Multimodal Fusion with LLMs for Engagement Prediction in Natural Conversation

Cheng Ma*
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
ccma@cs.cmu.edu

Kevin Hyekang Joo*
Institute for Creative Technologies
University of Southern California
Los Angeles, California, USA
Thomas Lord Department of Computer Science
University of Southern California
Los Angeles, California, USA
khjoo@usc.edu

Alexandria K. Vail*
The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
avail@cs.cmu.edu

Sunreeta Bhattacharya
The Neuroscience Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
sunreetb@andrew.cmu.edu

Álvaro Fernández García
The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
alvarof@andrew.cmu.edu

Kailana Baker-Matsuoka
Department of Electrical Engineering
Stanford University
Stanford, California, USA
kailana@stanford.edu

Sheryl Mathew
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
sherylm@andrew.cmu.edu

Lori L. Holt
Department of Psychology
University of Texas at Austin
Austin, Texas, USA
lori.holt@austin.utexas.edu

Fernando De La Torre
The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
ftorre@cs.cmu.edu

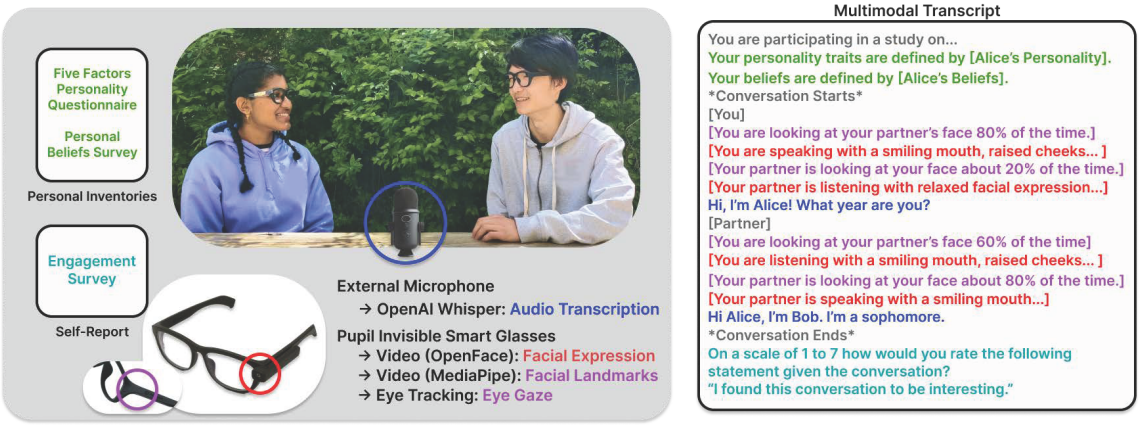

Figure 1: Visual representation of recorded behavior modalities during casual conversation and a sample of the multimodal transcript illustrating their fusion as introduced in this work. The goal is to predict engagement from this multimodal data. Color-coded modality names correspond to lines of the same color in the multimodal transcript.

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 enable the analysis of non-verbal behavior during natural face-to-face interactions. Our focus lies in predicting engagement in dyadic interactions by scrutinizing verbal and non-verbal cues, aiming to detect signs of disinterest or confusion.

*These authors contributed equally to this work.

ICMI Companion '25, Canberra, ACT, Australia

ACM ISBN 979-8-4007-2076-5/25/10
https://doi.org/10.1145/3747327.3764904

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 at the end of each conversation. 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. Remarkably, this method achieves performance comparable to established fusion techniques even in its preliminary implementation, indicating strong potential for further research and optimization. This fusion method is one of the first to approach "reasoning" about real-world human behavior through a language model. 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.

CCS Concepts

• **Human-centered computing** → Ubiquitous and mobile computing; *Collaborative and social computing*; • **Computing methodologies** → *Machine learning*.

Keywords

Engagement, Multimodal Machine Learning, Multimodal Fusion, Human Behavior, Affective Computing, Gaze Tracking

ACM Reference Format:

Cheng Ma, Kevin Hyekang Joo, Alexandria K. Vail, Sunreeta Bhattacharya, Álvaro Fernández García, Kailana Baker-Matsuoka, Sheryl Mathew, Lori L. Holt, and Fernando De La Torre. 2025. Multimodal Fusion with LLMs for Engagement Prediction in Natural Conversation. In *Companion Proceedings of the 27th International Conference on Multimodal Interaction (ICMI Companion '25), October 13–17, 2025, Canberra, ACT, Australia.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3747327.3764904

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 many potential applications, such as aiding navigation for the visually impaired or enhancing cues for those with difficulty reading nonverbal signals.

Although there has been substantial prior research in laboratory settings [8, 59, 71] and human-agent interaction [5, 11, 41], there are still many rich, unexplored opportunities in natural social contexts, for which smart glasses offer unique capabilities for study. Smart glasses enable the capture of social interactions in everyday settings, beyond the constraints of a lab, as people seek help, share, learn, and connect face-to-face. 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 is fundamental in maintaining group cohesion, preserving mental health, fostering academic learning, and supporting developmental growth.

Engagement has been recognized as a key determinant of communication success. While lacking a precise definition, engagement can be loosely defined as an individual's attentional and emotional investment during communication [51]. The ability to captivate in conversation can determine life-changing interactions, whether acing a job interview or making a favorable impression on a first date. The depth of our engagement and that of our partner shape 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. Engagement is conveyed through verbal and nonverbal cues—tone, expressions, gestures—or even silence and lack of gaze. Such engagement is hard to predict due to the dynamic and context-dependent nature of social exchanges. Thus, techniques that can 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 [41], MEISD [20], MELD [54], SEMPI [63], or NoXi [8], naturalistic dyadic interactions captured from an *egocentric* viewpoint are scarce. In the past few years, as smart glasses have become more widely accessible, research has begun to gather egocentric recordings for other tasks, such as skilled human activity (Ego-Exo4D, [25]) and user gaze anticipation [36], though less focused on interpersonal behavior. These factors challenge the development of socially-aware systems that respond authentically. 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 with communication disorders. The potential applications are extensive.

The contribution of the present work is twofold. We introduce a novel dataset including recordings of natural, unscripted conversations among unfamiliar dyads wearing the Pupil Invisible smart glasses with an *egocentric* camera built in, as illustrated in the left-hand segment of Figure 1. It contains conversations between 19 unique dyads, including video and audio recordings, gaze tracking, and self-reported information on demographic, political, and personality factors from the participants.

The second contribution presents an analysis of this dataset, focusing on predicting participant engagement levels through post-session self-reports. We compare audio-visual classical fusion techniques [75, 76] with our novel proposed fusion approach, which uses a large language model (LLM) as a reasoning engine to fuse behavioral measures of multiple modalities into a multimodal textual representation, a sample of which is displayed in the right-hand

segment of Figure 1. A fundamental methodological innovation in our approach involves creating a comprehensive "persona" for the LLM to simulate. Rather than merely incorporating additional modalities as supplemental feature sets, we explicitly integrate personality, beliefs, and behavioral data to form a cohesive textual representation. This persona-driven strategy enables the LLM to "act as" the participant by feeding on the textual representations of multimodal data, grounding its predictive reasoning in a simulated human perspective. This methodologically transparent approach not only enriches interpretability but also underscores the foundational significance of personality and belief dimensions in shaping engagement, rather than treating them merely as auxiliary data. Our results indicate that this approach achieves performance comparable to the established fusion techniques even in this early implementation. This approach is a powerful, simple, and flexible framework for future work on modeling human behavior and developing socially intelligent technologies.

2 Prior Work

2.1 Classical Fusion

Curhan and Pentland used speech features in the first 5 minutes of a simulated negotiation to predict the negotiation outcomes [15]. These features predicted 30% of the variance, demonstrating the value of speech in conversational dynamics. This suggests speech features are also important for predicting engagement. Activity level and mirroring had differing relationships with the 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 [51]. The study showed the value of visual and audio features, achieving the best results with the prosodic-acoustic and mimo-gestural modalities. Achieving similar results to studies using annotator-defined segments demonstrated that annotating engagement at a turn level can be effective. Others have also parallelly leveraged visual and linguistic features to detect engagement [11, 28, 31, 32].

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 [23, 58]. 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 [23]. 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 also 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 [58]. By closely analyzing the timing of blink inhibition in response to unfolding scene content, they found that they could uncover the viewers' unconscious, subjective evaluations of the importance and engagement level of what they observe. 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 presented to them.

2.2 Large Language Models (LLMs)

LLMs' accessibility has enabled many applications, especially in human-subject fields like psychology. They range from creating synthetic datasets of LLM-generated responses in human-less experiments [17] to providing automated feedback to clinicians [64].

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]. Tavast et al. evaluate the human-likeness of responses on the PANAS questionnaire generated by `GPT-3` [66]. The feasibility of using LLMs to replace human participants is further explored in [18, 27]. Park et al. introduce generative agents powered by LLMs that simulate believable human behavior in a virtual environment [50], also similarly seen in [79]. 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 concerning values [33, 62].

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 [55]. Prior work has proposed various benchmarks and methods to evaluate an agent's theory of mind [35, 60, 61].

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 human social interactions, such as predicting a person's responses to a survey that measures engagement. We argue that this dimension 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 in-the-wild social interactions. Given the potential of LLMs to advance socially intelligent technologies, incorporating in-the-wild social interactions into research is essential.

3 Data Set

We recorded dyadic conversations conducted in a controlled room setting, recorded from the viewpoint of each participant through a pair of smart glasses.

3.1 Population

Our study contained a total of 34 unique participants and 19 unique dyads. Within the participants, two participants appeared in multiple

dyads, but all dyads were unique. 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 a local university via 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

Each session lasted about 15 minutes, including introductions and closing. Each participant wore smart glasses (see subsection 3.3 for specifications) to capture 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 [69], so we ran a calibration procedure for each participant before the beginning of the session. Participants were encouraged to begin with a shared topic, COVID-19 experiences, but conversations were unconstrained. Following the session, participants completed questionnaires on their beliefs, personality, and engagement during this interaction (see subsection 3.4 for 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 smart glasses manufactured by Pupil Labs [69], 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 hinge 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 [69].

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 (Zoom H4N Pro) 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, 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 on previous studies on participant perception of interaction quality [14]: detailed statistics for the engagement questionnaire items are provided in Appendix A. The participants were also asked to complete the Big Five Inventory [39] for personality information and a handcrafted questionnaire on personal beliefs. This questionnaire was based on a set of socio-cultural issues studied to gauge polarization along the political spectrum [52].

4 Feature Extraction

Initially, we adjusted the video to eliminate the radial distortion introduced by the scene camera's lens. This was achieved by applying the distortion coefficients provided by the manufacturer [56]. Given 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.

4.1 Facial Expression

Facial action units (FAU) from the processed video were extracted with OpenFace 2.0 [4]. 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 requirement. For each frame, we used MediaPipe [38] (version: 0.9.1) 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 1080×1080px. If no face was detected in a particular frame, the location of the face in the previous frame was used. Interpolation was not used to fill missing frames, as our dataset rarely encountered either long sequences of missing frames or rapid movements that would necessitate interpolation.

4.2 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, 45]. 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. This adjustment aimed to account for the potential inaccuracies in the device's gaze prediction. The average error reported for the device assumes fine-tuned user calibration and ideal conditions (e.g., glasses worn correctly), neither of which were applicable to our recording setup [69].

4.3 Dialogue Transcription

OpenAI's Whisper [57] (version: large-v20230918) was used to transcribe the recording from each session. Whisper outputs fine-grained segments with start and stop times around a few seconds long. A speaker was assigned to each segment. If the segment contained speech from both speakers, the speaker who spoke the most was assigned. Diarization tools like PyAnnote [6, 53] and

source separation tools performed poorly with audio from our dataset, so manual labeling was chosen.

5 LLM Fusion

In this work, we explore the use of large language models (LLMs) to "reason" about a social interaction using multimodal information. Our method involves prompting an LLM to simulate a study participant and answer the end-of-session engagement questionnaire as though it were the participant themselves.

5.1 Socratic Models

Interpreting machine learning models is a well-known challenge. Typically, models encode behavioral features into a high-dimensional, abstract vector space, which is then mapped onto the target space. To understand a model's inner workings, we usually project these *intermediate* data into a space that is more understandable to humans, often through *visualization* techniques. However, consider the possibility of the inverse: rather than allowing the model to *obscure information – of multiple modalities – 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 the nuance and ambiguity inherent in these fields. Furthermore, controlling the input text also allows for privacy-sensitive adjustments.

Socratic Models, named for the ancient Greek philosopher's teaching method through cross-examination, use language to integrate information from a diverse set of modalities [77]. Within this framework, pre-trained models fine-tuned toward specific modalities or behaviors translate their interpretations of inputs into natural language. This output becomes a language prompt that guides the LLM's reasoning. 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 pre-trained models, each specialized in a distinct domain, to perform downstream multimodal tasks without 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 the capabilities of the zero-shot state-of-the-art ZeroCap [68]. However, when compared to fine-tuned models such as ClipCap [43], 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 [77].

This concept of "many-to-one" alignment has also been explored from other angles. 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 [80]. 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 [29, 30], and, as in the present work, engagement [3]. Various computational methods have been employed to extract this information from language, from bag-of-words approaches to neural network models [65, 70]. Recent advancements, however, have seen a considerable rise in LLMs adapted to augment tasks requiring social intelligence: notable applications have included refining persuasive communication for public health campaigns [13, 34] and identifying adverse social determinants of health within free-form clinical notes [26]. 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 [77].

5.2 Algorithms for LLM Fusion

The novel LLM fusion approach that we introduce enables an LLM to emulate a participant by creating a multimodal prompt: a dialogue transcript of the recording session augmented with textual representations of non-verbal behavior. These textual representations are formed from the data collected by the smart glasses, multiple pre-trained models, and personality questionnaires, but this method can be extended to contain any number of additional behavioral cues. We aim to evaluate whether this multimodal transcript effectively captures the dynamics of social interaction and can enable an LLM to predict self-reported engagement levels effectively. This work focuses on OpenAI's models `GPT-4` and `GPT-3.5`, but the technique could be applied to any LLM; we fixed at versions `GPT-4-0613` and `GPT-3.5-turbo-0613` for consistency.

5.2.1 Modalities. As described in section 4, this analysis included information from speech, gaze, and facial expression modalities, given their straightforward translation into text form and their established significance in signaling engagement.

The **speech** modality serves as the foundation of the multimodal transcript: its representation consists of the dialogue transcript augmented with speaker-labeled segments as described in subsection 4.3. The **gaze** modality is represented by a string indicating the proportion of time a speaker's gaze remains on their partner's face, rounded to the nearest 10% for brevity.

The **facial expression** modality is represented by a text description of the dominant emotional expression for each speaker-labeled segment of the recording following the methods of existing research [67] and applications (iMotion's Affectiva; [40]), these emotional expressions were defined by the facial action units measured by OpenFace 2.0: *happy*, *sad*, *surprise*, *fear*, *anger*, *disgust*, *contempt*, or *neutral* [19]. The *neutral* label was assigned if none of these labels were applicable. The emotional labels were translated into text as described by Zhao and Patras, which was generated by prompting ChatGPT, achieving state-of-the-art performance on the Dynamic Facial Expression Recognition problem [78].

Participant responses to the **personality** and **beliefs** questionnaires were also included as part of the *system message* [42], providing additional speaker-specific context, as personal characteristics are known to affect a person's social behavior [11].

5.2.2 Multimodal Transcript Generation. The messages provided to GPT use the discrete segments in 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 segment of speech forms the basis of a message provided to OpenAI's ChatCompletion API [48]. For each message, the *role* is assigned to the *assistant* if the segment is spoken by the simulated participant or to the *user* if spoken by the partner. The final *user* message is always a questionnaire item introduced by the "experimenter" (see Appendix A for questionnaire details). The final *assistant* message is generated by GPT as a response to the introduced questionnaire item. Prompted transcripts were truncated to five minutes, as previous literature has established that the first five minutes of a conversation is enough information for humans to predict its outcome successfully [15]. This limitation brought the added benefit of reducing the cost of the experiment.

6 Experiments

We conducted two experimental series to predict engagement based on the post-session questionnaires, outlined in subsection 3.4. The first series assessed multimodal fusion using classical models (subsection 6.1), while the second series used LLMs (subsection 6.2).

These models were evaluated against two baselines: a static mean prediction and a Bayesian regression model. The static mean prediction serves as a simple benchmark by predicting the average engagement score across all participants, while the Bayesian regression model is a probabilistic approach that updates prior beliefs with data to provide distributions of model estimates and predictions. The performance of these baselines is shown in Table 1.

6.1 Classical Fusion

Five standard machine learning techniques were employed to establish a comparative baseline: k-nearest neighbors (KNN), support vector machines (SVM), random forests (RF), bidirectional long short-term memory networks (Bi-LSTM), and multi-layer perceptrons (MLP). Each model was trained using per-turn behavioral features alongside the corresponding self-report ratings for each session, described in subsection 3.4. The KNN, SVM, and RF models implemented either the multivariate sequence kernel or the global alignment kernel (GAK; [16]) to facilitate the comparison of sequences of varying lengths, as these models are not inherently designed to process sequential or variable-length input. Conversely, the MLP and Bi-LSTM models followed canonical architectures specific to their respective methodologies.

The representations of the behavioral features provided to these models were designed to reflect the information presented to the LLM in subsection 6.2. Facial expression was denoted by a label indicating the predominant perceived emotion, while gaze direction was quantified as the proportion of time an individual directed their attention towards their partner's face. These representations parallel the descriptions provided to the LLM via the multimodal transcript. Dialogue text was encoded using sentence embeddings generated through the SimCSE framework [21]. For additional information on the extraction of these features, refer to section 4.

Models were evaluated with leave-one-dyad-out cross-validation. To detail: one session was allocated as the test set, while the remaining 16 sessions served as the training set. Within the training set, hyperparameters were optimized through 16 cross-validation folds. The final performance metrics were derived from the held-out dyad #17. This process was systematically repeated for each of the 17 dyads, guaranteeing that every dyad was used as the test set exactly once. Despite the small dataset, this ensured robust evaluation and mitigated overfitting.

As in the evaluation of the large language models (LLMs) in subsection 6.2, each model was trained using all three input modalities, as well as through an ablation study involving various subsets of these modalities, detailed in Table 1. The results suggest that the Support Vector Machine (SVM) achieved the best performance across the majority of subsets, with the Random Forest (RF) model closely following. While these two models outperformed the LLM variants, they were the only models to do so: the remaining three models generally underperformed compared to the LLM variants. Nearly all models outperformed the static mean and Bayesian regression baselines, except for the MLP model: this underperformance may be attributed to the MLP's tendency to overfit due to the limited size of the dataset.

6.2 LLM Fusion

GPT-4 was provided with the multimodal transcript paired with each of the survey items of the engagement questionnaire. Note that a few items on the questionnaire explicitly reference laughing or eye contact: despite not providing the model with explicit information on these behaviors, we included these items to explore the capability of the model to infer these behaviors with limited information.

We performed a set of ablation experiments to explore the significance of various feature sets, notated as follows:

- **4**: This model was given the raw dialogue transcription alone.
- **S**: The transcription is preceded by participant survey responses to the personality and beliefs questionnaires as (**S**)ystem instructions.
- **G**: The transcription is enhanced with descriptions of each participant's (**G**)aze behavior during each speaking turn.
- **F**: The transcription is enhanced with descriptions of each participant's (**F**)acial expression during each speaking turn.

Table 1: Prediction performance of classical vs. LLM fusion models when provided data from a limited set of modalities: RMSE mean and standard deviation across validation folds (lower is better). LLM-4/4S refers to ablations with GPT-4. A static mean prediction baseline performs at 1.913 (1.075).

First-Person Ratings (subsection 3.4) — lower RMSE scores are better ↓

Behavior Features	BayesReg	KNN	SVM	RF	Bi-LSTM	MLP	LLM-4	LLM-4S
Gaze-Only	1.637 (0.427)	1.556 (0.313)	**1.281 (0.310)**	1.355 (0.298)	1.588 (0.340)	1.881 (0.360)	—	—
Face-Only	1.610 (0.439)	1.530 (0.322)	**1.328 (0.333)**	1.390 (0.288)	1.563 (0.353)	2.090 (0.454)	—	—
Text-Only	1.591 (0.391)	1.512 (0.287)	**1.280 (0.309)**	1.301 (0.287)	1.478 (0.358)	2.069 (0.432)	1.669 (0.396)	**1.376 (0.381)**
Face + Gaze	1.578 (0.401)	1.500 (0.294)	**1.296 (0.339)**	1.314 (0.339)	1.517 (0.331)	1.833 (0.389)	—	—
Text + Gaze	1.638 (0.390)	1.557 (0.286)	**1.291 (0.305)**	1.409 (0.289)	1.466 (0.307)	1.988 (0.405)	1.418 (0.394)	**1.338 (0.378)**
Text + Face	1.600 (0.394)	1.521 (0.289)	**1.287 (0.337)**	1.305 (0.339)	1.572 (0.352)	1.945 (0.475)	1.477 (0.425)	**1.368 (0.417)**
Text + Face + Gaze	1.638 (0.409)	1.557 (0.300)	1.327 (0.342)	**1.303 (1.290)**	1.592 (0.355)	1.773 (0.356)	1.442 (0.423)	**1.364 (0.387)**

In three instances, the length of the multimodal transcript with added descriptions exceeded the input constraints of GPT-4 (two for **4SGF** and one for **4GF**); in these cases, the transcript was truncated. A t-test comparing the residuals of the truncated sessions with those of the non-truncated sessions yielded p-values of 0.186 and 0.648, suggesting no significant difference between the two groups. Future studies may benefit from exploring the impact of different truncation lengths and the ability of the technique to perform with shorter observation times.

The *temperature* parameter was set at 0 to ensure sampling from the most likely responses to the questionnaire. In cases where GPT-4 did not provide a numeric response, we selected the highest-likelihood numeric response from the top 20 generations for the first output token (see Appendix C).

6.2.1 LLM Fusion Results. We evaluated this technique through two labeling tasks: predicting participants' exact responses and predicting the valence/arousal of their responses. An "exact" response refers to the participant's original numeric rating (1–7). The valence/arousal model categorizes responses based on emotional dimensions: valence is defined as the positive or negative degree of emotion (e.g., pleasure), and arousal is defined as the intensity of emotion (e.g., high) [44]. We define valence in terms of the "disagree" range, a score of 1 ("strongly disagree") through 3 ("slightly disagree"), a neutral score of 4 ("neither agree nor disagree"), or the "agree" range, a score of 5 ("slightly agree") through 7 ("strongly agree"). Arousal is calculated as the distance of the participants' rating from the neutral score of 4, i.e., $|\text{response} - 4|$.

Exact Response As seen in Table 2, GPT-4's zero-shot performance of this technique is comparable to the baseline and classical early fusion models, evaluated via RMSE. Krippendorff's alpha metric, used to assess the reliability of agreement between multiple raters, indicated a moderate level of agreement between the model's predictions and the participants' responses, ranging within $[0.470, 0.543]$ across questions [37, 74]. This suggests that while the zero-shot technique may not outperform the more advanced models, it still holds potential for applications where computational resources are limited. Furthermore, the findings highlight the importance of evaluating various methodologies in diverse contexts, as different tasks may yield varying levels of effectiveness. Results demonstrated significant improvement over the baseline models, with GPT-4 matching the performance of classical models, suggesting that the multimodal transcript approach may serve as a viable alternative to conventional fusion methods. We also experimented with GPT-3.5; however, given its significantly poorer performance against GPT-4, we chose to exclude its results from further analysis.

Valence When restricting the labeling task to valence only, GPT-4 predictions agree statistically significantly with the study's participants. As presented in Table 2, all Krippendorff's alpha scores fall within an error interval of $[0.61, 0.80]$ [37].

Upon closer inspection of the valence predictions of the LLM-4S ablation model (which achieved the strongest performance in labeling exact responses; see Appendix B), we observe that GPT-4 reliably labels participant's "agree" responses, with a balanced accuracy of 91.8%. However, GPT-4 is less reliable in predicting participants' "disagree" responses, achieving a balanced accuracy of 66.1%. Notably, GPT-4 performs significantly poorly in labeling participant "neutral" responses, with a balanced accuracy of 12.7%. Given that the label range is smaller – one possible value (4), as opposed to three values in "agree" (5, 6, 7) or "disagree" (1, 2, 3) ranges – poor performance may be expected. We conjecture that GPT-4's process of reinforcement learning from human feedback (RLHF) [49] to reduce toxicity may result in overly "positive" responses from GPT-4, inadvertently introducing bias against "negative" responses. Further study is needed to determine the extent and sources of this potential bias.

Arousal Across all ablations, GPT-4 performs poorly in labeling arousal, only marginally better than chance, given Krippendorff's alpha scores in the range of $[0.047, 0.071]$ (see Table 2). While GPT-4 appears able to predict the general attitude of the participant towards a questionnaire statement (valence), it cannot reliably determine the strength of the participant's feelings (arousal). We attribute this performance gap primarily to two intertwined factors: the inherent limitations of our textual representation of multimodal data and the subtle, continuous nature of arousal. Valence often aligns closely with explicit linguistic cues – words and phrases explicitly indicate positivity or negativity – making valence naturally suited for our language-based LLM framework. In contrast, arousal reflects emotional intensity and is conveyed more implicitly through continuous and subtle behavioral signals, including prosody and nuanced facial or physiological responses. Our method converts continuous behaviors into discrete, categorical textual descriptions (e.g., emotion labels), thereby losing subtle variations crucial for predicting arousal.

6.2.2 Contribution per modality. To study the impact of each behavior modality (described in 5.2.1), we conducted a two-tailed paired t-test of each model's residuals against those of the baseline. The results suggest that each modality group added to the **LLM-4** baseline provides a statistically significant positive contribution ($p < 0.05$) to model performance.

In contrast, the additional modalities worsen the performance of the **4S** baseline on labeling exact scores but improve the performance

on labeling valence. In a paired t-test of residuals comparing exact predictions, the addition of facial expression descriptions in the **4SF** and **4SGF** ablations worsened performance significantly ($p = 0.003$ and $p = 0.021$, respectively); however, gaze did not have a notable impact ($p = 0.164$).

6.2.3 Performance across individual survey questions. 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%).

Prediction performance on the questions about laughter and eye contact is relatively poor, addressing our earlier hypothesis regarding the ability of the model to infer this behavior. In general, while the transcript did not explicitly contain descriptions of laughter, `GPT-4` tends to respond with the assumption that laughter did occur. Although numerous caveats apply to these results, they generally reflect the opinions of our study's participants.

7 Conclusion

Engagement is fundamental to all human interactions, representing the intrinsic interest or emotional investment of the individuals involved. Despite our intuitive grasp of engagement, computationally measuring it remains challenging. 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 propose a novel fusion method using LLMs, generating a "multimodal transcript" of the conversation to prompt an LLM to predict the participants' self-reported engagement levels. While most research has focused on using raw behavioral data for prediction tasks, to the best of our knowledge, our work is the first to integrate LLMs with behavioral features by transforming these features into textual summaries. This approach not only achieves comparable results but also improves the interpretability and generalizability of the prediction model. Furthermore, this approach facilitates holistic modeling of affective states while enhancing privacy through its use of language-based representations that humans have control over.

However, it is 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, 47, 72, 73]. They are also sensitive to the wording of the 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 [24, 46] should also be taken into account. 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.

Table 2: Krippendorff's alpha scores, mean and standard deviation, for each ablation (higher is better).

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)

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.

In addition, while our fusion method aggregates modalities into a textual form, it does not yet explicitly model cross-modal incongruity – e.g., when facial expressions contradict verbal tone. Future work could explore whether LLMs can be prompted or fine-tuned to recognize such misalignments as predictive signals.

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.

Acknowledgments

This work is supported by the James F. McDonnell Foundation - Understanding Human Cognition #103771720223937.

Ethical Impact Statement

This study was approved by the IRB at Carnegie Mellon University. Participants were provided with information on the study's purposes and then asked to give informed consent for the recording of their behavior and the sharing of this data with other researchers for future analysis. Permissions were presented separately to ensure transparency. The study posed minimal risk beyond casual conversation. We acknowledge long-term concerns regarding bias in computer vision, LLMs, our own feature selection and encoding process. Though rooted in foundational tools, we recognize our role in shaping their impact. We are committed to addressing these biases and advancing fair, responsible affective computing.

References

[1] Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23, Vol. 202)*. Honolulu, Hawaii, USA, 337–371.

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

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

[4] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *Proceedings of the Thirteenth IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. 59–66. doi:10.1109/FG.2018.00019

[5] Atef Ben-Youssef, Chloé Clavel, Slim Essid, Miriam Bilac, Marine Chamoux, and Angelica Lim. 2017. 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)*. Association for Computing Machinery, New York, NY, USA, 464–472. doi:10.1145/3136755.3136814

[6] Hervé Bredin. 2023. Pyannote.Audio 2.1 Speaker Diarization Pipeline: Principle, Benchmark, and Recipe. In *INTERSPEECH 2023*. ISCA, 1983–1987. doi:10.21437/Interspeech.2023-105

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

[8] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The NoXi Database: Multimodal Recordings of Mediated Novice-Expert Interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI '17)*. Association for Computing Machinery, New York, NY, USA, 350–359. doi:10.1145/3136755.3136780

[9] Patrick C. Carmody, Julio C. Mateo, Drew Bowers, and Mike J. McCloskey. 2017. 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 (Sept. 2017), 140–144. doi:10.1177/1541931213601518

[10] Mehmet Celepkolu and Kristy Elizabeth Boyer. 2018. Predicting Student Performance Based on Eye Gaze During Collaborative Problem Solving. In *Proceedings of the Group Interaction Frontiers in Technology* (Boulder, CO, USA) *(GIFT'18)*. Association for Computing Machinery, New York, NY, USA, Article 7, 8 pages. doi:10.1145/3279981.3279991

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

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

[13] Samuel Rhys Cox, Ashraf Abdul, and Wei Tsang Ooi. 2023. 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)*. Association for Computing Machinery, New York, NY, USA, 378–380. doi:10.1145/3623809.3623931

[14] Ronen Cuperman and William Ickes. 2009. 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 (2009), 667–684. doi:10.1037/a0015741

[15] Jared R. Curhan and Alex Pentland. 2007. Thin Slices of Negotiation: Predicting Outcomes from Conversational Dynamics within the First 5 Minutes. *Journal of Applied Psychology* 92, 3 (2007), 802–811. doi:10.1037/0021-9010.92.3.802

[16] Marco Cuturi. 2011. Fast global alignment kernels. In *Proceedings of the 28th International Conference on International Conference on Machine Learning* (Bellevue, Washington, USA) *(ICML'11)*. Omnipress, Madison, WI, USA, 929–936.

[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. 2023. Using Large Language Models in Psychology. *Nature Reviews Psychology* 2, 11 (Nov. 2023), 688–701. doi:10.1038/s44159-023-00241-5

[18] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences* (2023).

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

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

[21] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6894–6910. doi:10.18653/v1/2021.emnlp-main.552

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

[23] Charles Goodwin. 1981. *Conversational Organization: Interaction Between Speakers and Hearers*. Academic Press, New York.

[24] Calbert Graham and Nathan Roll. 2024. Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. *JASA Express Letters* 4, 2 (2024).

[25] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. 2024. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19383–19400.

[26] 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. 2024. Large Language Models to Identify Social Determinants of Health in Electronic Health Records. *npj Digital Medicine* 7, 1 (Jan. 2024), 1–14. doi:10.1038/s41746-023-00970-0

[27] Jacqueline Harding, William D'Alessandro, NG Laskowski, and Robert Long. 2023. AI language models cannot replace human research participants. *Ai & Society* (2023), 1–3.

[28] Yuyun Huang, Emer Gilmartin, and Nick Campbell. 2016. Conversational Engagement Recognition Using Auditory and Visual Cues.. In *Interspeech*. 590–594.

[29] Molly E. Ireland and James W. Pennebaker. 2010. Language Style Matching in Writing: Synchrony in Essays, Correspondence, and Poetry. *Journal of Personality and Social Psychology* 99, 3 (2010), 549–571. doi:10.1037/a0020386

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

[31] Kevin Hyekang Joo. 2025. Modeling Social Dynamics from Multimodal Cues in Natural Conversations. In *Proceedings of the 27th International Conference on Multimodal Interaction* (Canberra, Australia) *(ICMI '25)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3716553.3750822

[32] Kevin Hyekang Joo, Zongjian Li, Yunwen Wang, Yuanfeixue Nan, Mina Kian, Shriya Upadhyay, Maja Mataric, Lynn Miller, and Mohammad Soleymani. 2025. Multimodal Behavioral Characterization of Dyadic Alliance in Support Groups. In *Proceedings of the 27th International Conference on Multimodal Interaction* (Canberra, Australia) *(ICMI '25)*. Association for Computing Machinery. doi:10.1145/3716553.3750818

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

[34] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. 2023. 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 (April 2023), 116:1–116:29. doi:10.1145/3579592

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

[36] Bolin Lai, Fiona Ryan, Wenqi Jia, Miao Liu, and James M Rehg. 2023. Listen to look into the future: Audio-visual egocentric gaze anticipation. *arXiv preprint arXiv:2305.03907* (2023).

[37] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (March 1977), 159–174. doi:10.2307/2529310 jstor:2529310

[38] 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. 2019. MediaPipe: A Framework for Building Perception Pipelines. doi:10.48550/arXiv.1906.08172 arXiv:1906.08172 [cs]

[39] Robert R. McCrae and Paul T. Costa Jr. 1999. A Five-Factor Theory of Personality. In *Handbook of Personality: Theory and Research, 2nd Ed.* Guilford Press, New York, NY, US, 139–153.

[40] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana El Kaliouby. 2016. 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*. ACM, San Jose California USA, 3723–3726. doi:10.1145/2851581.2890247

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

[42] Microsoft. [n. d.]. System Message. https://microsoft.github.io/Workshop-Interact-with-OpenAI-models/Part-2-labs/System-Message/

[43] Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. ClipCap: CLIP Prefix for Image Captioning. doi:10.48550/arXiv.2111.09734 arXiv:2111.09734 [cs]

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

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

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

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

[48] OpenAI. [n. d.]. Create chat completion. https://platform.openai.com/docs/api-reference/chat/create

[49] 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. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems (NeurIPS '22)*, Vol. 35. 27730–27744.

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

[51] Arthur Pellet-Rostaing, Roxane Bertrand, Auriane Boudin, Stéphane Rauzy, and Philippe Blache. 2023. A Multimodal Approach for Modeling Engagement in Conversation. *Frontiers in Computer Science* 5 (March 2023). doi:10.3389/fcomp.2023.1062342

[52] Pew Research Center. 2021. *Beyond Red vs. Blue: The Political Typology*. Technical Report. Pew Research Center, Washington, DC, USA.

[53] Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

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

[55] David Premack and Guy Woodruff. 1978. Does the Chimpanzee Have a Theory of Mind? *Behavioral and Brain Sciences* 1, 4 (Dec. 1978), 515–526. doi:10.1017/S0140525X00076512

[56] Pupil-Labs. [n. d.]. Recording format. https://docs.pupil-labs.com/invisible/data-collection/data-format/

[57] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23, Vol. 202)*. JMLR, Honolulu, Hawaii, USA, 28492–28518.

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

[59] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. 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)*. 1–8. doi:10.1109/FG.2013.6553805

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

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

[62] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality Traits in Large Language Models. doi:10.48550/arXiv.2307.00184 arXiv:2307.00184 [cs]

[63] Maksim Siniukov, Yufeng Yin, Eli Fast, Yingshan Qi, Aarav Monga, Audrey Kim, and Mohammad Soleymani. 2024. SEMPI: A Database for Understanding Social Engagement in Video-Mediated Multiparty Interaction. In *Proceedings of the 26th International Conference on Multimodal Interaction* (San Jose, Costa Rica) *(ICMI '24)*. Association for Computing Machinery, New York, NY, USA, 546–555. doi:10.1145/3678957.3685752

[64] 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. 2024. Large Language Models Could Change the Future of Behavioral Healthcare: A Proposal for Responsible Development and Evaluation. *npj Mental Health Research* 3, 1 (April 2024), 1–12. doi:10.1038/s44184-024-00056-z

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

[66] Mikke Tavast, Anton Kunnari, and Perttu Hämäläinen. 2022. Language models can generate human-like self-reports of emotion. In *27th International Conference on Intelligent User Interfaces*. 69–72.

[67] Julian Tejada, Raquel Meister Ko Freitag, Bruno Felipe Marques Pinheiro, Paloma Batista Cardoso, Victor Rene Andrade Souza, and Lucas Santos Silva. 2022. Building and Validation of a Set of Facial Expression Images to Detect Emotions: A Transcultural Study. *Psychological Research* 86, 6 (2022), 1996–2006. doi:10.1007/s00426-021-01605-3

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

[69] Marc Tonsen, Chris Kay Baumann, and Kai Dierkes. 2020. A High-Level Description and Performance Evaluation of Pupil Invisible. doi:10.48550/arXiv.2009.00508 arXiv:2009.00508 [cs]

[70] Lyn M. Van Swol and Aimée A. Kane. 2019. Language and Group Processes: An Integrative, Interdisciplinary Review. *Small Group Research* 50, 1 (Feb. 2019), 3–38. doi:10.1177/1046496418785019

[71] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin. 2009. Canal9: A Database of Political Debates for Analysis of Social Interactions. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. 1–4. doi:10.1109/ACII.2009.5349466

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

[73] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 9440–9450. https://aclanthology.org/2024.acl-long.511

[74] Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-Replication Reliability - An Empirical Approach to Interpreting Inter-rater Reliability. In *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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 7053–7065. doi:10.18653/v1/2021.acl-long.548

[75] Abudukelimu Wuerkaixi, Kunda Yan, You Zhang, Zhiyao Duan, and Changshui Zhang. 2022. DyViSE: Dynamic Vision-Guided Speaker Embedding for Audio-Visual Speaker Diarization. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP '22)*. 1–6. doi:10.1109/MMSP55362.2022.9948860

[76] Eric Zhongcong Xu, Zeyang Song, Satoshi Tsutsui, Chao Feng, Mang Ye, and Mike Zheng Shou. 2022. AVA-AVD: Audio-visual Speaker Diarization in the Wild. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*. Association for Computing Machinery, New York, NY, USA, 3838–3847. doi:10.1145/3503161.3548027

[77] 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. 2022. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. In *The Eleventh International Conference on Learning Representations (ICLR '22)*. https://openreview.net/forum?id=G2Q2Mh3avow

[78] Zengqun Zhao and Ioannis Patras. 2023. Prompting Visual-Language Models for Dynamic Facial Expression Recognition. In *British Machine Vision Conference (BMVC '23)*. https://papers.bmvc2023.org/0098.pdf

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

[80] 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. 2024. LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. In *The Twelfth International Conference on Learning Representations (ICLR '23)*. https://openreview.net/forum?id=QmZKc7UZCy

A 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.

Please use this 7-point rating scale to share your impressions of the conversation with your partner.

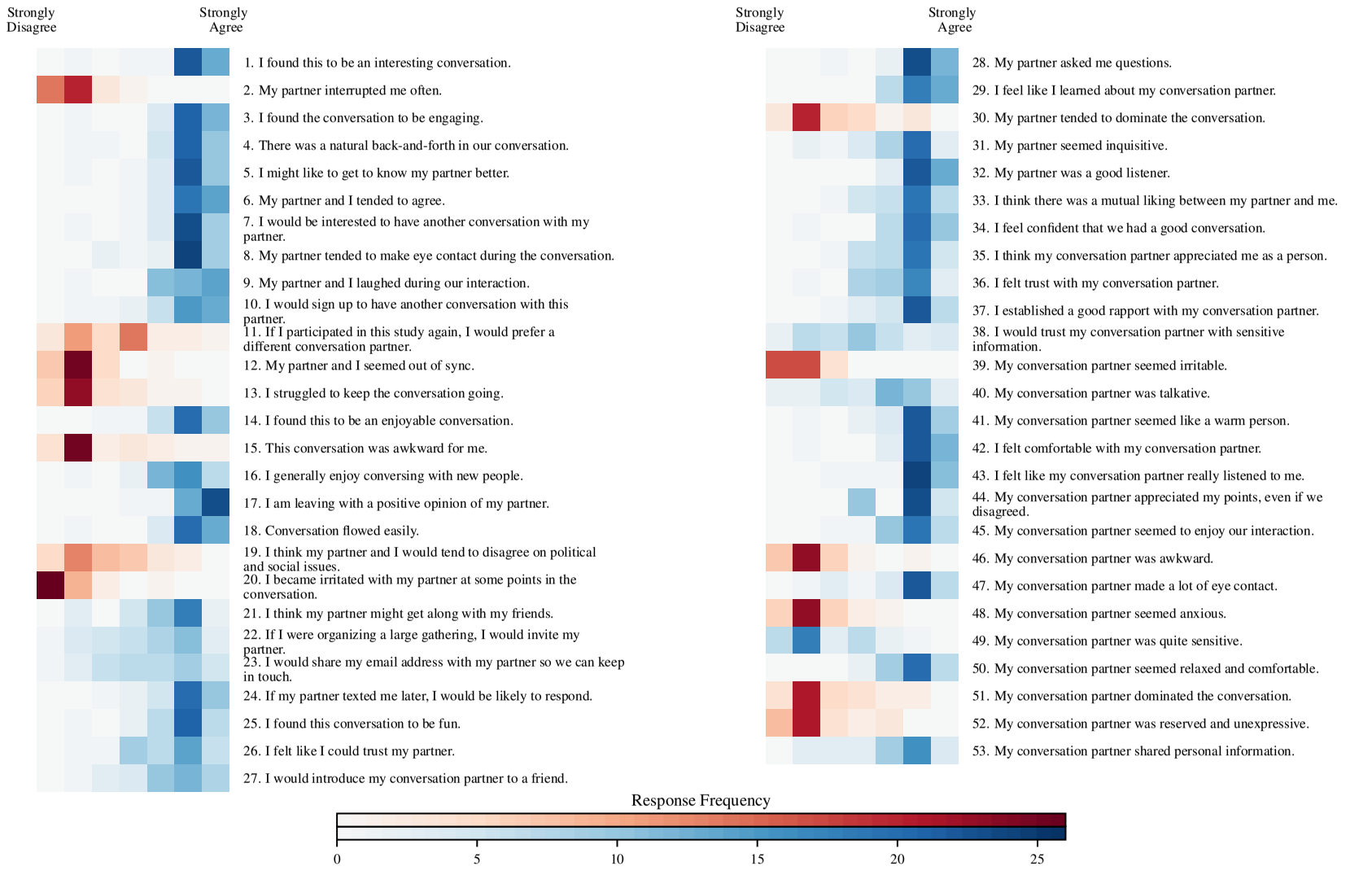

Table 3: 4S Valence Prediction Confusion Matrix: Responses are categorized as (Dis)agree (1–3), (Neu)tral (4), or (Agr)ee (5–7). Balanced accuracy is also reported.

	Predicted				
Actual	**Agr.**	**Neu.**	**Dis.**	**All**	**Bal. 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

B LLM Fusion: Valence Prediction

One of the experiments described in subsection 6.2 involves an evaluation of the LLM's ability to predict response valence rather than exact answers. Results for model **4S**, using raw transcripts and the participant personal characteristics, are presented in Table 3.

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*]

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 4.

Table 4: Sample top 20 tokens from a questionnaire response by the LLM where the first response is non-numeric.

	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

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.`"

D Multimodal Transcript Template

This appendix contains a detailed version of the sample multimodal transcript depicted in Figure 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

E Belief Questionnaire

Each participant completed the following questionnaire at the end of the recording session.

Please select the answer which most represents your beliefs.

Environmental Protection

- I am very much against environmental protection.
- I am against environmental protection.
- I am mildly against environmental protection.
- I am mildly in favor of environmental protection.
- I am in favor of environmental protection.
- I am very much in favor of environmental protection.

Careers for Women

- I am very much against women pursuing careers.
- I am against women pursuing careers.
- I am mildly against women pursuing careers.
- I am mildly in favor of women pursuing careers.
- I am in favor of women pursuing careers.
- I am very much in favor of women pursuing careers.

Belief in God

- I strongly believe that there is a God.
- I believe there is a God.
- I feel that perhaps there is a God.
- I feel that perhaps there is no God.
- I believe there is no God.
- I strongly believe there is no God.

Ranking of Schools

- I am very much against the ranking of schools.
- I am against the ranking of schools.
- I am mildly against the ranking of schools.
- I am mildly in favor of the ranking of schools.
- I am in favor of the ranking of schools.
- I am very much in favor of the ranking of schools.

Abortion

- I am very much against abortion.
- I am against abortion.
- I am mildly against abortion.
- I am mildly in favor of abortion.
- I am in favor of abortion.
- I am very much in favor of abortion.

Death Penalty

- I am very much against the death penalty.
- I am against the death penalty.
- I am mildly against the death penalty.
- I am mildly in favor of the death penalty.
- I am in favor of the death penalty.
- I am very much in favor of the death penalty.

Gay Marriage

- I am very much against gay marriage.
- I am against gay marriage.
- I am mildly against gay marriage.
- I am mildly in favor of gay marriage.
- I am in favor of gay marriage.
- I am very much in favor of gay marriage.

Money

- I strongly believe that money is one of the most important things in life.
- I believe that money is one of the most important things in life.
- I feel perhaps that money is one of the most important things in life.
- I feel perhaps that money is not one of the most important things in life.
- I believe that money is not one of the most important things in life.
- I strongly believe that money is not one of the most important things in life.

Divorce

- I am very much against divorce.
- I am against divorce.
- I am mildly against divorce.
- I am mildly in favor of divorce.
- I am in favor of divorce.
- I am very much in favor of divorce.

Smoking

- I am very much against smoking in public places like bars.
- I am against smoking in public places like bars.
- I am mildly against smoking in public places like bars.
- I am mildly in favor of smoking in public places like bars.
- I am in favor of smoking in public places like bars.
- I am very much in favor of smoking in public places like bars.

Spanking Children

- In general, I am very much in favor of spanking children.
- In general, I am in favor of spanking children.
- In general, I am mildly in favor of spanking children.
- In general, I am mildly against spanking children.
- In general, I am against spanking children.
- In general, I am very much against spanking children.

Climate Change

- I strongly believe that climate change has not been accelerated by humans.
- I believe that climate change has not been accelerated by humans.
- I mildly believe that climate change has not been accelerated by humans.
- I mildly believe that climate change has been accelerated by humans.
- I believe climate change has been accelerated by humans.
- I strongly believe that climate change has been accelerated by humans.

Health Care

- I strongly believe that humans are not entitled to health care.
- I believe that humans are not entitled to health care.
- I mildly believe that humans are not entitled to health care.
- I mildly believe that humans are entitled to health care.
- I believe that humans are entitled to health care.
- I strongly believe that humans are entitled to health care.

Social Safety Net

- I strongly believe the government should not provide funds to support individuals' welfare.
- I believe the government should not provide funds to support individuals' welfare.
- I mildly believe the government should not provide funds to support individuals' welfare.
- I mildly believe the government should provide funds to support individuals' welfare.
- I believe the government should provide funds to support individuals' welfare.
- I strongly believe the government should provide funds to support individuals' welfare.

College

- I strongly believe the government should not pay for college students' tuition.
- I believe the government should not pay for college students' tuition.
- I mildly believe the government should not pay for college students' tuition.
- I mildly believe the government should pay for college students' tuition.
- I believe the government should pay for college students' tuition.
- I strongly believe the government should pay for college students' tuition.

[Local University]

- I strongly believe that [local university] is a welcoming university environment.
- I believe that [local university] is a welcoming university environment.
- I mildly believe that [local university] is a welcoming university environment.
- I mildly believe that [local university] is not a welcoming university environment.
- I believe that [local university] is not a welcoming university environment.
- I strongly believe that [local university] is not a welcoming university environment.