Large Language Model Aided Modeling of Dyadic Engagement

Cheng Charles Ma

CMU-CS-24-105

May 2024

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:

Fernando De la Torre	*Carnegie Mellon University*, Chair
Daphne Ippolito	*Carnegie Mellon University*
Lori L. Holt	*University of Texas, at Austin*

Submitted in partial fulfillment of the requirements for the Master's degree in Computer Science.

Keywords: Dyadic Engagement, Large Language Models, Affective Computing, Multimodal Applications, Smart Glasses, Prompt Engineering, Commonsense Reasoning

Abstract

Over the past decade, wearable computing devices ("smart glasses") have undergone remarkable advancements in sensor technology, design, and processing power, ushering in a new era of opportunity for high-density human behavior data. Equipped with wearable cameras, these glasses offer a unique opportunity to analyze non-verbal behavior in natural settings as individuals interact. Our focus lies in predicting engagement in dyadic interactions by scrutinizing verbal and non-verbal cues, aiming to detect signs of disinterest or confusion. Leveraging such analyses may revolutionize our understanding of human communication, foster more effective collaboration in professional environments, provide better mental health support through empathetic virtual interactions, and enhance accessibility for those with communication barriers.

In this work, we collect a dataset featuring 34 participants engaged in casual dyadic conversations, each providing self-reported engagement ratings, augmented with external raters' assessments of engagement. We introduce a novel fusion strategy using Large Language Models (LLMs) to integrate multiple behavior modalities into a "multimodal transcript" that can be processed by an LLM for behavioral reasoning tasks. This fusion method is one of the first to approach "reasoning" about real-world human behavior through a language model. This work also explores the creation and features derived from LLMs for multimodal models to aid the task of engagement modeling. Smart glasses provide us the ability to unobtrusively gather high-density multimodal data on human behavior, paving the way for new approaches to understanding and improving human communication with the potential for important societal benefits. The features and data collected during the studies will be made publicly available to promote further research.

Acknowledgments

I would like to thank my advisor, Professor Fernando De La Torre, for providing me with the opportunity to work on an exciting project with amazing people. His guidance on this project and mentorship has been invaluable. I would also like to thank Professor Lori Holt and Professor Daphne Ippolito for their expertise and support.

I would like to thank other members of the Human Sensing Lab who worked on this project, including Dr. Alexandria Vail, Álvaro Fernández García, Kevin Hyekang Joo, Sunreeta Bhattacharya, Kailana Baker-Matsuoka, Sheryl Mathew, without whom this project would not be possible. Dr. Vail's advice and guidance has been particularly impactful in shaping this work.

Additionally, I'd like to also thank Professor Rema Padman and Professor Yi-Chin Lin for sparking my passion for research.

Finally, I'd like to thank my friends, in particular Aden Fiol, Parth Maheshwari, Mehul Agarwal, my girlfriend Brianna Fan, and my family for their unwavering support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Wearable computing devices, also known as "smart glasses," offer new approaches to quantifying and understanding human behavior through unobtrusive, high-density behavior tracking. Equipped with sensors such as a video scene camera to monitor the wearer's view, an eye camera to estimate gaze, a microphone to record speech, and an inertial measurement unit to measure head orientation, smart glasses can capture and respond to human behavior as it unfolds in real-time and real-world contexts. There are numerous potential future applications for such systems: for example, facilitating navigation among the visually impaired, or augmenting social cues for individuals with difficulties reading nonverbal signals.

Although there has been substantial prior research in laboratory settings [8, 62, 76] and human-agent interaction [6, 11, 44], there are still many rich, unexplored opportunities in natural social contexts, for which smart glasses offer unique capabilities for study. With smart glasses, we can capture authentic social interactions that are not constrained by the artificial settings of a laboratory, but rather occur in the natural course of daily life, as we seek help, share information, learn, and maintain social bonds through face-to-face communication. These interactions are rich, nuanced, and impacted moment-by-moment by multimodal cues both overt and subtle. The stakes can be high: human conflict — between couples, among friends and families, in leadership and governing bodies, and even among societies — occurs when communication breaks down. Face-to-face communication plays a fundamental role in maintaining group cohesion, preserving mental health, fostering academic learning, and supporting developmental growth.

The concept of engagement has been long recognized as a key determinant of communication success. Engagement, while lacking a precise definition, can be loosely defined as the level of attentional and emotional investment that an individual puts forth during communication [55]. The numerous other ways of defining and annotating engagement are explored in section 2.1. The ability to captivate in conversation can determine life-changing interactions, whether it's acing a job interview or making a favorable impression on a first date. The depth of our engagement and that of our partner shapes the outcomes of many social, educational, and professional activities.

For the most part, humans automatically and implicitly pick up on the subtle, variable cues that convey engagement in a conversation. Yet, building systems that accurately measure and gauge conversational engagement remains a formidable challenge. Difficulties arise with the complexity and subtlety of human behavior, its context-dependence, and its variability across personal histories and cultural backgrounds. Further complicating matters, social communica-

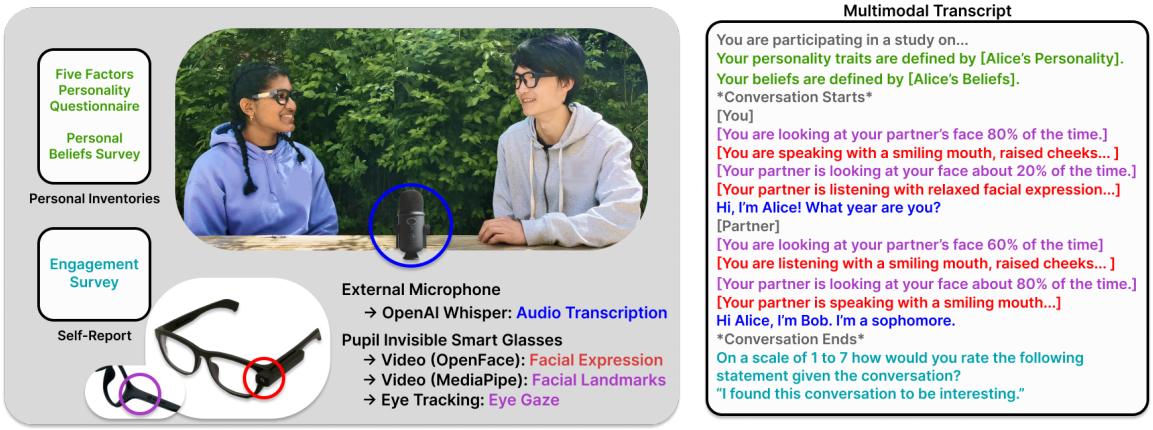

Figure 1.1: Visual representation of recorded behavior modalities during casual conversation and a sample of the multimodal transcript illustrating their fusion as introduced in this work. The goal is to predict engagement from this multimodal data. Color-coded modality names correspond to lines of the same color in the multimodal transcript.

tion is inherently multifaceted, with engagement likely to be conveyed across verbal content, nonverbal cues like tone of voice, facial expressions, hand and head gestures, and also through the absence of overt signals, such as extended periods of silence or few gazes to a partner's face. The unpredictable and dynamic nature of social exchanges likewise makes predicting engagement difficult, as patterns of engagement can shift rapidly and vary widely across contexts. Thus, techniques able to perform effectively with minimal or no in-domain training are of particular interest.

The dearth of relevant data presents another challenge. Although there is an abundance of openly available datasets of dyadic interactions from a third-person viewpoint, such as IEMOCAP [7], SEMAINE [44], MEISD [20], MELD [57], or NoXi [8], naturalistic dyadic interactions captured from an egocentric viewpoint are scarce. Collectively, these factors pose significant challenges for building socially-aware artificial systems that accurately interpret and respond in a manner that feels authentic and engaging to humans. Nonetheless, there is good reason to work to meet these challenges. Imagine a system that can gauge audience engagement with a teacher's lecture and provide on-the-fly feedback they can use to better engage their students. Or consider assistive technologies that can offer alternative presentations of challenging social signals for those living with communication disorders. The potential applications are extensive.

There are three main contributions to this work. We firstly introduce a novel dataset including recordings of natural, unscripted conversations among unfamiliar dyads described in chapter 3. This is defined as a pair of people, wearing Pupil Invisible smart glasses, as illustrated in the left-hand segment of Figure 1.1. This dataset contains conversations between 19 unique dyads, including video and audio recordings, eye tracking, self-reported information on demographic, political, and personality factors from the participants. Each session also contains first-person ratings of engagement, captured by each participant's answers on a multi-item survey, and third-person ratings of engagement, captured by ratings of engagement by third-party annotators for 10 second thin slices of the session.

The second contribution is a novel fusion approach described in chapter 4, which uses a large language model (LLM) as a "reasoning" engine to predict post-session self-report measures of engagement. This is achieved by fusing behavioral measures into a multimodal textual representation. A sample transcript can be seen in Figure 1.1 and the full sample transcript is displayed in Appendix A. This approach is a powerful, simple, and flexible framework for future work on modeling human behavior and developing socially intelligent technologies.

The third contribution of this work explores the potential of features derived from LLMs in order to predict external rater evaluations at 10-second intervals, described in chapter 5. Given the relationship between first-person self-reports of engagement and third-party ratings, we explore whether an LLM's prediction of self-reported engagement assists in prediction of third-party labels. This opens the possibility of deriving semantically meaningful features from LLMs that may assist in downstream tasks.

Chapter 2

Related Work

This section provides an overview on the definitions of engagement, the types of engagement that been explored, and existing work on understanding analyzing human behavior with classical machine learning models and large language models.

2.1 Definitions of Engagement

Engagement is a well studied concept in various domains that range from human-computer interaction to psychology [52]. There is a large amount of variation, ambiguity, and overlap in the definitions that do exist in literature. The range of definitions highlight the complex and multi-faceted nature of engagement. However, one can observe that these definitions are not contradictory but rather complementary and capture different aspects of this intuitive notion of engagement we hold in addition to the definition presented in chapter 1.

Within fields like affective computing, human-computer interaction, and human-agent interaction, engagement is typically defined as a state or process [52]. Inoue et al. presents engagement in a binary manner, whereby a person either has no interest or is following the conversation [31]. One of the most commonly used definition of engagement is "the process by which interactors start, maintain, and end their perceived connections to each other during an interaction" originally defined in [66] and used in [18, 30, 67]. There is also a distinction based on the entity receiving the engagement: social engagement refers to engagement between human-human or human-agent interactions, and task engagement refers to engagement in the task the human and agent are involved in [52].

It's also important to note the definitions of engagement in other domains. In a review of engagement at work in business and health, Simpson presents one definition of engagement as "employing or expressing of oneself physically, cognitively, and emotionally during work role performances" [68]. In the context of student engagement, Wong and Liem defines it as "psychological state of activity that affords them to feel activated, exert effort, and be absorbed during learning activities" [80].

2.1.1 Types of Engagement Annotations

Engagement is typically annotated in two ways: first-person and third-person perceptions of engagement[52].

The first is a session level, first-person annotations. This are typically collected through surveys designed to capture aspects of engagement and captures the participant's perception of the entire interaction. Jaques et al. designed a model to predict bonding in novel conversations between dyads [34]. They utilized post-study surveys completed by each participant that captured different aspects of bonding, and a median split of scores on this survey corresponded to the high or low bonding labels used to train their model. The creation of such a study is also used in Psychology in works like [15]. Self-report measures are also common means to assess student and patient engagement [21, 25].

The third-person perception of engagement employs an external rater or raters to annotate the interaction. These annotations are done on a frame level [18, 23] or "thin-slice" level, which refers to segments ranging from 5 to 30 seconds [40], which provides a temporal dimension to the engagement annotations and attempts to capture conversational dynamics.

Finally, there is a smaller body of work that utilizes another metric as a proxy for engagement. Gray and Perkins employ student attendance as a means to gauge student engagement [27], and Naik and Kamat use logs from an Moodle, an online learning platform instead[48].

2.2 Multimodal Fusion

This section presents existing work on modeling social interactions and human behavior with machine learning, and different multimodal fusion techniques to achieve this. Multimodal Fusion refers specifically to integrating information from multimodal modalities with the goal of predicting an outcome measure [4], like engagement in this setting.

Curhan and Pentland used speech features (conversational engagement, prosodic emphasis, and vocal mirroring) in the first five minutes of a simulated negotiation to predict the outcomes of the negotiation [16]. Using these features, they predicted 30% of variance in negotiation outcomes, demonstrating the value of speech features in conversational dynamics. This result suggests that speech features have a similar importance in predicting conversational engagement. Activity level and mirroring had differing relationships with outcome depending on the assigned position of participants showing that perceived status can affect how conversational dynamics relate to negotiation success. This interaction poses the question of how status affects how features predict conversational engagement.

Pellet-Rostaing et al. used prosodic-acoustic, prosodic-temporal, mimo-gestural, and linguistic features to predict the engagement level of the target participant while holding the speaking turn [55]. The study showed the value of using both visual and audio features achieving the best results with the prosodic-acoustic, prosodic-temporal, and mimo-gestural modalities. Achieving similar results to studies using annotator-defined segments demonstrated that annotating engagement at a turn level can be effective.

In our study, we attempted to use gaze as a means of gauging dyadic interaction, along with other modalities, as it is evidenced by some to have correlations with engagement [24, 61].

Goodwin emphasizes the interconnected nature of gaze behavior among participants in a conversation and points out that the way individuals direct their gaze is not a solitary or random act but is deeply intertwined with the social dynamics of the interaction [24]. This gaze behavior acts as a nuanced signal of a participant's level of attention and engagement, reflecting whether they are actively participating or disengaging from the conversation. Furthermore, Goodwin explores the concept of gaze withdrawal as a strategic communicative gesture that participants use to signal their intentions within the conversation, such as making a bid for closure or expressing a particular understanding of the conversation's trajectory.

Moreover, Ranti et al. underscore the potential of utilizing eye-blink measures as a reliable indicator of an individual's subjective engagement with various stimuli [61]. By closely analyzing the timing of blink inhibition in response to unfolding scene content, they found that they can uncover the viewers' unconscious, subjective evaluations of the importance and engagement level of what they are observing. A notable observation is that a slower blinking rate is often associated with a higher degree of engagement, suggesting that individuals are more absorbed and attentive to the conversation or content in front of them.

Given the dominance and success of the transformer architecture and an extremely wide range of domains, Transformer-based multimodal learning has become a widely studied topic in the field of Multimodal Machine Learning [81]. Particularly interesting architectures include methods proposed by Nagrani et al., in which the authors propose a novel method "bottleneck fusion" layer which utilizes the attention mechanism to force information between different modalities to pass through a small number of bottleneck latents, which in turn requires the model to learn important information relevant to each modality [47]. They also contrast this approach with vanilla cross attention for multimodal fusion. Another inspiring architecture was proposed by Pramanick et al., in which they combine cross-attention and an optimal transport kernel to perform multimodal fusion.

2.3 Large Language Models (LLMs)

The capabilities and accessibility of LLMs have opened up a wide range of potential applications, particularly the case in fields related to human subjects like psychology. They range from creating synthetic datasets of LLM-generated responses in humanless experiments [17] to providing automated feedback to clinicians [69].

One application involves exploring the ability of LLMs to mimic human behavior, because of their potential to reduce the need for human subject experiments and power realistic, interactive interactions. Aher et al. explore the ability of LLMs to reproduce human subjects' behavior in classic experiments, such as the "Wisdom of Crowds" [1]. Argyle et al. investigate the potential of LLMs as proxies for human sub-populations in social science research [2]. Park et al. introduce generative agents powered by LLMs that simulate believable human behavior in a virtual environment [54], also similarly seen in [84]. There is also a body of work on understanding the personality of LLMs, identifying ways to manipulate the personality embodied by an LLM, and injecting personality into LLMs to predict human responses with respect to values [35, 65].

Another application involves exploring the ability of LLMs to understand human behavior. This line of work involves evaluating their theory of mind abilities, which refers to the ability to

understand the mental states of others, such as purpose or intention [59]. Prior work has proposed various benchmarks and methods to evaluate an agent's theory of mind [37, 63, 64].

These works are essential to assessing the ability of LLMs to simulate and understand human behavior. However, they are all limited to static benchmarks or simplified, virtual interactions. There is a lack of work exploring the ability of LLMs to simulate and predict the outcomes of a real, human social interaction, such as the task of predicting a person's responses to a survey that measures engagement. We argue that this is another dimension that should be considered when developing LLMs to simulate and understand behavior.

Our work proposes a dataset and method for unifying the work on simulating and understanding engagement in social interactions with LLMs grounded in real, in-the-wild social interactions data. Given that LLMs provide a promising path towards developing, utilizing real social interactions in research on the social intelligence of LLMs and AI systems more broadly is a crucial next step towards the development of models that can successfully model engagement be socially intelligent.

2.3.1 Socratic Models

Understanding and interpreting the reasoning of machine learning models is widely recognized to be a significant challenge. Typically, models encode behavioral features into a high-dimensional, abstract vector space, which is then mapped onto the target prediction space. To understand a model's inner workings, we usually project these intermediate data into a space that is more comprehensible to humans, often through visualization techniques. However, consider the possibility of the inverse — rather than allowing the model to obscure information into abstract dimensions, we could direct its operation into a universally interpretable space: the domain of language itself. When studying a topic like human behavior from a computational perspective, AI systems like LLMs that utilize language to "reason" about said topics are worth further study because the language allows for nuance and ambiguity that inherently exists in these fields.

Socratic Models, named for the ancient Greek philosopher's method of teaching through cross-examination, use language to integrate information from a diverse set of modalities [82]. Within this framework, pre-trained models fine-tuned toward specific modalities or behaviors translate their interpretations of inputs into natural language. This translation is formulated into a language prompt to direct the reasoning of an LLM. This approach allows a set of pre-trained models to "discuss" various multimodal information, akin to asking and answering questions in a Socratic dialogue. By framing the task as a language-driven exchange, the Socratic Model framework allows a set of pre-trained models, each specialized in a distinct domain, to perform downstream multimodal tasks without the need for further training or fine-tuning.

Thus far, there have been only a few early attempts at applying this framework for prediction. In the domain of image captioning, one study revealed that an ensemble of models within the Socratic Models framework generated captions that substantially improve upon the capabilities of the zero-shot state-of-the-art ZeroCap [72]. However, when compared to fine-tuned models such as ClipCap [45], performance was not as impressive; yet, this performance gap narrowed considerably when the ensemble was provided a small set of example captions from the training set, suggesting its potential in few-shot learning scenarios [82].

This concept of "many-to-one" alignment has been explored from other angles as well.

ImageBind, for instance, develops a multimodal representation through a set of image-paired modalities [22] while LanguageBind extends video-language pre-training to a broader range of language-paired modalities [85]. However, both of these models still face the challenge of abstracting information. ImageBind and LanguageBind create "bindings" centered around a specific modality but do not explicitly work within that modality itself. Instead, they map a primary modality into an abstract space and then align information from other modalities to this space, resulting in a multimodal representation that resembles the embedding of the primary modality. While this approach has proven effective at abstract tasks such as video-text alignment and image-text retrieval, it is less effective in providing human users with a coherent understanding of its reasoning. Our research aims to follow a similar path but with a crucial distinction: our embedding space is designed to be language itself, which may offer a more direct and interpretable framework for multimodal learning.

Previous studies have established the value of the language modality in understanding complex social phenomena, such as rapport [9], affinity [32, 33], and, as in the present work, engagement [3]. A variety of computational methods have been employed to extract this information from language, from rudimentary bag-of-words approaches to more sophisticated neural network models [70, 74]. Recent advancements, however, have seen a considerable increase in LLMs adapted to augment tasks requiring social intelligence: notable applications have included refining persuasive communication for public health campaigns [13, 36] and identifying adverse social determinants of health within free-form clinical notes [28]. One of the objectives of the present work is to explore the utility of LLMs for behavior analysis of social interactions: in our case, estimating the conversational engagement of speakers in a dyadic interaction. The proposed approach centers around employing OpenAI's `GPT` models to impersonate each participant in the conversation by responding to the self-reported questionnaire in a zero-shot manner. This is achieved through reconstructing the conversation using multimodal-informed prompting that combines behavioral information, inspired by the Socratic Models framework proposed by Zeng et al. [82].

Chapter 3

Dyadic Interaction Dataset

In our work, we collected and studied recordings of pairs of strangers wearing smart glasses conversing in a recording room, depicted in Figure 3.1. While the initial prompt encouraged discussion about personal experiences during COVID-19 as a shared ice-breaker, they were informed that they were not obligated to limit the conversation to that topic and could discuss whatever topic they felt comfortable with.

3.1 Population

Our study contained 34 unique participants and 19 unique dyads[1]. Demographically, 14 participants identified as male, 19 identified as female, and one identified as non-binary; 47% identified as Asian and 38% identified as White/Caucasian. All participants were 18–35 years of age but were primarily in their early twenties. Participants were recruited from Carnegie Mellon university through various physical and digital media and word-of-mouth. Participants were required to be fluent in English and have normal or corrected vision with contact lenses (to avoid conflict with the smart glasses).

3.2 Procedure

The entire recording session lasted approximately 15 minutes, including introductions and closing. Each participant was equipped with a pair of smart glasses (refer to section 3.3 for detailed specifications) capturing their field of vision, head motion, and gaze. While the smart glasses are advertised to work well across recording sessions without calibration, they benefit from calibration when changing users [73][2], so we conducted a calibration procedure for each participant before the beginning of the session. At the start of the session, participants were suggested an ice-breaker: their experience during COVID-19, a universally shared topic, but the conversation was not constrained to this topic. Following the session, participants completed questionnaires

[1]Two participants appeared in multiple dyads, but all dyads were unique.

[2]Note that the referenced study was conducted by the manufacturer of the device.

Figure 3.1: Side view of dyad in dataset

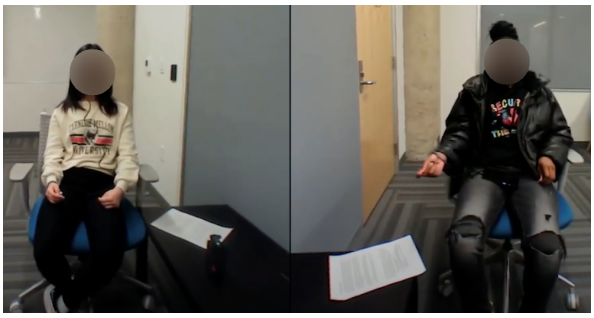

Figure 3.2: Frames from egocentric video

on their beliefs, personality, and engagement during this interaction (refer to section 3.4 for details on the questionnaires).

3.3 Recording Instruments

Each session was recorded using Pupil Invisible smart glasses worn by all participants and a centrally placed external microphone to record the dialogue.

3.3.1 Pupil Smart Glasses

Each participant was equipped with Pupil Invisible[3] smart glasses manufactured by Pupil Labs, specially designed to closely resemble regular eyeglasses for user comfort and a discreet appearance. The key features of these smart glasses that we leverage in our work include the following:

- **Scene camera:** A detachable camera mounted on the left arm of the glasses frame captures the wearer's field of view with an $82° \times 82°$ viewing angle, at a resolution of 1088×1080 pixels and a frame rate of 30 Hz.
- **Eye gaze tracking:** Two IR cameras, positioned near the would-be hinges of the glasses frame, record eye movements at a resolution of 192×192 pixels and a frame rate of 200 Hz. Post-processing software provided by the manufacturer converts this data into 2D gaze points at 120 Hz in scene camera coordinates. This system is advertised to achieve an uncalibrated accuracy of approximately $4.6°$, but calibration per user can enhance accuracy [73].

3.3.2 Stereo Microphone

In addition to the recordings captured by the smart glasses' scene camera, we used an external high-quality stereo microphone to record the conversation at a standard 44.1 kHz sampling rate. This decision was made after determining that the quality of the audio captured by the smart glasses scene camera was insufficient for acoustic analysis. To synchronize the media streams,

[3] `https://pupil-labs.com/products/invisible`

participants were instructed to perform a hand clap at the start of each session, emulating the clapperboard technique commonly used in film production.

3.4 Self-Report Questionnaires

The participants were asked to complete a questionnaire that measured self-reported engagement after each interaction. The engagement questionnaire consisted of 53 items based primarily those used in previous studies on perception of interaction quality [14]: refer to Appendix A for detailed information and statistics regarding the items in the engagement questionnaire. The participants were also asked to complete the Big Five Inventory [42] for personality information and a handcrafted questionnaire on political views. We based our political typology questionnaire on a set of socio-cultural issues that have been studied to gauge polarization along the political spectrum [56].

3.5 Observer Ratings of Engagement

In addition to gathering self-reported data from participants after each session, we collected third-party annotations of perceived engagement, a similar but distinct concept that has also been the focus of prior literature [52]. The recordings were segmented into ten-second intervals, which were subsequently provided to the annotators in randomized order, presenting both members of the dyad simultaneously but preventing exposure to the temporally surrounding context. The annotators watch paired egocentric video, similar to what is seen in Figure 3.2, except the faces remained visible.

Annotators were tasked with rating the participants' engagement within each conversation segment using a Likert scale ranging from 1 (not engaged at all) to 7 (extremely engaged). We deliberately provided limited instruction to capture the annotators' instinctive sense of engagement through their personal interpretation of behavior cues. This method aimed to gather a genuine and diverse range of perspectives on this definition.

To assess the reliability of our annotation approach for subsequent analysis, we conducted a pilot inter-annotator agreement study. Seven annotators independently rated a representative sample of 19 segments, one segment per session, constituting roughly 4% of the total dataset. Upon completion, we calculated the level of inter-annotator agreement using the ordinal variation of Krippendorff's alpha [29]. The resulting score of approximately 0.70 reflected a moderate level of agreement across annotators, a satisfactory outcome given the highly subjective nature of the engagement construct and the intentionally unstructured annotation approach. With this level of agreement, we were confident in the robustness of our annotation protocol, and the remaining data was annotated by a single annotator.

3.6 Feature Extraction

After gathering the recordings, the data required pre-processing. Initially, we adjusted the video to eliminate the radial distortion introduced by the lens of the scene camera. This was achieved

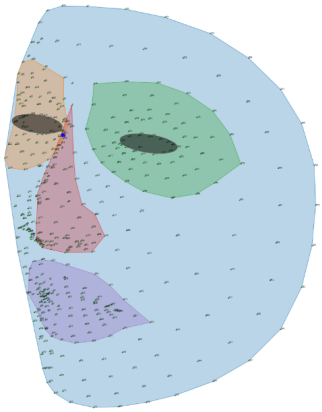

Figure 3.3: Convex hull and features from Mediapipe

by applying the distortion coefficients provided by the manufacturer[4]. Due to the differing frame rates between the eye-tracking camera and the egocentric scene-view camera, we also synchronized the data to a unified 30 fps timestamp.

3.6.1 Facial Expression

Facial action units (FAU) from the processed video were extracted with OpenFace 2.0 [5]. Since OpenFace achieves optimal performance when the face in the image exceeds a width of 100px, we needed to upscale our data to meet this suggestion. For each frame, we used MediaPipe [41] to identify the location of the face in the image, then cropped and rescaled the image to ensure that the face was centered and was at least 240px wide and the final dimensions were 1080x1080px. If no face was detected for a particular frame, the location of the face in the previous frame was used.

3.6.2 Gaze Tracking

For every frame, we determined whether a participant's gaze is directed towards their partner's face, recognizing the significance of gaze in forecasting engagement [10, 49]. This was accomplished by creating a convex hull using the 478 2-dimensional face landmarks extracted from MediaPipe to outline the face. A gaze point captured by Pupil smart glasses was deemed to be on the face if it fell within the convex hull (including its boundary) or within 30% of the width of the face's convex hull to account for the potential possible inaccuracy of gaze prediction from Pupil smart glasses described in its specifications. The convex hull is visualized in Figure 3.3 with the blue dot corresponding to the gaze location.

[4]For details on the information provided by the manufacturer, see Pupil Lab Invisible recording and export instructions: `https://docs.pupil-labs.com/invisible/data-collection/data-format/`.

3.6.3 Dialogue Transcription

OpenAI's Whisper [60] was used to transcribe the recording from each session. Whisper outputs fine-grained segments with start and stop times around a few seconds long. For each segment, a speaker was assigned. If the segment contained speech from both speakers, the speaker who spoke most was assigned. Diarization tools like PyAnnote and source separation tools performed poorly with audio from our dataset, so manual labeling was chosen.

3.6.4 Audio Features

Each session's main audio recording was processed with Torch Audio to extract MFCCs, which are 39 dimensional features. They are created from 13 MFCC features, 13 delta MFCC features which are the first order derivatives of the MFCCs, 13 delta-delta MFCC which are the second order derivatives of the MFCCs, which are typically used in ASR systems [38].

Chapter 4

LLM Fusion

This section discussed how we to used large language models (LLMs) to "reason" about a social interaction using multimodal information. Our method involves prompting an LLM to simulate a study participant, answering the end-of-session engagement questionnaire as though it were the participant themselves.

One of the objectives of the present work is to explore the utility of LLMs for behavior analysis of social interactions: in our case, estimating the conversational engagement of speakers in a dyadic interaction. The proposed approach centers around employing OpenAI's GPT models to impersonate each participant in the conversation by responding to the self-reported questionnaire in a zero-shot manner. This is achieved through reconstructing the conversation using multimodal-informed prompting that combines behavioral information, inspired by the Socratic Models framework proposed by Zeng et al. [82].

4.1 Algorithms for LLM Fusion

The LLM Fusion approach enables an LLM to impersonate a participant by creating a multimodal prompt augmented with textual representations of non-verbal behavioral information. These textual representations are formed from data provided by the smart glasses, multiple pretrained models, and personality information. This multimodal transcript effectively captures the dynamics of a social interaction. This method can be extended to contain any number of other behavioral cues. The multimodal prompt enables an LLM to effectively gauge the level of self-reported engagement. OpenAI's chat models `GPT-4` and `GPT-3.5` [1] were used in this work, but any other LLM could be used.

4.1.1 Modalities:

The gaze, facial expression, and audio modalities were used because of their ability to be represented by text and significance to modeling engagement.

The *speech* modality forms the backbone of the prompt. This information comes from the transcript augmented with speaker labeled segments described in subsection 3.6.3.

[1] Fixed to be `GPT-4-0613` and `GPT-3.5-turbo-0613` for consistency.

The *gaze* modality is a string that containing the percentage of time a speaker's gaze is on the face of the other person. The percentage is rounded to the nearest 10% for brevity.

The *face* modality facial action unit is converted into a description of the dominant facial expression for a given snippet of video. Using the FAUs obtained from OpenFace 2.0 described previously, each frame was mapped to one of the following emotions {happy, sad, surprise, fear, anger, disgust, contempt, neutral} with the table from [19], which is a method used in other publications like [71] and software like iMotion's Affectiva [43]. Neutral was assigned if none of these emotions were applicable. The emotion that occurred most often was assigned to a snippet, and the text descriptions of the emotion are from [83], which was obtained from prompting ChatGPT and achieved state-of-the-art performance on the Dynamic Facial Expression Recognition problem.

The participants' responses to the Big-5 personality survey and political beliefs survey were also incorporated as part of the system message in different ablations, which provides additional context to guide model behavior. This information was included because a participant's personality and beliefs will likely influence their behavior during the dyadic interaction and response to the engagement questionnaire.

4.1.2 Creating the Multimodal Transcript

The messages provided to GPT use the start and end times from segments of Whisper's transcription as atomic units to which information from other modalities is added. Consecutive segments with the same speaker are merged to combine speech and other modalities into a larger temporal window.

`GPT` imitates each participant using the following procedure. Each merged Whisper segment forms a message in the messages field of OpenAI's ChatCompletion API. For each message in the list of messages, the *role* field is set to *assistant* if the segment is uttered by the participant being imitated and *user* if the segment is uttered by their conversation partner. The last *user* message before `GPT`'s response will always be from the perspective of the experimenter who introduces a question from the engagement questionnaire (see Appendix A).

The last assistant message is the response to the engagement questionnaire question and is generated by `GPT`. The experiments also only used the transcripts where the end timestamp is less than or equal to 5 minutes in accordance with traditional psychology literature, whereby the first 5 minutes of a conversation enables people to relatively successfully predict its outcome [16]. This also helped reduce the cost.

4.2 Experiments

For each question on the engagement questionnaire, `GPT-4` was provided the multimodal transcript and prompted to predict the participant's response. A few of the items on the questionnaire explicitly reference laughing or eye contact; we decided to include these questions to explore whether `GPT-4` could infer information about modalities with limited information. Note that only 17 sessions were used out of the 19 because they had satisfactory transcriptions by Whisper. This corresponds to 34 sets of responses, but five sets of responses came from two participants

as they appeared in five out of the 17 sessions already in dataset. We performed a set of ablation experiments to explore the significance of various feature sets:

- **4**: Raw transcript only.
- **4S**: Raw transcript plus participant personality and belief questionnaire responses (provided as **(S)ystem dialogue**).
- **4G/4SG**: Addition of **(G)aze** features to **4/4S**
- **4F/4SF**: Addition of **(F)acial expression** features to **4/4S**.
- **4GF/4SGF**: Adds both **(G)aze** and **(F)acial expression** features to **4/4S**.

The length of the multimodal transcript exceeded the input constraints of `GPT-4` in three cases: two sessions for **4SGF** and one session for **4GF**). In these cases, the transcript was truncated: a t-test comparing the residuals of the truncated sessions with those of the other sessions yielded p-values of 0.186 and 0.648, indicating no significant difference between the full and truncated sessions. Based on this evidence, we felt confident including all sessions in subsequent analysis.

The *temperature* parameter, which determines the LLM's decoding strategy, was set to 0 to ensure sampling from the most-likely responses to the questionnaire. In cases where `GPT-4` did not return a numeric response, we selected the most-likely numeric response from the top 20 tokens by log probability for the first generated token (see Appendix C for more details on these cases).

4.2.1 LLM Fusion Results

Our analysis explores the ability of `GPT-4` to predict each participant's exact answer to survey items, as well as its ability to predict the valence and arousal associated with their responses. Valence refers to the positive or negative degree of emotion (e.g., pleasure/displeasure), while arousal refers to the intensity of emotion (high or low) [46]. In the context of our study, valence is defined by a response in the disagree range (1–3), neutral (4), or agree range (5–7), and arousal is defined as the "absolute value" from neutral, i.e., $|\text{response} - 4|$. Although `GPT-3.5` was tested as well, its results have been omitted due to significantly poorer performance compared to `GPT-4`. Full results are displayed in Table 4.1.

Exact Response Notably, when considering the Krippendorff's alpha metric, `GPT-4` achieves a "moderate" level of agreement with our study's participants on average and falls within the range of $[0.470, 0.543]$. However, we observe that when considering the precise level of agreement, it is still below the typically acceptable threshold level of around 0.6 [79].

Valence

Most remarkably, the valence of `GPT-4`'s responses achieve a significant level of agreement with the study's participants. As demonstrated in Table 4.1, all responses fall within the interval of $[0.61, 0.80]$ outlined in [39].

Upon closer inspection of the valence predictions of the `LLM-4S` model, the ablation with the highest value of inter-annotator agreement for exact responses (presented in detail in Table 4.2), we can observe that `GPT-4` reliably responds "agree" to participant "agree" responses, with an extremely high class accuracy of 91.8%. `GPT-4` is less reliable to respond "disagree" to participant "disagree" responses, achieving a class accuracy of 66.1%. Notably, `GPT-4` rarely

Ablation	Exact	Valence	Arousal
4	0.470 (0.209)	0.634 (0.246)	0.055 (0.169)
4S	0.518 (0.217)	0.687 (0.252)	**0.071 (0.174)**
4F	0.513 (0.203)	0.686 (0.250)	0.053 (0.164)
4GF	0.520 (0.212)	0.695 (0.259)	0.066 (0.180)
4S	**0.543 (0.206)**	0.680 (0.244)	0.054 (0.185)
4SG	0.535 (0.210)	0.702 (0.247)	0.039 (0.172)
4SF	0.532 (0.202)	0.698 (0.247)	0.055 (0.170)
4SGF	0.531 (0.193)	**0.703 (0.248)**	0.047 (0.180)

Table 4.1: Krippendorff's Alpha scores, mean and standard deviation, for each ablation (higher is better).

	Predicted				
Actual	**Agr.**	**Neu.**	**Dis.**	**All**	**Cl. Acc.**
Agr.	1072	44	52	1168	91.8
Neu.	91	18	33	142	12.7
Dis.	105	62	325	492	66.1
All	1268	124	410	1802	56.9

Table 4.2: 4S Valence Prediction Confusion Matrix: Responses are categorized as (Dis)agree (1–3), (Neu)tral (4), or (Agr)ee (5–7). Class accuracy is also reported.

responds with "neutral" to participant "neutral" responses, only achieving a class accuracy of 12.7%. While worse performance on the neutral response may be expected given that the label range is smaller (the response must match "4" exactly), we can observe that `GPT` consistently generates answers in the agree/disagree class, with a tendency for the agree class. We hypothesize that `GPT`'s process of reinforcement learning from human feedback (RLHF) [53] to follow instructions and reduce toxicity may contribute to overly positive responses from `GPT-4`. This calibration may inadvertently introduce bias against "negative" responses.

Arousal On the other hand, across all ablations, `GPT-4` is unable to accurately predict the level of arousal of a participant's response, performing only marginally better than chance with Krippendorff's alpha scores in the range of $[0.047, 0.071]$ (see Table 4.1 for detail). While `GPT-4` appears able to predict the general attitude of the participant towards a questionnaire statement, it is not able to reliably determine the strength of the participant's feelings.

Contribution of each modality

We study the impact of each modality described in subsection 4.1.1 on the exact predictions of `GPT-4` by conducting a two-tailed paired t-test of each ablation's residuals against those of the `LLM-4` text baseline. The addition of every single/combination of modalities results in a statistically significant positive contribution ($p < 0.05$) to the performance of `GPT-4`.

However, when comparing the 4SG, 4SF, 4SGF ablations against the **4S** ablation, the addition of gaze or facial expression information or both appear to negatively impact the ability of GPT-4 to predict the raw response, while improving its ability to match the valence. In a paired t-test of residuals comparing exact predictions, the addition of facial expressions in the 4SF and 4SGF ablations worsened performance significantly ($p = 0.003$ and $p = 0.021$, respectively), whereas gaze did not have a notable impact ($p = 0.164$).

Performance across questions and following up with GPT

The following statements achieved the *best* performance across all ablations (mean accuracy and standard deviation):

1. *I felt like my conversation partner really listened to me*
   (mean 64.0%, std. dev. 7.1%);
2. *I became irritated with my partner at some points in the conversation* (mean 60.7%, std. dev. 5.9%); and
3. *My conversation partner seemed like a warm person*
   (mean 53.7%, std. dev. 6.2%).

The following statements achieved the *worst* performance across all ablations (mean accuracy and standard deviation):

1. *My conversation partner was quite sensitive*
   (mean 4.0%, std. dev. 1.6%);
2. *I would trust my conversation partner with sensitive information* (mean 8.8%, std. dev. 5.2%); and
3. *My partner and I laughed during our interaction*
   (mean 10.3%, std. dev. 4.1%).

Unsurprisingly, prediction performance on the question about laughter is quite poor, as the transcript does not explicitly include laughter. This situation is similar to the case of GPT-4 and statements about eye contact in ablations without gaze information. Interestingly, while the transcript did not explicitly contain laughter information, GPT-4 responds with the assumption that laughter did occur, given that its responses were not consistently negative. Although numerous caveats apply to these results, they nonetheless seem to reflect the opinions of our study's participants.

Chapter 5

LLMs as feature extractors

GPT-4 can reasonably predict each participant's first-person perception of engagement. Given that there exists a relationship between the scores from the self-reported measures and the third-party annotator's labels, this raises another key question: Is it possible to derive and use features from GPT-4's predicted responses in combination raw modalitites to improve the ability to predict the level of dyadic engagement? Namely, for behavioral related problems, can we utilize Large Language Models (LLMs) to extract semantically meaningful features? This is because there exist very few pretrained models designed to extract behavioral data, and as shown in chapter 4, LLMs like GPT-4 are already capable of predicting behavioral features to a reasonable degree.

5.1 Experiments

This section explores models to predict the third-party annotator's perception of engagement that utilize the predicted self-report scores from GPT-4 as a feature. Note that of the 19 dyads, only 15 dyads were used [1]. 1 participant appeared in 3 sessions and 1 appeared in 2 sessions.

The two novel models I propose are slight variants of each other, which firstly utilize the early fusion of the self-reported scores to "augment" the raw representation of each modality, and the late fusion of each "augmented" modality to form the final prediction. Given the small size of the dataset, I attempted to design a model to be as expressive as possible while balancing the complexity to account for the bias-variance trade-off, so they needed to be relatively simple to avoid overfitting.

Cross-attention is the main mechanism I explored to to fuse modalities. As described above, cross-modality attention, is a variant of cross-attention proposed in the original transformer [75] paper, whereby the query matrix and key matrix come from different modalities as a method to map different modalities into a common feature space, and thus learns inter-modality interaction. Since this is a transformer based model, I employed sinusoidal positional embeddings to add temporal information to each raw modality before passing it into the model.

Each 10 second interval of a dyadic interaction from our dataset is accompanied by the third party annotator's labels, which is an integer in the range $[1, 7]$ where 1 means the dyad is not

[1] 2 sessions were omitted because of labeling errors and 2 sessions were omitted because of poor transcription

engaged at all and 7 represents the dyad being highly engaged. Each participant has the following features from four separate modalities for each participant:

- Gaze: this is a 6 dimensional vector, which comprise of the features if the participant's gaze is on their partner's face, the distance of the gaze to their participant's face, the distance of their gaze to their partner's left eye, right eye, nose, and mouth.
- Facial Action Units: a 35 dimensional vector, which are either real valued numbers in the range $[0, 5]$ or indicator variables directly extracted from OpenFace [5].
- Mel-frequency cepstral coefficients (MFCCs): a 39 dimensional vector that represents the participant's speech.
- Predicted self-report engagement : a feature of dimension (53×1) or (53×7). The first representation corresponds to the numeric response from the 4S ablation section 4.2 prediction of self-report because it resulted in the highest Krippendorff's alpha value when predicting the exact response. The second (53×7) representation comes from using the probabilities over the tokens corresponding to each value in the $[1-7]$ likert scale collected from `GPT-4`'s responses.

The gaze (*g*), facial expressions, represented by facial actions units from OpenFace [5] (*f*), and audio represented by MFCCs (*a*) were selected because they are crucial predictors of engagement seen in related work described below and previous experiments are.

Most notably, there is a relationship between true participant's responses' relationship to the third-party annotator's labels. Data from each modality $m \in \{g, f, a, gpt\}$ will be denoted $X_m \in \mathbb{R}^{(L_m, d_m)}$ where L_m, d_m refers to the length and dimensionality of each modality. Note that $L_f = L_g = 300$, which is the number of frames in a 10 second slice.

5.1.1 Early Fusion

The cross-attention mechanism used in this work is described in Equation 5.1.

$$CrossAttention(Q_{GPT,m}, K_m, V_m) = \text{softmax}\left(Q_{GPT}K_m^T/\sqrt{d}\right)V_m \tag{5.1}$$

Where $m \in \{g, a, f\}$ represents on of the modalities we utilize. Typically, given the input X_m, the query, key, and value matrices are formed by a linear projection of X_m. However, here

$$Q_{GPT} = X_{GPT}W_{Q,m}, K_m = X_mW_{K,m}, V_m = X_mW_{V,m}$$

The intuition behind this that I wanted to explore the possibly of using the predicted self-reported scores to *augment* the raw modalities and somehow add behavioral information. Fortunately, the self-attention mechanism is one candidate method to do this. We learn a query matrix for the GPT scores for each modality. Then, as an analogue to the original self-attention mechanism, one element $(Q_{GPT}K_m^T)_{i,j}$ represents the similarity between a behavioral feature and the raw modality's feature at a certain time during the interval being processed. The softmax should therefore form a representation of the raw modality during a 10 second interval that is weighed with behavioral information.

The cross-attention mechanism is in the cross-attention block, which takes inspiration from a simplified transformer block. The last layer-norm layer is removed to reduce the complexity of

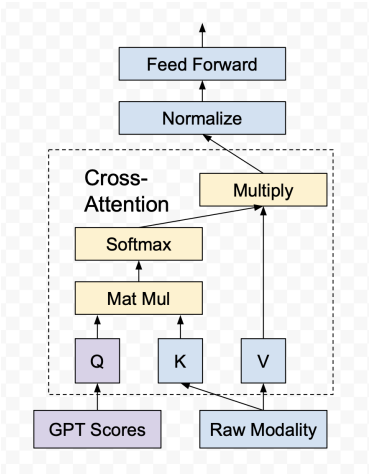

Figure 5.1: Early Fusion Cross Attention Block with GPT Features

the model. The entire block is displayed in the following fiture, and is repeated for each modality in the ablations containing all modalities.

5.1.2 Late Fusion

Each model uses an MLP as a method to predict the raw scores. For ablations that combine multiple modalities, inspired by the works of Pramanick et al. 58, the final feature vector passed into the MLP is of the following form, which is the concatenation of the tensors weighed with a learned multiplicative constant

$$Z_{multimodal} = [w_g Z_g, w_a Z_a, w_f Z_f] \tag{5.2}$$

Where $Z_{g,a,f}$ are the outputs of the modules containing the cross attention layers described in Equation 5.1, and $w_{g,a,f}$ are weights initialized to 1. The idea behind this is that they will dynamically adapt to the importance of each modality.

The desirable feature of self-attention enables each modality's representation $Z_{g,a,f}$ to be of the same shape so when weighing each modality's representation, it could be more fair.

5.1.3 Multimodal model

I've developed the *single* and *dyadic* multimodal model, which entail different methods of handling and fusing the data from each participant in a dyadic interaction as visualized in 5.2

1. Single Multimodal model: Each participant's features are concatenated together, and passed onto a single cross-attention module that attends to each modality. That is, each modality from both participant's are passed through the same module. Finally, the outputs are concatenated together with the late fusion approach described in subsection 5.1.2.
2. Dyadic Multimodal Model: Inspired by the work of Dermouche and Pelachaud 18, the features from each participant are attended to separately. The model implemented here differs in the sense that there is no module that attends to features from both participants

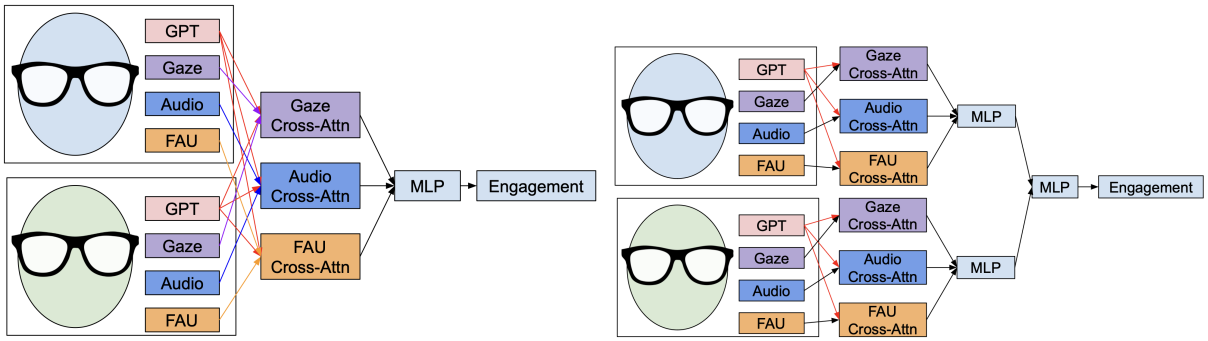

Figure 5.2: Single and Dyadic multimodal model. Left: Single, Right: Dyadic. Each Modality Cross-Attn is the Early Fusion Cross Attention Block shown in Figure 5.1

as originally proposed. The architecture of the model that processes a single participant's data is identical to that of the singular multimodal model with the exception of the last MLP's layer, and the output of both models are concatenated together and passed through a final MLP to predict the engagement score.

The model details are as follows:

- Attention Dimension: All models have the attention dimension (The second dimension of the matrix that forms Q, K, V) be 256.
- Single Multimodal Model: There are 3 layers of dimension 512, 128, 1.
- Dyadic Multimodal Model: The MLP for each dyad has two layers and dimensions 512, 256, and final MLP has two layers of dimensions 512, 256, 1.

5.2 Results

These parameters gave rise to a model of around 20 million parameters. All experiments used Mean Squared Error (MSE) loss, because the task is now regression and MSE effectively handles outliers. Adam was the optimizer used across all experiments, including baselines. The learning rate was fixed to be $1e-3$.

The full results are outlined in Table 5.1. They are reported from leave-one dyad out cross validation, that is a separate model is trained on 14 dyads and evaluated on all the intervals from the 1 dyad left out, and the results are the average of each model evaluated on the left out dyad without picking the best performing epoch so as to avoid cherrypicking results. This process is repeated for every dyad in the dataset. It's important to note why this is done instead of the traditional k-fold cross validation where the dyad the interval comes from is not taken into account. The MLP baseline numbers for each of the combinations of modalities were ommitted because they yielded similar/worse performance than the Bi-LSTM.

The single and dyadic multimodal models do appear to be promising ways to fuse representations from GPT and raw modalities in order to predict engagement. We can observe that the new Single and Dyadic multimodal model using the GPT and MFCC features achieve the lowest RMSE loss as well as the highest value of Krippendorff alpha. The Krippendorff alpha value is

Model	Modalities	RMSE (SD)	Acc	Class Acc
Bi-LSTM	Gaze	1.789 (0.787)	0.298	0.134
Bi-LSTM	FAU	2.415 (0.961)	0.275	0.156
Bi-LSTM	MFCC	2.273 (0.959)	0.243	0.14
Single	Gaze	1.672 (0.647)	0.3	0.166
Single	FAU	2.136 (1.039)	0.261	0.15
Single	**Audio**	**1.586 (0.772)**	**0.303**	**0.172**
Single	Gaze, FAU, Audio	1.755 (0.832)	0.316	0.176
Single (Probs)	Gaze, FAU, Audio	2.880 (1.414)	0.192	0.154
Dyadic	Gaze	1.766 (0.857)	0.303	0.162
Dyadic	FAU	2.545 (1.492)	0.26	0.141
Dyadic	**Audio**	**1.541 (0.549)**	**0.318**	**0.18**
Dyadic	Gaze, FAU, Audio	1.862 (0.878)	0.288	0.161
Dyadic (Probs)	Gaze, FAU, Audio	2.368 (1.284)	0.211	0.16

Table 5.1: Table of results: Top 2 performing models **bolded**. Lower is better for RMSE, higher is better for Krippendorff Alpha. Single and Dyadic refer to the Single and Dyadic Multimodal Model described in section 5.1. (prob) refers to the usage probabilities from `GPT-4` described in section 5.1. Class Acc refers to the average class accuracy (accuracy averaged across all classes).

calculated from rounding the prediction to the nearest integer, then passing it through the function used to calculate this metric because it only handles exact integer values. Furthermore, the ordinal version of Krippendorff alpha is used because our values are recorded on a Likert scale.

While the best model's performance with an average RMSE of 1.541 and standard deviation of 0.549 may appear reasonable on a subjective task and scale of 1-7, and it does outperform the baseline models, there is still a lot of room for improvement given the large standard deviation. Given that there are 7 classes, most models are better than random guessing $1/7 \approx 14\%$. The average class accuracy tell a similar story.

While it does seem like the features derived from GPT-4 through the self-reported scores do help the prediction of engagement, it's difficult to assert this with confidence given the small dataset and model.

Extensive experiments were run in order to investigate which modality or combination of modalities were most helpful in predicting engagement. It's interesting to note that for the Bi-LSTM baseline, the Gaze modality was the most helpful. However, the Audio modality represented by MFCCs were most helpful in the Single and Dyadic multimodal models. Furthermore, it's also interesting to note that combining all modalities never yielded better performance that using just a single modality. This may be a result of noise arising from combining multiple modalities, but is worth further exploring because this social interactions as perceived by humans include all modalities used in these experiments.

This observation indicates that firstly, the contribution of the self-reported scores predicted by GPT-4 do appear to help the prediction of dyadic engagement. Secondly, the way in which the late fusion of different modalities are performed need to be changed or improved through hyperparameter optimization or changing the architecture all together.

Chapter 6

Conclusion

6.1 Potential Applications

In addition to using LLMs to simulate human participants in future research, the opened ended responses of LLMs could provide researchers access to more meaningful and explainable results in studies. For example, when `GPT-4` was unconstrained when responding to *My conversation partner appreciated my points, even if we disagreed.*, one generated response was *I would rate it a 7. Even though we come from different academic backgrounds and have had different experiences, I felt that my conversation partner was genuinely interested in my responses and shared her perspective openly.* During the time of writing, LLMs are unable to reason about their generations, so directly asking for explanations now yields marginal value. However, it's likely that future models will be more interpretable and, therefore, asking the LLM for explanations can provide meaningful insight. This offers an intriguing means to provide researchers with better insight into their studies. Consider the possibility of being able to ask for possible reasons a human participant answered a particular way in a study whenever a researcher desires. Future experiments could include including recording participants' reasoning about each response, and comparing the alignment between the LLM "reasoning" each participant. This could provide deeper insight into how well the textual model is capturing outcomes.

6.2 Limitations

The first main limitation is the size of the dataset. 17 and 15 dyadic interactions for the two experiments is still a very limited number of recorded interactions for a model to learn from and to generalize to unseen dyads, and to make strong claims about the ability of LLMs to infer the mental states of participants in real social interactions. However, given that the LLM Fusion method achieved such performance in a zero-shot manner, this is worth further exploration.

Given the limited size and variance in demographics of our participants and engagement experiences within our dataset, it also raises the question of how well LLMs can simulate engagement questionnaire responses for different populations and conversational experiences. It's also possible that a person's responses to the Big Five Inventory and belief questionnaire may not accurately reflect their true personality and beliefs.

It is also crucial to acknowledge the limitations and biases associated with the models used. LLMs inadvertently learn and incorporate positional, racial, gender, and other social biases [12, 51, 77, 78]. They are also sensitive to the wording of provided prompts. Furthermore, given that our multimodal transcript relies on pre-trained models such as OpenFace, MediaPipe, and Whisper, possible issues of bias and robustness in those models [26, 50] should also be taken into consideration. Additional noise may be created from the usage of multiple pre-trained models. The ability of the multimodal transcript to accurately represent the conversation is inherently limited by the accuracy of the pre-trained models used.

LLMs such as `GPT-4` have been fine-tuned with RLHF to produce responses that are safer and better aligned with the user's intent. While this process reduces response toxicity and improves the ability to follow instructions, we note that this calibration may interfere with the ability of the LLM to emulate human-like responses in a research setting.

6.3 Conclusion

Engagement is fundamental to all human interactions, representing the intrinsic interest or emotional investment of the individuals involved. Despite humans' intuitive understanding of engagement, developing computational systems capable of recognizing and measuring engagement remains a significant challenge. Our work studies this core element of communication through smart glasses worn by participants in natural conversation. We collected a dataset of casual conversations between pairs of strangers, each outfitted with a pair of smart glasses, to capture behavioral cues such as facial expressions, eye contact, and verbal exchanges. We introduce a novel fusion method using large language models (LLMs) to include behavioral features, generating a "multimodal transcript" of the conversation to prompt an LLM to predict the participants' self-reported engagement levels. Our work is one of the first to use language to "reason" about real world human behavior, laying the groundwork for promising directions for future research in computational behavior analysis. Finally, we explore the usage of features derived from `GPT-4`'s prediction of first-person engagement to aid the prediction of third-party engagement ratings.

There are a few key directions for future work. The first direction involves developing a more comprehensive dataset of dyadic engagement and interactions, which contain a more diverse range of experiences. A path could involve recording dyads engaging in a game to gather more negative experiences. The second direction concerns research on studying and improving the social intelligence of AI systems. We encourage future research on the social intelligence of AI systems to benchmark performance with respect to real human interactions in addition to existing static and virtual benchmarks. While existing benchmarks are necessary, they are only proxies for understanding human behavior in-the-wild, which is the setting these systems will inevitably be deployed in. Finally, the third direction involves exploring this paradigm of using foundation models to extract abstract features such as ones relating to behavior in order to augment or supplement raw modalities. The rapid development of foundation models raises the question of how can we utilize them to supplement existing or create new models that handle novel modalities. Exploring this path could prove to be a fruitful avenue for future research in the field of affective computing, multimodal machine learning, Psychology, and a wide range of other domains.

Appendix A

Self Report Engagement Questionnaire

The following questionnaire was completed by each participant at the end of the recording session. Also displayed is the distribution of responses received in our participant sample; red rows indicate negatively-coded items.

Figure A.1: Participants completed the questionnaire described above

Appendix B

LLM Fusion: Multimodal Transcript Template

This appendix contains a detailed version of the sample multimodal transcript depicted in Figure 1.1. Magenta text corresponds to information from personal inventories. Red text corresponds to information from OpenFace. Violet text corresponds to information from MediaPipe and Pupil Invisible eye tracking. Blue text corresponds to information from the Whisper transcription. Green text corresponds to information from the post-session engagement survey. Black text is always present. The last row with "assistant" is what the LLM generates.

Role	Content
System	You are a student at ... You are participating in a psychology study that aims to understand how people communicate, and you are participating in a conversation with ... as part of this study. There will be a questionnaire at the end of this conversation. Others will read what you answer; your goal is to convince them it was answered from the perspective of the persona that participated in the following conversation. Your personality traits are defined by the scores to the following statements. The scores range from 1 to 5, where 1 means strongly disagree and 5 means strongly agree. [Alice's personality defined by responses to the big-5 personality survey.] Your political beliefs are defined by the following statements: [Alice's beliefs defined by responses to the beliefs survey.]
Assistant	[You] [You are looking at your partner's face about 80% of the time. You are speaking with a smiling mouth, raised cheeks... Your partner is looking at your face about 80% of the time. Your partner is listening with relaxed facial expression...] Hi, I'm Alice! What year are you?
User	[Partner] [You are looking at your partner's face about 60% of the time. You are listening with a smiling mouth, raised cheeks... Your partner is looking at your face about 80% of the time. Your partner is speaking with a smiling mouth, raised cheeks...] Hi Alice, I'm Bob. I'm a sophomore.
	[*five minutes of conversation*]
User	[Experimenter] On a scale of 1 to 7, where 1 means strongly disagree and 7 means strongly agree, how would you rate the following statement given the conversation you just had? I found this conversation to be interesting. Your answers will be kept private and your conversation partner will not see the responses, so please be as honest as possible. Provide your answer in the form of an integer between 1 and 7.
Assistant	7

Appendix C

LLM Fusion: Non-Numeric Responses

There were 80 cases across all ablation experiments where the first generated token returned by the LLM was non-numeric (1.1% of all responses). For 57 of these questions, `GPT-4` declined to respond because it was prompted to answer questions that referenced modalities not explicitly provided in the ablation feature sets. Sample responses are included below for illustration.

Q8: *My partner tended to make eye contact during the conversation.*

- "`As an AI, I don't have the ability to make or observe eye contact.`"
- "`As this conversation was text-based and not in person, I cannot accurately rate my partner's tendency to make eye contact. Therefore, I would rate it as N/A.`"
- "`As this conversation was text-based and didn't involve any form of video or face-to-face interaction, I cannot rate my partner's eye contact. Therefore, I would rate it as N/A.`"

Q47: *My conversation partner made a lot of eye contact.*

- "`As an AI, I don't have the ability to make or perceive eye contact. Therefore, I can't provide a rating for this statement.`"
- "`As an AI, I don't have the ability to make or perceive eye contact. Therefore, I would rate this statement as 1, strongly disagree.`"
- "`Given the nature of the conversation, it's hard to determine the level of eye contact as it was a text-based interaction. However, if we consider the level of engagement and attentiveness as a form of "eye contact" in this context, I`" [*text cut off*]

	Token	Prob.		Token	Prob.
1	`As`	0.316	11	`Sorry`	0.002
2	`[`	0.283	12	`Because`	0.002
3	`Since`	0.214	13	`The`	0.001
4	`I`	0.104	14	`5`	0.001
5	`Given`	0.042	15	`4`	0.001
6	`Considering`	0.007	16	`It`	0.001
7	`This`	0.007	17	`Without`	0.001
8	`Unfortunately`	0.004	18	`N`	0.001
9	`Ap`	0.003	19	`3`	0.001
10	`Due`	0.003	20	`My`	0.001

Table C.1: Sample top 20 tokens from a questionnaire response by the LLM where the first response is non-numeric.

For example, consider the following response to Q8: "`As this conversation was text-based, I cannot provide a rating for eye contact`". A sample of the top 20 tokens with highest probability are displayed in Table C.1.

The other 23 responses exceeded 50 generated tokens and were cut off. This occurred often in the $4F$ ablation experiments when the `GPT-4` would prefix its answers with the facial expression string, such as the following example.

"`[You] [You are speaking mostly with relaxed facial muscles, a straight mouth, a smooth forehead, and unremarkable eyebrows. Your partner is listening to you mostly with relaxed facial muscles, a straight mouth, a smooth forehead, and unremark`" [*text cut off*].

It's interesting to note that not all `GPT` models are able to impersonate a participant. For example, nearly all experiments with `gpt-4-1106-preview` would result in an example similar to the following:

"`As an AI language model, I don't have personal experiences or opinions. However, if I were to simulate a response for the scenario described where a participant has engaged in an interesting conversation that touched on computer science, philosophy of neuroscience, differences between cities, and personal experiences, they might rate the conversation on the higher end of the scale indicating that they found it to be engaging and intellectually stimulating.`"

Bibliography

[1] Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pages 337–371, Honolulu, Hawaii, USA, July 2023. JMLR.org. 2.3

[2] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351, July 2023. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2023.2. 2.3

[3] Meghan J. Babcock, Vivian P. Ta, and William Ickes. Latent Semantic Similarity and Language Style Matching in Initial Dyadic Interactions. *Journal of Language and Social Psychology*, 33(1):78–88, January 2014. ISSN 0261-927X. doi: 10.1177/0261927X13499331. 2.3.1

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 2.2

[5] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *Proceedings of the Thirteenth IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66, May 2018. doi: 10.1109/FG.2018.00019. 3.6.1, 5.1

[6] Atef Ben-Youssef, Chloé Clavel, Slim Essid, Miriam Bilac, Marine Chamoux, and Angelica Lim. UE-HRI: A new dataset for the study of user engagement in spontaneous human-robot interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI '17, pages 464–472, New York, NY, USA, November 2017. Association for Computing Machinery. ISBN 978-1-4503-5543-8. doi: 10.1145/3136755.3136814. 1

[7] Carlos Busso, Murtaza Bulut, Chi Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4), 2008. doi: 10.1007/s10579-008-9076-6. 1

[8] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. The NoXi database: Multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM*

International Conference on Multimodal Interaction, ICMI '17, pages 350–359, New York, NY, USA, November 2017. Association for Computing Machinery. ISBN 978-1-4503-5543-8. doi: 10.1145/3136755.3136780. 1

[9] Patrick C. Carmody, Julio C. Mateo, Drew Bowers, and Mike J. McCloskey. Linguistic Coordination as an Unobtrusive, Dynamic Indicator of Rapport, Prosocial Team Processes, and Performance in Team Communication. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 61(1):140–144, September 2017. ISSN 1071-1813. doi: 10.1177/1541931213601518. 2.3.1

[10] Mehmet Celepkolu and Kristy Elizabeth Boyer. Predicting student performance based on eye gaze during collaborative problem solving. In *Proceedings of the Group Interaction Frontiers in Technology*, GIFT'18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360777. doi: 10.1145/3279981.3279991. URL `https://doi.org/10.1145/3279981.3279991`. 3.6.2

[11] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. Multimodal Human-Human-Robot Interactions (MHHRI) Dataset for Studying Personality and Engagement. *IEEE Transactions on Affective Computing*, 10(4):484–497, October 2019. ISSN 1949-3045, 2371-9850. doi: 10.1109/TAFFC.2017.2737019. 1

[12] Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.84. 6.2

[13] Samuel Rhys Cox, Ashraf Abdul, and Wei Tsang Ooi. Prompting a Large Language Model to Generate Diverse Motivational Messages: A Comparison with Human-Written Messages. In *Proceedings of the 11th International Conference on Human-Agent Interaction*, HAI '23, pages 378–380, New York, NY, USA, December 2023. Association for Computing Machinery. ISBN 9798400708244. doi: 10.1145/3623809.3623931. 2.3.1

[14] Ronen Cuperman and William Ickes. Big Five predictors of behavior and perceptions in initial dyadic interactions: Personality similarity helps extraverts and introverts, but hurts "disagreeables". *Journal of Personality and Social Psychology*, 97(4):667–684, 2009. ISSN 1939-1315. doi: 10.1037/a0015741. 3.4

[15] Ronen Cuperman and William Ickes. Big five predictors of behavior and perceptions in initial dyadic interactions: Personality similarity helps extraverts and introverts, but hurts "disagreeables". *Journal of personality and social psychology*, 97(4):667, 2009. 2.1.1

[16] Jared R. Curhan and Alex Pentland. Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology*, 92(3): 802–811, 2007. ISSN 1939-1854. doi: 10.1037/0021-9010.92.3.802. 2.2, 4.1.2

[17] Dorottya Demszky, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margarett Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, Michaela Jones, Danielle Krettek-Cobb, Leslie Lai, Nirel JonesMitchell,

Desmond C. Ong, Carol S. Dweck, James J. Gross, and James W. Pennebaker. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701, November 2023. ISSN 2731-0574. doi: 10.1038/s44159-023-00241-5. 2.3

[18] Soumia Dermouche and Catherine Pelachaud. Engagement modeling in dyadic interaction. In *2019 international conference on multimodal interaction*, pages 440–445, 2019. 2.1, 2.1.1, 2

[19] Paul Ekman and Wallace V. Friesen. Facial Action Coding System, 1978. 4.1.1

[20] Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. MEISD: A Multimodal Multi-Label Emotion, Intensity and Sentiment Dialogue Dataset for Emotion Recognition and Sentiment Analysis in Conversations. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4441–4453, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.393. 1

[21] Jennifer A. Fredricks and Wendy McColskey. *The Measurement of Student Engagement: A Comparative Analysis of Various Methods and Student Self-report Instruments*, pages 763–782. Springer US, Boston, MA, 2012. ISBN 978-1-4614-2018-7. doi: 10.1007/978-1-4614-2018-7_37. URL https://doi.org/10.1007/978-1-4614-2018-7_37. 2.1.1

[22] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind One Embedding Space to Bind Them All. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '23)*, pages 15180–15190, June 2023. doi: 10.1109/CVPR52729.2023.01457. 2.3.1

[23] Patricia Goldberg, Ömer Sümer, Kathleen Stürmer, Wolfgang Wagner, Richard Göllner, Peter Gerjets, Enkelejda Kasneci, and Ulrich Trautwein. Attentive or not? toward a machine learning approach to assessing students' visible engagement in classroom instruction. *Educational Psychology Review*, 33:27–49, 2021. 2.1.1

[24] Charles Goodwin. *Conversational Organization: Interaction Between Speakers and Hearers*. Language, Thought, and Culture. Academic Press, New York, 1981. ISBN 978-0-12-289780-1. 2.2

[25] Guendalina Graffigna, Serena Barello, Andrea Bonanomi, and Edoardo Lozza. Measuring patient engagement: development and psychometric properties of the patient health engagement (phe) scale. *Frontiers in psychology*, 6:132788, 2015. 2.1.1

[26] Calbert Graham and Nathan Roll. Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2), 2024. 6.2

[27] Cameron C Gray and Dave Perkins. Utilizing early engagement and machine learning to predict student outcomes. *Computers & Education*, 131:22–32, 2019. 2.1.1

[28] Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L. Chaunzwa, Idalid Franco, Benjamin H. Kann, Shalini Moningi, Jack M. Qian, Madeleine Goldstein, Susan Harper, Hugo J. W. L. Aerts, Paul J. Catalano, Guergana K. Savova, Raymond H. Mak, and Danielle S.

Bitterman. Large language models to identify social determinants of health in electronic health records. *npj Digital Medicine*, 7(1):1–14, January 2024. ISSN 2398-6352. doi: 10.1038/s41746-023-00970-0. 2.3.1

[29] Andrew F. Hayes and Klaus Krippendorff. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1):77–89, April 2007. ISSN 1931-2458. doi: 10.1080/19312450709336664. 3.5

[30] Joey Chiao-yin Hsiao, Wan-rong Jih, and Jane Yung-jen Hsu. Recognizing continuous social engagement level in dyadic conversation by using turn-taking and speech emotion patterns. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012. 2.1

[31] Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara. Engagement recognition by a latent character model based on multimodal listener behaviors in spoken dialogue. *APSIPA Transactions on Signal and Information Processing*, 7:e9, 2018. 2.1

[32] Molly E. Ireland and James W. Pennebaker. Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of Personality and Social Psychology*, 99 (3):549–571, 2010. ISSN 1939-1315, 0022-3514. doi: 10.1037/a0020386. 2.3.1

[33] Molly E. Ireland, Richard B. Slatcher, Paul W. Eastwick, Lauren E. Scissors, Eli J. Finkel, and James W. Pennebaker. Language Style Matching Predicts Relationship Initiation and Stability. *Psychological Science*, 22(1):39–44, January 2011. ISSN 0956-7976, 1467-9280. doi: 10.1177/0956797610392928. 2.3.1

[34] Natasha Jaques, Daniel McDuff, Yoo Lim Kim, and Rosalind Picard. Understanding and predicting bonding in conversations using thin slices of facial expressions and body language. In *Intelligent Virtual Agents: 16th International Conference, IVA 2016, Los Angeles, CA, USA, September 20–23, 2016, Proceedings 16*, pages 64–74. Springer, 2016. 2.1.1

[35] Dongjun Kang, Joonsuk Park, Yohan Jo, and JinYeong Bak. From Values to Opinions: Predicting Human Behaviors and Stances Using Value-Injected Large Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP '23)*, pages 15539–15559, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.961. 2.3

[36] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):116:1–116:29, April 2023. doi: 10.1145/3579592. 2.3.1

[37] Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.890. 2.3

[38] Kshitiz Kumar, Chanwoo Kim, and Richard M Stern. Delta-spectral cepstral coefficients

for robust speech recognition. In *2011 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 4784–4787. IEEE, 2011. 3.6.4

[39] J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, March 1977. ISSN 0006-341X. doi: 10.2307/2529310. 4.2.1

[40] Tianlin Liu and Arvid Kappas. Predicting engagement breakdown in hri using thin-slices of facial expressions. In *Workshops at the thirty-second AAAI conference on artificial intelligence*, 2018. 2.1.1

[41] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A Framework for Building Perception Pipelines, June 2019. 3.6.1

[42] Robert R. McCrae and Paul T. Costa Jr. A Five-Factor theory of personality. In *Handbook of Personality: Theory and Research, 2nd Ed*, pages 139–153. Guilford Press, New York, NY, US, 1999. ISBN 978-1-57230-483-3. 3.4

[43] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana El Kaliouby. AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 3723–3726, San Jose California USA, May 2016. ACM. ISBN 978-1-4503-4082-3. doi: 10.1145/2851581.2890247. 4.1.1

[44] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Transactions on Affective Computing*, 3(1): 5–17, January 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2011.20. 1

[45] Ron Mokady, Amir Hertz, and Amit H. Bermano. ClipCap: CLIP Prefix for Image Captioning, November 2021. 2.3.1

[46] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, January 2019. ISSN 1949-3045. doi: 10.1109/TAFFC.2017.2740923. 4.2.1

[47] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=KJ5h-yfUHa`. 2.2

[48] V Naik and VV Kamat. Predicting engagement using machine learning techniques. 2018. 2.1.1

[49] Yukiko I. Nakano and Ryo Ishii. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10, page 139–148, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605585154. doi: 10.1145/1719970.1719990. URL

https://doi.org/10.1145/1719970.1719990. 3.6.2

[50] Shushi Namba, Wataru Sato, and Sakiko Yoshikawa. Viewpoint robustness of automated facial action unit detection systems. *Applied Sciences*, 11(23):11171, 2021. 6.2

[51] Roberto Navigli, Simone Conia, and Björn Ross. Biases in Large Language Models: Origins, Inventory, and Discussion. *Journal of Data and Information Quality*, 15(2):10:1–10:21, June 2023. ISSN 1936-1955. doi: 10.1145/3597307. 6.2

[52] Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. Engagement in Human-Agent Interaction: An Overview. *Frontiers in Robotics and AI*, 7:92, August 2020. ISSN 2296-9144. doi: 10.3389/frobt.2020.00092. 2.1, 2.1.1, 3.5

[53] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS '22)*, volume 35, pages 27730–27744, December 2022. 4.2.1

[54] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, pages 1–22, New York, NY, USA, October 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763. 2.3

[55] Arthur Pellet-Rostaing, Roxane Bertrand, Auriane Boudin, Stéphane Rauzy, and Philippe Blache. A multimodal approach for modeling engagement in conversation. *Frontiers in Computer Science*, 5, March 2023. ISSN 2624-9898. doi: 10.3389/fcomp.2023.1062342. 1, 2.2

[56] Pew Research Center. Beyond Red vs. Blue: The Political Typology. Technical report, Washington, DC, USA, November 2021. 3.4

[57] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL '19)*, pages 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. 1

[58] Shraman Pramanick, Aniket Roy, and Vishal M Patel. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3930–3940, 2022. 2.2, 5.1.2

[59] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, December 1978. ISSN 1469-1825, 0140-525X. doi: 10.1017/S0140525X00076512. 2.3

[60] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya

Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pages 28492–28518, Honolulu, Hawaii, USA, July 2023. JMLR.org. 3.6.3

[61] Carolyn Ranti, Warren Jones, Ami Klin, and Sarah Shultz. Blink Rate Patterns Provide a Reliable Measure of Individual Engagement with Scene Content. *Scientific Reports*, 10(1): 8267, May 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-64999-x. 2.2

[62] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG '13)*, pages 1–8, April 2013. doi: 10.1109/FG.2013.6553805. 1

[63] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense Reasoning about Social Interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. 2.3

[64] Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP '22)*, pages 3762–3780, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.248. 2.3

[65] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality Traits in Large Language Models, September 2023. 2.3

[66] Candace L Sidner, Christopher Lee, and Neal Lesh. Engagement when looking: behaviors for robots when collaborating with people. In *Diabruck: Proceedings of the 7th workshop on the Semantic and Pragmatics of Dialogue*, pages 123–130. Citeseer, 2003. 2.1

[67] Candace L Sidner, Christopher Lee, Cory D Kidd, Neal Lesh, and Charles Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164, 2005. 2.1

[68] Michelle R Simpson. Engagement at work: A review of the literature. *International journal of nursing studies*, 46(7):1012–1024, 2009. 2.1

[69] Elizabeth C. Stade, Shannon Wiltsey Stirman, Lyle H. Ungar, Cody L. Boland, H. Andrew Schwartz, David B. Yaden, João Sedoc, Robert J. DeRubeis, Robb Willer, and Johannes C. Eichstaedt. Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. *npj Mental Health Research*, 3(1): 1–12, April 2024. ISSN 2731-4251. doi: 10.1038/s44184-024-00056-z. 2.3

[70] Yla R. Tausczik and James W. Pennebaker. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54, March 2010. ISSN 0261-927X. doi: 10.1177/0261927X09351676. 2.3.1

[71] Julian Tejada, Raquel Meister Ko Freitag, Bruno Felipe Marques Pinheiro, Paloma Batista Cardoso, Victor Rene Andrade Souza, and Lucas Santos Silva. Building and validation of a set of facial expression images to detect emotions: A transcultural study. *Psychological Research*, 86(6):1996–2006, 2022. doi: 10.1007/s00426-021-01605-3. 4.1.1

[72] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. ZeroCap: Zero-Shot Image-to-Text Generation for Visual-Semantic Arithmetic. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '22)*, pages 17897–17907, June 2022. doi: 10.1109/CVPR52688.2022.01739. 2.3.1

[73] Marc Tonsen, Chris Kay Baumann, and Kai Dierkes. A High-Level Description and Performance Evaluation of Pupil Invisible, September 2020. 3.2, 3.3.1

[74] Lyn M. Van Swol and Aimée A. Kane. Language and Group Processes: An Integrative, Interdisciplinary Review. *Small Group Research*, 50(1):3–38, February 2019. ISSN 1046-4964. doi: 10.1177/1046496418785019. 2.3.1

[75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5.1

[76] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin. Canal9: A database of political debates for analysis of social interactions. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–4, September 2009. doi: 10.1109/ACII.2009.5349466. 1

[77] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.243. 6.2

[78] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large Language Models are not Fair Evaluators, August 2023. 6.2

[79] Ka Wong, Praveen Paritosh, and Lora Aroyo. Cross-replication Reliability - An Empirical Approach to Interpreting Inter-rater Reliability. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7053–7065, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.548. 4.2.1

[80] Zi Yang Wong and Gregory Arief D Liem. Student engagement: Current state of the construct, conceptual refinement, and future research directions. *Educational Psychology Review*, 34(1):107–138, 2022. 2.1

[81] Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023. doi: 10.1109/TPAMI.2023.3275156. 2.2

[82] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. In *The Eleventh International Conference on Learning Representations (ICLR '22)*, 2022. URL https://openreview.net/forum?id=G2Q2Mh3avow. 2.3.1, 4

[83] Zengqun Zhao and Ioannis Patras. Prompting Visual-Language Models for Dynamic Facial Expression Recognition. In *British Machine Vision Conference (BMVC '23)*, pages 1–14. arXiv, October 2023. 4.1.1

[84] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. In *The Twelfth International Conference on Learning Representations (ICLR '24)*. arXiv, March 2024. doi: 10.48550/arXiv.2310.11667. 2.3

[85] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. In *The Twelfth International Conference on Learning Representations (ICLR '23)*. arXiv, January 2024. doi: 10.48550/arXiv.2310.01852. 2.3.1