accuracy but a good 'intuition' for aligning within the less common scale levels in the data (i.e., 1 and 5), as exhibited in the BA score. The more balanced configuration was achieved by the model that includes all features with MAE and Accuracy scores that surpassed all GPT3.5 models and a BA and F1 score that approached the best model.

Regarding correlations with the user self-reported perceptions, the only GPT4 model that showed any was All Features. It showed the strongest correlations from all tested models and configurations. It displayed a strong positive significant correlation with user satisfaction ($r = 0.682, p < 0.001$), a moderate positive significant correlation with perceived fun ($r = 0.444, p = 0.026$), a moderate positive **not** significant correlation with the interesting item ($r = 0.365, p = 0.073$), and a moderate positive significant correlation with perceived strangeness ($r = 0.473, p = 0.017$).

5.4 Multimodal LLM-based Annotator

While the previous subsections focused on the textual content of the human-robot interactions, we will now focus on using a multimodal LLM (Google's Gemini 1.5 Pro) to include videos (with audio) within the prompt. Due to its long context length, and joint training on text, image, audio, and video, Gemini can do multimodal reasoning [11, 12], making it suitable for enjoyment detection with the multimodal cues that annotators used for the scale. Similar to GPT4, we evaluated the model with the same four text configurations, with additional five multimodal configurations (Table 1 Part 4).

We complement our multimodal test configurations in two ways: (1) providing two videos (the frontal/robot and side cameras) similarly to the data that the human annotators had available when annotating with all multimodal cues in the prompt, and (2) with configurations that would be available if implemented on an autonomous system that solely use the frontal camera view from the robot's perspective. The same leading configuration of having all features without Reasoning excelled in MAE and Accuracy, whereas the text model with only History and Reasoning was on par with the video models on BA, slightly exceeding them. However, the most balanced configuration with a leading F1 score was the one with all features and the frontal video only. The side camera appears to add little benefit. The reason we included both cameras for the original annotators was to give a better assessment of posture. However, by inspecting the reasoning of both models throughout the 590 exchanges, All Features + frontal video still had 32 mentions of posture compared to 36 mentions in the All features + both videos configuration. This alongside the superior performance in most metrics indicates that multiple videos is not impactful in our setup.

Also, similarly to the previous models, some Gemini Pro 1.5 configurations achieved significant correlations with self-reported user perceptions. The History+Reasoning model ($r = 0.471, p = 0.017$), All Features text model ($r = 0.484, p = 0.014$), All Features excluding reasoning ($r = 0.455, p = 0.022$) and the multimodal All Features model that excluded reasoning ($r = 0.420, p = 0.037$) all had a moderate significant positive correlation with user satisfaction. The Gemini based text models also exhibited moderate positive correlations with reversed perceived strangeness from both the History+Reasoning model ($r = 0.466, p = 0.019$) and All Features model that excluded reasoning ($r = 0.420, p = 0.036$).

Fig. 2 shows the average per-class accuracy of the best GPT4 model and the same configuration in Gemini only in textual form and the multimodal version. It appears that the multimodal models

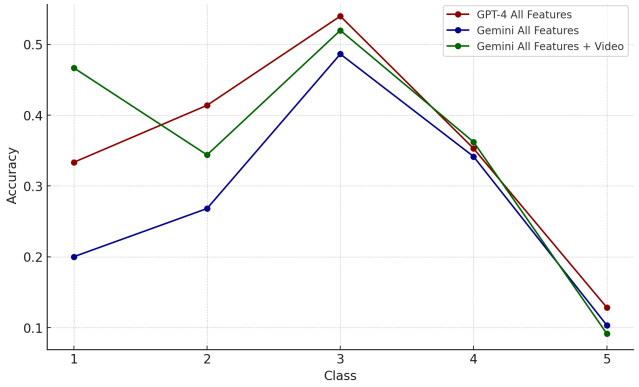

Figure 2: Average Per-Class Accuracies of Different Models

constantly outperform the text-only configurations in detecting very low engagement (Class 1). Incorporating multimodal cues can improve the evaluation of perceived enjoyment, bringing the models' reasoning capabilities similar to human understanding. For instance, in one of the exchanges in the dataset, the robot gave a disengagement cue, "I understand", which brought the conversation flow to a halt. In return, the participant looked confused, and waited 16 seconds before replying with hesitation, "Have we finished talking then, right? Can we continue?". GPT4 with All Features model provided the following reasoning for this excerpt, which indicates that it aligns well with the instructions, examples, history, and past results: *The person's question suggests a possible ambivalence about the continuation of the conversation, indicating either a loss of interest or seeking clarification on the status of the interaction. This could imply a dip in engagement or smoothness of the conversation but does not necessarily denote strong discomfort or frustration. The person is still participating and inquiring about the continuation, which keeps the exchange active. Therefore, the interaction seems to hover around a neutral enjoyment level due to the lack of clear positive engagement cues and the presence of slight uncertainty. [Score] 3.* However, without the temporal, audio, or visual cues that display that there was a clear interaction failure, GPT4 was oblivious to the extent of the dis-enjoyment in the conversation. On the other hand, multimodal models take that a step further and also reason about the extra modality that is even more relevant in detecting more extreme emotional reactions. The same exchange rated by the Gemini model with All Features and videos was: *The user appears to exhibit a level of boredom or disengagement with the conversation. Their question implies a lack of enthusiasm and a desire to move on from the current topic. The user's facial expression and tone of voice further support this assessment, as they seem neutral and unexcited. [Score] 2.* In this case, the three human annotators rated the exchange with a 3, a 2, and a 1.

5.5 Multimodal Supervised Annotator

Finally, we further attempt to enhance the results from our best text-based LLM (GPT4 with all features, given the strong correlation with self-reported user ratings) with two supervised learning techniques, namely categorical ensemble XGBoost classifiers and a LSTM regressor. This allows the comparison of integrating temporal, video, and audio modalities directly into the prompt with using solely a text-based model and complementing it with relevant multimodal features.

5.5.1 Video Features. The Py-feat library [3] was used to analyze the 10 frames-per-second videos from the robot's perspective.

- **Geometric Features:** Face Position (FaceRectX, FaceRectY) and head Orientation (pitch, roll, yaw)
- **FACS and emotion features:** 20 facial Action Coding System (FACS) Action Units and 7 facial expressions (anger, disgust, fear, happiness, sadness, surprise, neutrality)

To analyze variability and trends in these features, we computed the following statistical measures for each dialog exchange: minimum & maximum values, first & third quartiles, median, mean, standard deviation, and interquartile range (IQR). This resulted in 256 video features being considered per exchange.

5.5.2 Audio Based Features. We extracted 92 audio features by employing the hierarchical wavelet-based method described in [46] and the openSMILE toolkit (using the eGeMAPSv02 feature set) [9]:

- **Spectral Features:** Features representing characteristics of the speech signal's frequency spectrum (e.g., MFCCs).
- **Prosodic Features:** Features describing intonation, rhythm, and loudness of speech (e.g., pitch, loudness variations, F0).
- **Voice Quality Features:** Features capturing nuances in voice quality (e.g., jitter, shimmer).
- **Other Relevant Features:** Examples include statistics of formants, zero-crossing rate, and voicing-related features.

5.5.3 Temporal Features. These features help us understand the distribution of conversational time, rhythm, as well as potential hesitation or delays from interaction failures within the interaction. From manually corrected logs of the interaction, we calculated:

- **Turn Length-Exchange:** Total exchange duration.
- **Length-1stPause:** Pause duration before a turn.
- **Length-Robot:** Robot's turn duration.
- **Length-2ndPause:** Pause duration before participant's response.
- **Length-Person:** Person's turn duration.
- **4 Percentage features:** The above 4 durations normalized by the exchange length.

5.5.4 Model Training. To classify enjoyment levels, we started by employing a categorical ensemble XGBoost classifier [2]. We utilized leave-one-group-out cross-validation (LOGO CV) with participant-based splits for robust evaluation. Missing values were imputed with zeros, and features were scaled before training. Hyperparameters were tuned to control model complexity and overfitting, and early stopping was implemented to prevent overtraining and ensure generalization. Notably, we investigated class weights and tested optimizing for f1-score to address potential imbalances in enjoyment categories but found they had minimal impact. This observation, along with the desire to give stronger penalties to mispredictions at the extremes of the enjoyment scale, motivated us to explore a sequential model with a regression approach using mean squared error (MSE) as the loss function.

For enjoyment regression, we trained a sequential LSTM regressor [17]. We utilized the same LOGO CV with participant-based splits. The enjoyment labels were reshaped to match the LSTM's expected sequence format. A masking layer was employed to handle padded turn values. The LSTM architecture included dropout layers for regularization and multiple stacked LSTM layers for learning complex temporal patterns. To avoid overfitting, we used regularization, early stopping, and a TimeDistributed output layer for regression at each timestep. We used a custom masked MSE loss function that focused on valid timesteps.

For each of these two models, we created an additional feature for the annotator (integer in XGboost and one-hot encoded in LSTM) so that we could use the data from our three annotators. This tripled our dataset to a total of 1770 rated exchanges and allowed our models to learn each annotator separately. To sample the results for the next section we always generated 3 sets of predictions, one per annotator and averaged each metric.

5.5.5 Results. Contrary to expectations, the multimodal supervised learning model exhibited decreased performance in some of our metrics (see Table 1 Part 5) for predicting enjoyment levels compared to models relying solely on LLM outputs. The models that included multiple modalities, while outperforming both the LLM models and the human baseline in MAE and Accuracy scores, did so at the cost of damaging their F1 and BA scores. These models do not have access to the data from the participant they are evaluating but can better learn the distributions of each annotator, whereas the LLMs are completely blind to the annotation patterns of the annotators. This explains why in our XGBoost modality analysis even the model with only temporal features was able to outperform many LLMs in MAE and Accuracy scores. Nevertheless, we observed incremental improvements in detecting extreme levels of enjoyment with the integration of either audio or visual cues over temporal features and the integration of all + LLMs lead to improved performance in our XGBoost classifier. The sequential LSTM models demonstrated the highest accuracy and lowest MAE from all tested models, indicating a successful replication of human annotators' behaviors. From all these models, we only found one correlation with self-reported user perceptions. The LSTM with all modalities strongly correlates with strangeness ($r = 0.516, p = 0.008$), the only user rating that was correlated with the annotators.

6 Discussion

We formulated four hypotheses to guide our research:

Hypothesis 1: Specific prompting techniques enhance text-only LLMs' ability to interpret enjoyment levels. Our results support this hypothesis. After testing several models and assessing how they were affected by each different configuration, we counterintuitively found that including detailed instructions, examples, dialog history, previous results, and reasoning made the mean absolute error increase and accuracy decrease. The data shows that the exchanges are more frequently rated as neutral, which is also the majority baseline in our data. Adding more alignment features, especially in the stronger large language models, contributed to a better classification in the non-neutral levels, as reflected in the improved F1 and BA scores. Nevertheless, within these metrics the improvement was still modest compared to human annotators. This suggests that text-only prompts are still not able to fully capture the nuanced expressions of enjoyment that multimodal cues offer.

Hypothesis 2: Multimodal LLMs outperform text-only LLMs in interpreting human-robot interactions. This hypothesis was supported by (1) the Gemini multimodal model achieving the highest F1 score of all automated models, (2) by comparing the reasoning texts between models as exemplified in the detailed

example in Sect. 5.4, and (3) enhanced performance of that model in predicting lower enjoyment classes (see Fig. 2). The integration of multimodal cues directly into a prompt enhanced the model's ability to detect enjoyment more broadly, demonstrating the critical role of non-verbal communication in understanding human affective states. The accuracy improvement in detecting low enjoyment (Class 1) could be a crucial element for a conversational robot, since detecting dis-enjoyment in real-time could enable employing possible adaptive strategies to improve the conversation, such as changing the topic, injecting more humor, or proactively ending the interaction. The capability of analyzing facial expressions, tone of voice, and posture are incorporated into Gemini's reasoning and appear to be the cause of this over-performance. The fact that using only the camera from the robot's perspective provided similar performance to including a side camera also exemplifies that these kind of models can be used in real-time settings.

Hypothesis 3: Combining LLM outputs with temporal, visual, and audio cues in a supervised model improves prediction performance. Our data does not fully support this hypothesis. Similar to the multimodal LLMs, we found unbalanced class improvements when adding other modalities. Audio and visual features were more impactful than temporal features to distinguish between enjoyment classes. While training a sequential model to replicate annotators' enjoyment recognition patterns appears straightforward, we did not prioritize this analysis further. This decision aligns with observations in affective computing research, where it is argued that improvements in affect recognition accuracy alone are insufficient [14]. While most supervised models outperform human baselines and LLMs, they learn annotator-specific distributions at the expense of the LLM's broader predictive capabilities. This is evident in the sequential model trained solely on LSTM input from the same annotator: it excels in accuracy and MAE by learning that annotator's scale usage. Despite efforts to address imbalances and overfitting, models adapt to individual annotators rather than learning the nuances of enjoyment itself. Though LSTMs seemed promising for capturing enjoyment's dynamic nature, our data suggests that a unified multimodal model might be better for a goal beyond mimicking annotator distribution.

Hypothesis 4: Participants' self-reported enjoyment metrics will weakly correlate with automated enjoyment scores. We expected weak correlations between automated enjoyment predictions and self-reported ratings, similar to our human baseline. Surprisingly, this hypothesis was not supported. Automated scores correlated more strongly with self-reported enjoyment than human ratings did. This suggests that our LLM-based models, with their vast conversational data, may recognize enjoyment patterns that human annotators miss. Some text-only LLM models showed strong correlations with self-reported ratings, as did several multimodal LLMs. Although the supervised models used GPT4, which strongly correlates with self-reported metrics, only 'strangeness', in the all modalities LSTM model showed a strong correlation, similar to human baselines. This suggests that these models excel at mimicking annotators but may not fully capture true user enjoyment.

6.1 Limitations and Future Research

While we wanted to focus on natural open-domain contexts, it could be beneficial to evaluate enjoyment detection performance within scenarios that try to create a more balanced dataset by stimulating more extreme positive and negative enjoyment reactions. We also did not explore finetuning LLMs or designing textual prompts with information from other modalities. We instead focused on evaluating the capabilities of zero-shot multimodal LLMs that have their own way of tackling this issue. Translating low-level signals to text is a complex problem that deserves separate careful investigation.

6.2 Ethical Considerations

LLMs pose risks due to potential biases in their zero-shot reasoning, even with neutral prompts, which can lead to epistemic harm [23]. Personalization, where the LLM adapts to individual user cues, may mitigate this. While personalization can improve affective connection and trust, it introduces risks of selection bias (leading to less diverse and potentially less enjoyable conversations [26]) and privacy concerns if data is stored externally. Our aim was to create an autonomous enjoyment detector for real-time conversational adaptation, ultimately benefiting users. However, similar technologies could be misused for persuasion [1, 51]. As LLM capabilities advance, it's crucial for the society to understand both their potential benefits (as demonstrated here) and the associated risks.

7 Conclusion

Our research highlights the potential of LLMs and their integration with multimodal cues in enhancing the challenging task of detection of enjoyment in human-robot conversational exchanges. It was clear that stronger state-of-the-art LLMs, with well-designed prompts, clearly outperformed preceding models in fully using our enjoyment scale, even exceeding the human annotator baseline at correlating with users' self-reported scores. Despite the moderate agreement between human annotators, supervised learning techniques significantly surpassed annotator performance in predicting annotator scores, as exhibited by lower accuracy and mean absolute error metrics. However, this increase in performance was often at the cost of detecting less represented and perhaps more important (dis-enjoyment) levels. This suggests that LLMs trained on large conversational datasets may implicitly learn patterns of enjoyment expression that makes them arguably more consistent and objective at using our scale than the human annotators that help design it. Our findings suggest that LLMs are suitable for designing robots that proactively respond to user enjoyment levels, making them more capable of meeting the emotional and social needs of users.

Acknowledgments

This work was supported by KTH Digital Futures (Sweden) and the Swedish Research Council project 2021-05803.

References

[1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[2] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[3] Jin Hyun Cheong, Eshin Jolly, Tiankang Xie, Sophie Byrne, Matthew Kenney, and Luke J Chang. 2023. Py-feat: Python facial expression analysis toolbox. *Affective Science* 4, 4 (2023), 781–796.

[4] Neeraj Cherakara, Finny Varghese, Sheena Shabana, Nivan Nelson, Abhiram Karukayil, Rohith Kulothungan, Mohammed Farhan, Birthe Nesset, Meriam Moujahid, Tanvi Dinkar, Verena Rieser, and Oliver Lemon. 2023. FurChat: An Embodied Conversational Agent using LLMs, Combining Open and Closed-Domain Dialogue with Facial Expressions. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SigDIAL)*.

[5] Martin D Cooney, Takayuki Kanda, Aris Alissandrakis, and Hiroshi Ishiguro. 2011. Interaction design for an enjoyable play interaction with a small humanoid robot. In *2011 11th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 112–119.

[6] Mihaly Csikszentmihalyi, Reed Larson, et al. 2014. *Flow and the foundations of positive psychology*. Vol. 10. Springer.

[7] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.

[8] Jean Endicott, John Nee, Wilma Harrison, and Richard Blumenthal. 1993. Quality of Life Enjoyment and Satisfaction Questionnaire: a new measure. *Psychopharmacology bulletin* 29, 2 (1993), 321–326.

[9] Florian Eyben and Björn Schuller. 2015. openSMILE:) The Munich open-source large-scale multimedia feature extractor. *ACM SIGMultimedia Records* 6, 4 (2015), 4–13.

[10] Luke K. Fryer and Daniel L. Dinsmore. 2020. The Promise and Pitfalls of Self-report: Development, research design and analysis issues, and multiple methods. *Frontline Learning Research* 8, 3 (Mar. 2020), 1–9. https://doi.org/10.14786/flr.v8i3.623

[11] Gemini team, Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf

[12] Gemini team, Google. 2024. Gemini: A Family of Highly Capable Multimodal Models. arXiv:https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf

[13] Jonathan Ginzburg, Ellen Breitholtz, Robin Cooper, Julian Hough, and Ye Tian. 2015. Understanding laughter. In *20th Amsterdam Colloquium*.

[14] Hatice Gunes and Nikhil Churamani. 2023. Affective Computing for Human-Robot Interaction Research: Four Critical Lessons for the Hitchhiker. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1565–1572.

[15] Markku Haakana. 2010. Laughter and smiling: Notes on co-occurrences. *Journal of Pragmatics* 42, 6 (2010), 1499–1512.

[16] Marcel Heerink, Ben Kröse, Vanessa Evers, and Bob Wielinga. 2010. Assessing acceptance of assistive social agent technology by older adults: the almere model. *International Journal of Social Robotics* 2 (2010), 361–375.

[17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[18] Guy Hoffman and Keinan Vanunu. 2013. Effects of robotic companionship on music enjoyment and agent perception. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 317–324.

[19] Bahar Irfan, James Kennedy, Séverin Lemaignan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2018. Social Psychology and Human-Robot Interaction: An Uneasy Marriage. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA). ACM, 13–20. https://doi.org/10.1145/3173386.3173389

[20] Bahar Irfan, Sanna-Mari Kuoppamäki, and Gabriel Skantze. 2023. Between reality and delusion: challenges of applying large language models to companion robots for open-domain dialogues with older adults. (2023). https://doi.org/10.21203/rs.3.rs-2884789/v1

[21] Bahar Irfan, Jura Miniota, Sofia Thunberg, Erik Lagerstedt, Sanna Kuoppamäki, Gabriel Skantze, and André Pereira. 2024. Human-Robot Interaction Conversational User Enjoyment Scale (HRI CUES). *arXiv preprint arXiv:2405.01354* (2024).

[22] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards Mitigating Hallucination in Large Language Models via Self-Reflection. *arXiv preprint arXiv:2310.06271* (2023).

[23] Anoop K., Manjary P. Gangan, Deepak P., and Lajish V. L. 2022. *Towards an Enhanced Understanding of Bias in Pre-trained Neural Language Models: A Survey with Special Emphasis on Affective Bias*. Springer Nature Singapore, 13–45. https://doi.org/10.1007/978-981-19-4453-6_2

[24] Deborah Kendzierski and Kenneth J DeCarlo. 1991. Physical activity enjoyment scale: Two validation studies. *Journal of sport & exercise psychology* 13, 1 (1991).

[25] Weslie Khoo, Long-Jing Hsu, Kyrie Jig Amon, Pranav Vijay Chakilam, Wei-Chu Chen, Zachary Kaufman, Agness Lungu, Hiroki Sato, Erin Seliger, Manasi Swaminathan, Katherine M. Tsui, David J. Crandall, and Selma Sabanović. 2023. Spill the Tea: When Robot Conversation Agents Support Well-Being for Older Adults. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 178–182. https://doi.org/10.1145/3568294.3580067

[26] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence* 6, 4 (2024), 383–392.

[27] Michinari Kono and Koichi Araake. 2022. Is it Fun?: Understanding Enjoyment in Non-Game HCI Research. *arXiv preprint arXiv:2209.02308* (2022).

[28] Songpol Kulviwat, Gordon C Bruner II, Anand Kumar, Suzanne A Nasco, and Terry Clark. 2007. Toward a unified theory of consumer acceptance technology. *Psychology & Marketing* 24, 12 (2007), 1059–1084.

[29] Yoon Kyung Lee, Yoonwon Jung, Gyuyi Kang, and Sowon Hahn. 2023. Developing Social Robots with Empathetic Non-Verbal Cues Using Large Language Models. In *2023 32nd IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*.

[30] Florian Lingenfelser, Johannes Wagner, Elisabeth André, Gary McKeown, and Will Curran. 2014. An event driven fusion approach for enjoyment recognition in real-time. In *Proceedings of the 22nd ACM international conference on Multimedia*. 377–386.

[31] Zhiwei Liu, Kailai Yang, Tianlin Zhang, Qianqian Xie, Zeping Yu, and Sophia Ananiadou. 2024. EmoLLMs: A Series of Emotional Large Language Models and Annotation Tools for Comprehensive Affective Analysis. arXiv:2401.08508 [cs.CL]

[32] Benjamin Ma, Timothy Greer, Matthew Sachs, Assal Habibi, Jonas Kaplan, and Shrikanth Narayanan. 2019. Predicting human-reported enjoyment responses in happy and sad music. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 607–613.

[33] Kaixin Ma, Xinyu Wang, Xinru Yang, Mingtong Zhang, Jeffrey M Girard, and Louis-Philippe Morency. 2019. ElderReact: A Multimodal Dataset for Recognizing Emotional Response in Aging Adults. In *2019 International Conference on Multimodal Interaction* (Suzhou, China) *(ICMI '19)*. Association for Computing Machinery, New York, NY, USA, 349–357. https://doi.org/10.1145/3340555.3353747

[34] Elisa D Mekler, Julia Ayumi Bopp, Alexandre N Tuch, and Klaus Opwis. 2014. A systematic review of quantitative studies on the enjoyment of digital entertainment games. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 927–936.

[35] Jordan Miller and Troy McDaniel. 2022. I enjoyed the chance to meet you and I will always remember you: Healthy Older Adults' Conversations with Misty the Robot. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 914–918.

[36] Spencer Ng, Ting-Han Lin, You Li, and Sarah Sebo. 2024. Role-Playing with Robot Characters: Increasing User Engagement through Narrative and Gameplay Agency. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 522–532.

[37] Shogo Nishimura, Takuya Nakamura, Wataru Sato, Masayuki Kanbara, Yuichiro Fujimoto, Hirokazu Kato, and Norihiro Hagita. 2021. Vocal Synchrony of Robots Boosts Positive Affective Empathy. *Applied Sciences* 11, 6 (Mar 2021), 2502. https://doi.org/10.3390/app11062502

[38] Joanna Piasek and Katarzyna Wieczorowska-Tobis. 2018. Acceptance and Long-Term Use of a Social Robot by Elderly Users in a Domestic Environment. In *2018 11th International Conference on Human System Interaction (HSI)*. 478–482. https://doi.org/10.1109/HSI.2018.8431348

[39] Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev Psychopathol* 17, 3 (Summer 2005), 715–734.

[40] C. Price. 2021. *The Power of Fun: How to Feel Alive Again*. Random House Publishing Group. https://books.google.se/books?id=ZgclEAAAQBAJ

[41] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356 [eess.AS]

[42] Navin Raj Prabhu, Chirag Raman, and Hayley Hung. 2021. Defining and Quantifying Conversation Quality in Spontaneous Interactions. In *Companion Publication of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20 Companion)*. Association for Computing Machinery, New York, NY, USA, 196–205. https://doi.org/10.1145/3395035.3425966

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

[44] Hesam Sagha, Jun Deng, and Björn Schuller. 2017. The effect of personality trait, age, and gender on the performance of automatic speech valence recognition. In *2017 seventh international conference on affective computing and intelligent interaction (ACII)*. IEEE, 86–91.

[45] Mostafa Al Masum Shaikh, Helmut Prendinger, and Ishizuka Mitsuru. 2007. Assessing sentiment of text by semantic dependency and contextual valence analysis. In *Affective Computing and Intelligent Interaction: Second International Conference, ACII 2007 Lisbon, Portugal, September 12-14, 2007 Proceedings 2*. Springer, 191–202.

[46] Antti Suni, Juraj Šimko, Daniel Aalto, and Martti Vainio. 2017. Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language* 45 (2017), 123–136.

[47] Viswanath Venkatesh, James YL Thong, and Xin Xu. 2012. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS quarterly* (2012), 157–178.

[48] MA Walker, DJ Litman, CA Kamm, and A Abella. 1998. Evaluating spoken dialogue agents with PARADISE: Two case studies. *Computer Speech & Language* 12, 4 (1998), 317–347. https://doi.org/10.1006/csla.1998.0110

[49] Jin Wang, Liang-Chih Yu, K Robert Lai, and Xue-jie Zhang. 2015. A locally weighted method to improve linear regression for lexical-based valence-arousal prediction. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 415–420.

[50] Wenqing Wei, Sixia Li, Shogo Okada, and Kazunori Komatani. 2021. Multimodal User Satisfaction Recognition for Non-task Oriented Dialogue Systems. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montréal, QC, Canada) *(ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 586–594. https://doi.org/10.1145/3462244.3479928

[51] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (, Seoul, Republic of Korea,) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 214–229. https://doi.org/10.1145/3531146.3533088

[52] Jiaxin Xu, Chao Zhang, Raymond H Cuijpers, and Wijnand A IJsselsteijn. 2024. Affective and Cognitive Reactions to Robot-Initiated Social Control of Health Behaviors. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 810–819.

[53] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics* 50, 1 (03 2024), 237–291. https://doi.org/10.1162/coli_a_00502 arXiv:https://direct.mit.edu/coli/article-pdf/50/1/237/2367175/coli_a_00502.pdf