CONVERSE: Annotation Scheme and Dataset for Multimodal Conversational Engagement Analysis in Human-Human and Human-Robot Interaction

Ekaterina Torubarova[1], **Oskar Ljung**[2], **Julia Uddén**[2], **André Pereira**[1]
[1]KTH Royal Institute of Technology, Stockholm, Sweden
[2]Stockholm University, Stockholm, Sweden
ekator@kth.se, oskarljung93@gmail.com, julia.udden@psychology.su.se, atap@kth.se

Abstract

Creating conversational agents that can both understand and respond appropriately to users' engagement remains a major challenge, as conversation is one of the most universal yet complex human behaviors. Modeling conversational engagement requires a fine-grained understanding of how engagement unfolds dynamically in interaction. This paper introduces a novel turn-based annotation scheme for conversational engagement, together with the CONVERSE dataset that contains annotations of 25 hours of unscripted human–human and human–robot conversations with 48 native Swedish speakers. This dataset uniquely utilizes such an annotation scheme for both human and robot agents within the same study, allowing for direct comparison. Notably, this dataset builds upon our previous multimodal corpus, which includes brain imaging (fMRI), eye-tracking, and speech data, as well as personality and stance measures. This dataset opens a new perspective on conversational engagement through these behavioral annotations and the existing neural data at the intersection of multimodal machine learning, human-robot interaction, and cognitive neuroscience.

Keywords: Annotation, Engagement, Human-Robot Interaction, Multimodal Dataset, Cognitive Neuroscience

1. Introduction

Conversation is one of the most natural yet cognitively complex social behaviors humans engage in. It requires constant coordination of attention, emotion, and intention between partners, processes that humans manage seamlessly but that computational models still struggle to capture and reproduce. As these models increasingly underpin conversational agents and tools for studying interaction, understanding how engagement unfolds in dialogue should be considered a central challenge. These agents, therefore, must be *engaging* for the user and *engaged* with the user. Thus, modeling engagement has been a common aim in human-agent interaction (Sidner et al., 2005). This line of research draws inspiration from human-human interactions in daily life, as we are typically able to effortlessly and automatically monitor our own and the other's level of engagement to achieve the communicative goal efficiently as part of our pragmatic skillset (Prutting and Kittchner, 1987). However, despite the ubiquitous use of the term, perhaps due to the inherent complexity and subjectivity of this process in humans, the field is lacking a unified definition of engagement - an issue raised in multiple previous works (see Oertel et al. (2020); Pellet-Rostaing et al. (2023); Sorrentino et al. (2024)). These reviews show that the same term is used to refer to everything from the process of establishing a connection or maintaining it; to the procedural or social aspect of the interaction (task vs social engagement). When modeling engagement, automation relies on ground truth commonly provided by annotating engagement in an interaction. As a consequence of the vagueness of the term, various annotation methods have also been proposed, varying in the annotation scale (e.g. binary vs gradual), the unit of annotation (e.g. whole interaction, time window, action, turn, single utterance), and the relevant features guiding the annotations.

To reduce the ambiguity, here, we focus on *conversational engagement* as a particular instance of this concept. This highlights it as a concept that arises in a reciprocal, jointly managed verbal interaction in the absence of preallocated roles (Schegloff, 2007). We thus adopt the conversational engagement definition from recent work by Pellet-Rostaing et al. (2023): "*a state of attentional and emotional investment in contributing to the conversation by processing partner's multimodal behaviors and grounding new information*".

In this paper, we provide a novel annotation scheme and a dataset of turn-by-turn conversational engagement annotations. The aim of the dataset is to provide a new perspective on conversational engagement modeling, by coupling annotations serving as ground truth with multimodal data, including audio, eye tracking, speech embeddings, and uniquely, brain imaging. The original dataset that was annotated in this work includes functional MRI data of participants engaging in unrestricted conversations (Torubarova et al., 2025).

The contributions of the current paper are the following: 1) a novel fine-grained conversational engagement annotation scheme; 2) an open-source

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2787–2797
11–16 May 2026.

dataset of conversational engagement annotations of 25 hours of conversation (14 hours of human-human and 9.5 hours of human-robot data); 3) initial analysis of the annotations.

2. Background

Different approaches to annotating engagement have been attempted, using various units of observation and various scales (a detailed recent review is provided in Pellet-Rostaing et al. (2023)). Besides, most existing datasets that include engagement data have focused either on solely human-human (e.g. Ringeval et al. (2013); Cafaro et al. (2017); Reverdy et al. (2022); Borghesi et al. (2022); Pellet-Rostaing et al. (2023)) or human-agent data (e.g. McKeown et al. (2011); Ben-Youssef et al. (2017); Kesim et al. (2023); Lee et al. (2025)). Few datasets comprise data from both human-human (HHI) and human-robot (HRI) interaction in one study, which was one of our main interests. Stefanov and Beskow (2016) provides an audio-visual multiparty dataset with different combinations of human and robot partners, focused on visual attention in the interaction. This dataset, however, is focused on task-based object manipulation rather than conversation. Celiktutan et al. (2017) provides personality and engagement annotation data from human-human and human-robot conversations. In this dataset, however, engagement was self-annotated by experiment participants post-study as a single value for the whole interaction. None of the existing engagement datasets satisfy all of the following features: 1) uses both human-human and human-agent data from unscripted conversations; 2) provides conversational engagement annotations at a linguistically motivated scale; 3) uses a fine-grained annotation scheme.

Pellet-Rostaing et al. (2023) provided a 5-level annotation scheme, taking a conversational turn as a unit of annotation, due to its naturally emerging boundaries (Sacks et al., 1974). Using this scheme, they annotated an audio-visual 8-hour human-human conversational corpus with the purpose of implementing multimodal engagement detection. A set of multimodal behavioral cues, involving prosodic, morpho-syntactic and mimo-gestural information, was found to be the most informative set of features. When focusing on conversational scenarios, multiple works have investigated how other modalities contribute to modeling engagement, assuming linguistic information as the bare minimum of verbal interaction. The visual modality, e.g. eye movements, posture, facial expressions, and other behavioral cues have shown successful contribution to automatic engagement detection (Ishii et al., 2013; Huang et al., 2016; Sun et al., 2017; Dermouche and Pelachaud, 2018; Ben-Youssef et al., 2019). Several studies suggested employing physiological data for engagement modeling, due to its known correlation with cognitive processes (Sarkar and Etemad, 2020; Singh et al., 2024). This data thus might better reflect a person's internal state (such as engagement) compared to visual data. The current dataset includes engagement annotations of conversations coupled with functional MRI (fMRI) data. This non-invasive brain imaging method allows for continuous recording of the participant's brain activity during a task, such as conversation. In an fMRI setup, a participant is lying in the fMRI scanner while perceiving experimental stimuli (e.g. visual, auditory, or olfactory). This method inherently limits participant's freedom of movement (see Limitations), but does not require invasive sensors that would interfere with cognitive processes. Both affective and linguistic processes have been widely studied with fMRI, which makes it an excellent novel modality to investigate complex internal state such as engagement. Only one previous dataset includes fMRI data for human-human and human-robot conversations (Rauchbauer et al., 2020), however the interactions in this dataset are too short (1 minute) to establish engagement variation. The MRI setup, while adding a valuable novel modality, comes with additional challenges and limitations. Often, data for annotation comes from an excellently controlled experimental environment with highly sensitive microphones and several camera angles. Given the contribution showed by the visual modality, most if not all of the multimodal engagement annotations to a large extent rely on visual data. For instance, in the audio-visual study by Dermouche and Pelachaud (2018), speech was even filtered to be rendered incomprehensible while preserving prosodic features. The specifics of the MRI setup for this dataset created particular challenges for engagement annotations: as such, obtaining visual data was not possible, and audio data was noisy due to the MRI scanner background noise. Thus, while the annotators were on the one hand more limited in the cues they could rely on, they focused on the other hand on the details of participants' speech that expressed their internal state.

3. Method

3.1. Data

The data for the current study comes from a dataset collected by the current authors[1] (Torubarova et al., 2025). In this dataset, 50 participants engaged in three unscripted 10-minute conversations with a confederate in a between-subject design: 30 par-

[1] https://openneuro.org/datasets/ds004996

ticipants conversed with a human agent (HHI condition; mean age = 26.8, SD = 4.2, 15 female), and 20 with a robot agent remotely operated by the same confederate (HRI condition; mean age = 25.1, SD = 5.5, 10 female). All participants and the confederate were native Swedish speakers.

The study used a social robot Furhat (Al Moubayed et al., 2012) - a robotic head with a back-projected face, capable of a wide range of facial expressions and head movements. The robot was operated via a teleoperation interface, where the confederate was wearing a VR headset equipped with several cameras and an eye tracker, which captured the confederate's facial expressions, lip sync, eye movements, and head movements, and displayed them on the robot's mask and moved the robot's neck servos in real time. To approximate the content and voice quality in the two conditions, in both of them, the confederate was talking to the participants using her real voice.

The conversations were centered around ethical dilemmas, chosen to establish a potentially engaging 10-minute conversation. For each participant, the topics were pseudo-randomly sampled from a set of six, ensuring equal distribution: 1) Should we have robot judges that use machine learning to make decisions? 2) Would you take a DNA test before a first date with a potential partner? 3) If we invent a pill that stops aging, but only 5% of people can use it, should it be allowed? 4) Would you put a microchip in your arm with your personal data? 5) Should we have an app that allows parents and children to track each other? 6) Should we have a points system in society in which you can gain or lose social benefits for following or breaking the rules? The participants gave their opinion about the idea presented in the dilemma before and after conversation of a 5-point Likert scale, where 1 is "completely disagree" and 5 is "completely agree". The confederate's opinion on these topics was stable throughout the experiment, i.e. 1, 5 - pro, 2, 3 - neutral, 4, 6 - against.

To enable variation in conversational engagement, it was manipulated via the confederate's behavior across the three conversations with each participant using the cues established in previous studies: amount of backchannels, level of reactivity when addressed or asked a question, and level of proactivity in moving the conversation further. The manipulation created three distinct confederate's roles: Passive Listener, Active Listener and Engaged Communicator (for simplicity, we will refer to these conditions and low, medium, and high engagement levels throughout the paper). Thus, every participant experienced all three levels of engagement with one of the agents and a random set of three ethical dilemmas.

In the study, participants were lying in an fMRI scanner recording their brain activity. They were connected to the confederate located in a separate room, via a bidirectional audio link (using noise-cancelling earphones and a microphone) and a unidirectional video link. A participant could see the confederate on a screen, but not vice versa, due to hardware limitations: the MRI head coil placed over participant's head does not allow to record participant's full face, but allows the participant to have a non-obstructed view of the screen. Thus, to focus on participants' conversational engagement, only audio recordings of the conversations were used for annotation. Data from two participants in the HHI condition were not annotated due to poor audio quality, thus, data of 48 participants were annotated and analyzed.

3.2. Data Preprocessing

Conversations were recorded via two separate channels for the participant and the confederate. Since the fMRI scanner created a lot of background noise, the participant's audio was denoised online during the experiment to establish a natural conversation. Our dataset includes both the raw audio and denoised audio using Adobe Enhance Speech v2[2], which provided optimal speech quality for the annotations. The audio from both interlocutors was transcribed using Google Cloud Speech-to-Text[3], and the transcriptions were manually corrected by two native Swedish transcribers. Each conversation was assigned a unique random ID to avoid bias for the annotator, who was aware of the design of the study but did not know the confederate's level of engagement or participant ID in a given conversation.

3.3. Turn Definition

To define a turn, we followed the definition in Pellet-Rostaing et al. (2023): a turn consists of a series of consecutive intra-pausal units (*IPUs*, blocks of speech surrounded by silence of at least 200 ms) from the same speaker until the next speaker change (see Fig. 1). A turn ends when the current speaker stops speaking and the other one picks up the turn. In case of overlap, the turn is assigned to both interlocutors. To define the IPUs, the start and end timestamps of speech segments from both interlocutors were automatically extracted using Silero Voice Activity Detector (Silero Team, 2024) with 200 ms minimum silence threshold. Both recordings were aligned in time and IPUs were merged into turns using pympi, a Python module

[2] https://podcast.adobe.com/en/enhance

[3] https://cloud.google.com/speech-to-text

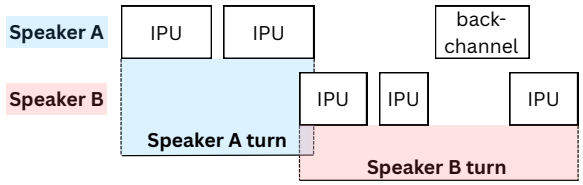


Figure 1: Example of a series of IPUs merged into a single turn.

for processing ELAN files (Lubbers and Torreira, 2018).

Turn definition in Pellet-Rostaing et al. (2023) attributed an IPU to the current speaker's turn unless it constituted a feedback, or a *backchannel*, since "backchanneling is by definition an activity of a listener". In our data, backchannels were not previously annotated and were instead tagged during engagement annotation process. Thus, the automatic turn parsing would occasionally merge a backchannel IPU with the other IPUs constituting a turn. For this reason, the boundaries of each turn were manually checked by the annotator and corrected if needed. In Pellet-Rostaing et al. (2023), turns shorter than 1 second were removed from the annotation process, but in our data, all turns no matter their duration, were preserved.

3.4. Annotation Process

Two annotators were employed: the main annotator worked on 100% of the data, and the second annotator did 10% of the data to assess annotation scheme reliability. Both were native Swedish speakers with an experimental linguistic background. Annotations were created in ELAN (Wittenburg et al., 2006). For each conversation, an EAF file was created containing both raw and denoised participant's audio (raw audio was included in case noise reduction made speech hard to comprehend), the confederate's audio, and speech transcriptions of both interlocutors. Three additional tiers were created: 1) an engagement annotation tier containing time boundaries of each participant's turn to be annotated; 2) an annotation confidence tier with the same boundaries; 3) an empty backchannel tier.

For each conversation, the annotator's task was to annotate the level of conversational engagement in each participant's turn on a scale from 1 to 5, following the annotation scheme in Pellet-Rostaing et al. (2023):

- Level 1: strongly disengaged
- Level 2: disengaged
- Level 3: neutral
- Level 4: engaged
- Level 5: strongly engaged

For this rating, the annotator was asked to keep in mind three guideline questions, as taken from Pellet-Rostaing et al. (2023):

1. How willing is the participant to contribute to the progress of the conversation?
2. How invested is the participant in what he/she is saying?
3. How interested is he/she in the conversation?

If a turn consisted of only a backchannel (for definition, see 3.5.2), the annotator was asked to move this turn to the backchannel tier and not assign an engagement score. The annotator was also informed that occasionally, the turn boundaries might erroneously include a backchannel. In this case, the annotator was asked to correct the turn boundaries (see 3.5.1). Only the main annotator completed these tasks, and the second annotator worked with already corrected data.

In addition, for each engagement annotation the annotator was asked to rate his own confidence in the given annotation on a scale from 1 (not confident at all) to 5 (completely confident).

During the annotation process, which spanned two months, the main annotator formulated more fine-grained criteria to support their annotation judgments, as presented in section 3.6.

3.5. Annotator's Guidelines

3.5.1. Correcting Turn Boundaries

The annotator manually checked automatic turn boundaries and corrected them if necessary, following these guidelines:

1. A turn starts when the participant starts speaking, and it does not constitute a backchannel.
2. A turn ends if the participant stops speaking and their partner starts speaking, and the partner's speech is not a backchannel.
3. A turn ends if the participant responds to something their partner is saying. In these cases, a turn ends and a separate turn begins shortly thereafter.

3.5.2. Backchannels Definition

The annotator manually tagged IPUs as backchannels. Since there is no single definition of what classifies as a backchannel (Schegloff, 1982; Kjellmer, 2009), the annotator followed these guidelines, comprising features of backchannels identified in previous research:

1. Backchannels are utterances made to indicate that a person is listening, and do not serve a communicative function beyond giving feedback.
2. They are mostly short vocalizations. They do not carry much, if any, semantic information.

Disengaged		Managing Conversation	Engaged	
1	**2**	**3**	**4**	**5**
Fails to engage	Struggles to engage	Seems to be potentially disengaging	Engages so that conversation flows naturally	Engages with vigor
Displays signs of discomfort	Takes long pauses	Uses the turn barely to support conversation structure	If displays struggle, not for lack of interest	Shows excitement or fascination
Repeats a lot of what has already been said	Moves the conversation forward, but not much	Uses the turn to exchange information necessary for conversation	Moves the conversation forward	Frequently introduces new ideas
Does not introduce new ideas	Does not display effort	Asks for clarifications	Introduces new ideas	Speaks louder than normal
Does not appear to be listening	Asks the interlocutor to help them by taking the turn	Helps the interlocutor to find words	Asks interlocutor for input out of interest in what they have to say	Exchanges turns quickly
Displays a lot of disfluencies	Displays increased amount of disfluencies	Uses phatic expressions	Finds the topic interesting	Displays a high speech rate

Table 1: Description and common indicators of 5 levels of speaker's conversation engagement.

3. They do not serve the intent to interrupt a person speaking or take over the turn.

It is not always obvious what is and isn't a backchannel. For example, a person might initiate their turn by making a vocalization that is identical to a backchannel. In these cases, it was up to the annotator's judgment and intuition to determine what was considered a backchannel.

3.6. Refined Conversation Engagement Annotation Scheme

To extend the annotation guidelines in (Pellet-Rostaing et al., 2023), the annotator comprised descriptions of each level of engagement, along with common indicators that usually occur at that level (Table 1). Indicators are neither sufficient nor necessary conditions for identifying any level of engagement. An engagement level might not exhibit all indicators simultaneously, and indicators from different levels might co-occur. The indicators are to guide the annotator's judgment, not supplant it.

- **Level 1: Strongly Disengaged**

A speaker is strongly disengaged when they do not contribute to the conversation in any substantial way. Strong disengagement does not necessarily indicate that the person is unwilling to engage, but rather describes their lack of ability to do so. However, if they are unwilling to contribute to the conversation, they are almost certainly strongly disengaged.

- **Level 2: Disengaged**

The speaker does partake in the conversation, but to a very limited extent. They show difficulty contributing to the conversation, in spite of a willingness to do so. The effort put into the conversation often seems forced. The speaker might employ some strategy to aid the conversation in place of a relaxed conversational flow, such as monologuing by themselves or searching for new topics to introduce.

- **Level 3: Managing Conversation**

At this level, the speaker displays neither high nor low engagement. It consists in large part of expressions that benefit the conversation (e.g., ensuring or establishing a common ground on the topic), but do not contribute to the meaningful parts of it or move it forward. Much of it consists of language that serves a barely functional role: to establish mutual understanding, to prevent misunderstandings, or as conventional gestures made during conversation with little meaning otherwise (i.e, phatic expressions). It also likely displays that the speaker is potentially shifting from a higher level of engagement to a lower one, e.g. if conversation enters a phase in which the current topic at hand is exhausted, but the interlocutors do not move on from it immediately, in case one of the parties finds more to say about it.

- **Level 4: Engaged**

The speaker is engaged in the conversation. It flows mostly unhindered and seems mostly effort-

less. The speaker contributes significantly to the conversation, is attentive to their interlocutor, or has much to say about the topic. The speaker might still struggle, but not due to lack of interest: e.g. pause to think or rephrase what they said several times, stutter, or search for the right word or expression.

- **Level 5: Strongly Engaged**

The speaker makes a strong contribution to the conversation and is an eager participant. They engage with the conversation more than would be expected, e.g. frequently jumping from topic to topic before finishing the previous one. Turn taking is rapid, and does not leave much if any silence between the turns. The interlocutors might have their speech overlap. The speaker might display a strong emotion (most often enjoyment, but possibly also frustration or anger), or speak louder than usual.

4. Dataset Description

The resulting dataset includes turn-by-turn annotations of conversational engagement from 28 participants in the human-human group and 20 in the human-robot group, corresponding to 840 and 570 minutes of dialogue, respectively (see Table 2).

Agent Condition	Eng. Condition	Turn Count
Human (N=28)	High	513
	Medium	538
	Low	350
	Total	**1,401**
Robot (N=20)	High	426
	Medium	437
	Low	325
	Total	**1,188**

Table 2: Distribution of annotated turns by agent and engagement conditions.

The distribution of the number and duration of turns in the dataset is presented in Figure 2, together with comparisons by agent condition (human/robot) and predefined engagement condition (high/medium/low).

4.1. Inter-Annotator Agreement

To check for annotation scheme reliability, the second annotator first annotated a 10% subset, consisting of three random participants from HHI condition and two from HRI condition, following the original guidelines provided in section 3.4. Then the two annotators had a two-hour meeting discussing the instances of disagreement (specifically, the second annotator used the scheme more conservatively, skewing towards higher scores), and the main annotator introduced the fine-grained annotation scheme. Then, they annotated two conversations together to practice using the annotation scheme. Then the second annotator annotated another 10% subset with a different set of participants, consisting of 270 turns, following the refined guidelines provided in section 3.6. In this subset, the main annotator used the labels 1 and 5 two and nine times, respectively, while the second annotator used label 1 once and never used label 5. Given this rarity and inconsistent use of the extremes, to calculate inter-annotator agreement, we collapsed levels 1+2 and 4+5. Moderate agreement was established between the annotators with Cohen's kappa $\kappa = 0.441$. The percentage agreement across classes is presented in Table 3.

Annotation Class	% Agreement	Cohen's kappa
Disengaged (1–2)	51.43	
Managing Conversation (3)	76.19	
Engaged (4–5)	73.44	
Overall	**65.19**	**0.441**

Table 3: Inter-annotator agreement on a 10% subset of the data, by annotation label.

4.2. Annotations Distribution

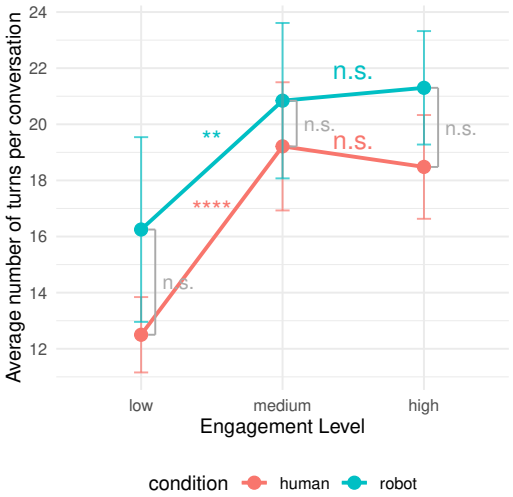
The distribution of engagement annotation labels is presented in Figure 3. Extreme labels (1 and 5) were rare, comprising 0–6% and 5–7% in the human condition and 0–1% and 2–3% in the robot condition, respectively. To investigate how the annotations correspond to predefined engagement conditions, we fitted a cumulative link mixed-effects model with *agent condition* (human/robot), *engagement condition* (high/medium/low), plus random intercepts for *participant* and *dilemma*, with coefficients on the log-odds scale. The estimates with comparisons across conditions are presented in Figure 4.

4.3. Backchannels Distribution

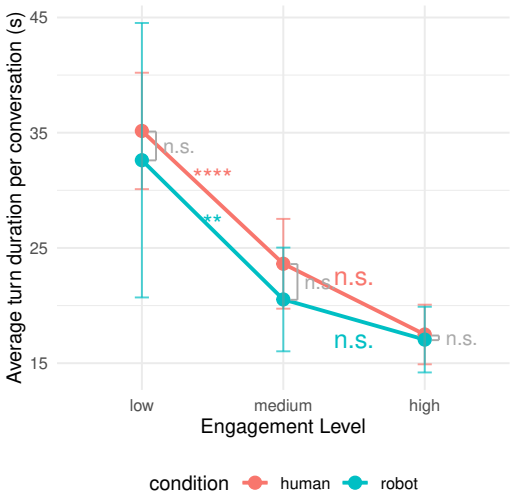
The dataset includes manually tagged participant's backchannels. While backchannels represent the activity of a listener and engagement annotations in this dataset represent the activity of a speaker, we were interested in how these two roles in a conversation might influence each other. The distribution of backchannel count with comparisons across conditions is presented in Figure 5.

4.4. Stance Adjustment Distribution

Previous analysis of the original dataset in Torubarova et al. (2025) showed that participants



(a) Average turn count estimated using a Poisson GLM with post-hoc Tukey's HSD.



(b) Average turn duration estimated using a linear mixed-effects model with post-hoc Tukey's HSD.

Figure 2: Model-estimated turn statistics by agent (human/robot) and engagement (low/medium/high) conditions. Error bars show 95% confidence intervals. Here and in Figures 4 and 5, colored stars show within-condition comparisons, grey brackets show between-condition comparisons. Significance: **** $p < .0001$, ** $p < .01$, n.s. $p > .05$.

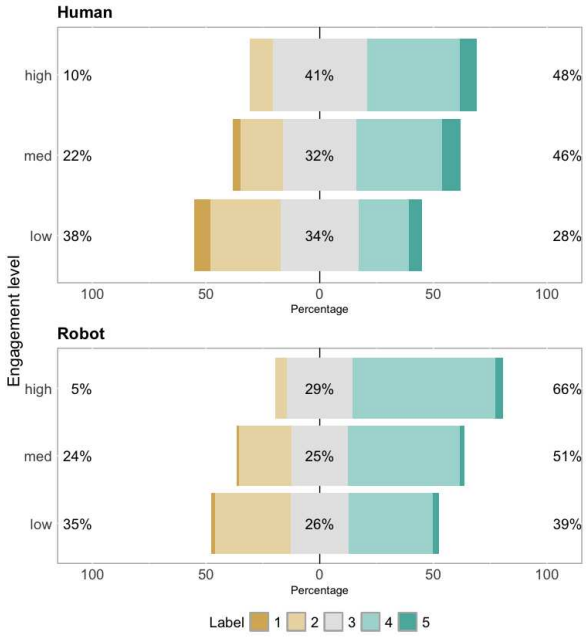


Figure 3: Engagement annotation distribution by agent and engagement conditions.

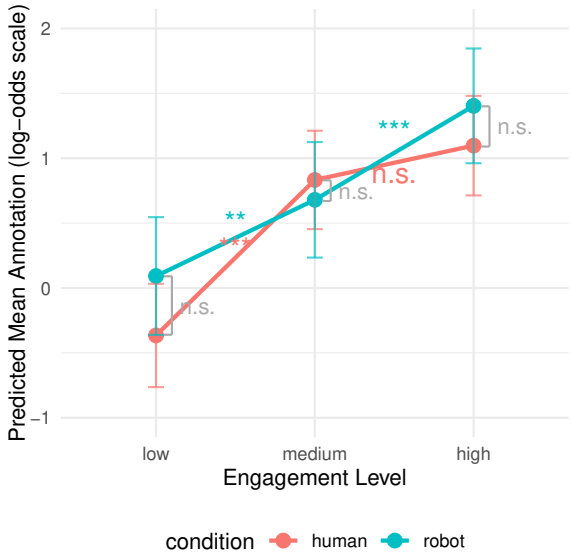


Figure 4: Predicted annotation by agent and engagement conditions from a cumulative link mixed model with post-hoc Tukey's HSD. Points are estimated marginal means on the latent log-odds scale with 95% CIs. Significance: *** $p < .001$, ** $p < .01$, n.s. $p > .05$

were more likely to change their opinion on the dilemma after a conversation with a human compared to a robot. Here, we investigated whether they were more likely to align or disagree with the confederate in different conditions. We encoded the confederate's stance on each dilemma on a 5-point scale as 1 - against, 3 - neutral, 5 - pro, and calculated whether the distance between participant's and confederate's opinions changed after having the conversation. One participant's opinion data was missing, thus 47 participants were analyzed. The distribution of stance adjustment is presented in Figure 6. For the cases when participants adjusted their stance, a mixed-effects logistic

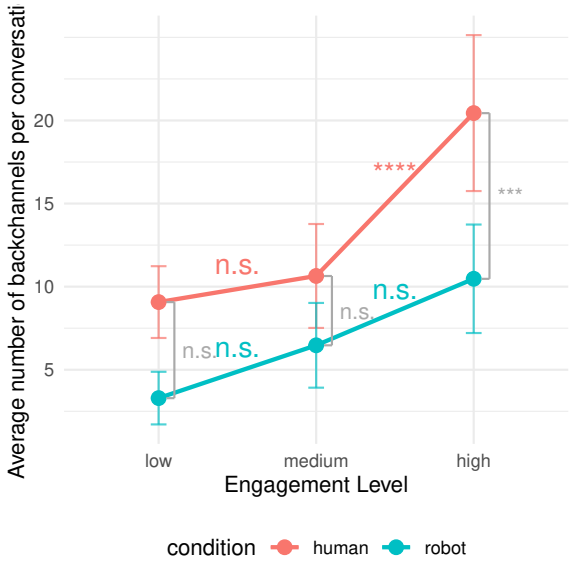


Figure 5: Average number of backchannels by agent and engagement conditions, estimated using a linear mixed-effects model with post-hoc Tukey's HSD. Error bars show 95% CIs. Significance: **** $p < .0001$, *** $p < .001$, n.s. $p > .05$.

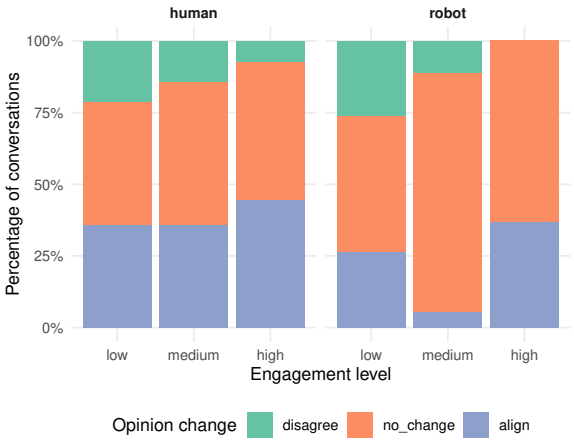


Figure 6: Distribution of stance adjustment before and after conversation with respect to the confederate's stance, by agent and engagement conditions.

regression model showed significantly higher odds for aligning with the confederate rather than disagreeing ($p = 0.003$)

5. Conclusion and Future Directions

In this paper, we introduced CONVERSE, a dataset of fine-grained conversational engagement annotations applied to 25 hours of interaction from the multimodal NeuroEngage corpus (Torubarova et al., 2025). To the best of our knowledge, this is the first dataset to annotate engagement on a turn level in human-human and human-robot data within the same study. Inclusion of both types of agents allows for a direct comparison to uncover the cues that build engagement with artificial agents, enabling the development of more socially aware robots.

The dataset includes transcribed and annotated conversational turns (in Swedish) and manually tagged verbal backchannels. The current dataset extends our previous multimodal corpus that includes brain imaging (fMRI) data alongside eye tracking, audio and speech features, personality and stance measures, offering the research community a novel resource for investigating engagement from behavioral and neural perspectives.

Our refined, fine-grained annotation scheme provides a robust framework for analyzing conversational engagement at the turn-by-turn level, showing moderate inter-annotator agreement ($\kappa = 0.44$), similar to previous datasets annotating complex affective states (e.g. Ringeval et al. (2013); Poria et al. (2018)), despite the challenging MRI setup and no access to visual data. For this dataset, the extreme engagement labels (1 and 5) were comparatively rare, likely due to the absence of visual cues that often signal extremes (e.g., smiles, gaze aversion, nods, posture shifts). Despite these constraints and our choice to elicit natural conversations rather than provoke extreme states, annotators achieved good agreement in neutral and high engagement classes. The annotation scheme can be applied to any conversational data, not limited to an MRI setup, as it includes various linguistic indicators of engagement, and can be further enhanced with visual indicators. Our results show that reliable engagement annotation is feasible from audio-only, in-scanner conversations. We maintain the 5-level scheme as the canonical release for fine-grained analyses, and we provide a 3-level mapping for robustness in low-data scenarios that are limited in visual modality such as ours.

This dataset opens several exciting avenues for researchers at the intersection of multimodal machine learning, human-robot interaction, and cognitive neuroscience. Our analysis, confirming both the successful manipulation of participants' engagement and differences in opinion change between partners, provides a rich foundation for this work. For machine learning, CONVERSE provides a unique opportunity to integrate brain imaging as a new modality, allowing researchers to explore how neural activity correlates with annotated engagement to create models that understand the internal cognitive signatures of engagement, not just its external cues.

Future work will build on these foundations. First, we will use the time-resolved engagement annotations to clarify the role of specific brain networks, such as the default mode network, to distinguish be-

tween mind-wandering due to disengagement and internal processing necessary for comprehension. By correlating our turn-by-turn annotations with neural data, we can build more sensitive models of the brain networks that support engagement in both speech production and comprehension. Second, the multimodal nature of CONVERSE opens the door for advanced cross-participant modeling, using methods like multimodal transformers and hyper-brain analyses to capture inter-brain coupling. We anticipate this work will stimulate interdisciplinary research, bringing the field closer to machines that not only recognize engagement behaviorally but also model it in ways informed by the human brain.

Data Availability

The dataset is available at https://github.com/tor-e-i/CONVERSE. It includes engagement and backchannels annotations alongside conversational transcriptions and study metadata in .csv and .json formats, detailed annotation instructions, the .eaf files used for annotation, and the supporting code.

Limitations

As the dataset contains annotations of MRI-based conversations, one must acknowledge the inherent limitation this method poses on the naturalness of the interaction. Due to hardware constraints, a solution for recording participant's face inside the MRI scanner has not yet been found. Because of that, the interaction was not fully reciprocal since the confederate could not see participant's face, and the dataset does not include participant's video. Despite no reported discomfort by the participants, it is important to keep in mind that MRI setup limits participants' movement, and background noise might have forced participants to speak louder than usual. The dataset offers both raw noisy and denoised audio. Although the audio was denoised using a state-of-the-art speech enhancing model, a certain level of artifacts is inevitable; thus this audio might not be reliable for analyzing prosodic features. The noisy raw audio, however, can be a testing ground for relevant algorithms.

Turn-based annotations provide a linguistically motivated scale. However, in the current dataset, turn duration varied across conditions and could have varied within a single conversation. Some turns spanned a long time, and engagement may vary within a single turn. Thus, further exploration into the most suitable engagement annotation unit is necessary.

The audio-only annotation setup resulted in sparsity at the annotation extremes (1, 5). While it reflects our design choice to favor ecological dialogue rather than targeted induction of very low/high engagement, this sparsity can reduce statistical power for class-based metrics and inflate variance. Collapsing to 3 levels improves reliability but sacrifices granularity; researchers should choose the best representation that matches their task.

Ethical Statement

This study builds on a previously collected dataset, and no new data were collected. The original protocol received ethical approval, and all participants provided informed consent prior to data collection, being informed about the terms of data storage and distribution. To protect personal data, we apply GDPR-compliant safeguards.

Researching conversational engagement has ethical significance beyond methodological issues. Because conversation is fundamental to social connection, empathy, and attentive interaction, studying and modeling engagement has important societal implications, particularly for the development of socially aware conversational agents. By understanding how engagement unfolds in both human and human-robot interactions, this work contributes to developing systems that can interact with users more responsibly. We also recognize that technologies capable of modeling internal states like engagement could be misused for manipulative or deceptive purposes if not developed with robust ethical safeguards. We acknowledge potential risks (e.g., re-identification, stigmatizing inferences) and limitations to discourage misuse, and we recommend responsible use of this data in human–human and human–robot studies.

6. References

Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. 2012. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive behavioural systems: COST 2102 international training school, dresden, Germany, february 21-26, 2011, revised selected papers*, pages 114–130. Springer.

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*, pages 464–472.

Atef Ben-Youssef, Giovanna Varni, Slim Essid, and Chloé Clavel. 2019. On-the-fly detection of user engagement decrease in spontaneous human–robot interaction using recurrent and deep neural networks. *International Journal of Social Robotics*, 11(5):815–828.

Daniele Borghesi, Andrea Amelio Ravelli, and Felice Dell'Orletta. 2022. What makes the audience engaged? engagement prediction exploiting multimodal features. In *NL4AI@ AI* IA*, pages 85–107.

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*, pages 350–359.

Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. 2017. Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing*, 10(4):484–497.

Soumia Dermouche and Catherine Pelachaud. 2018. From analysis to modeling of engagement as sequences of multimodal behaviors. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Yuyun Huang, Emer Gilmartin, and Nick Campbell. 2016. Conversational engagement recognition using auditory and visual cues. In *Interspeech*, pages 590–594.

Ryo Ishii, Yukiko I Nakano, and Toyoaki Nishida. 2013. Gaze awareness in conversational agents: Estimating a user's conversational engagement from eye gaze. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(2):1–25.

Ege Kesim, Tugce Numanoglu, Oyku Bayramoglu, Bekir Berker Turker, Nusrah Hussain, Metin Sezgin, Yucel Yemez, and Engin Erzin. 2023. The ehri database: a multimodal database of engagement in human–robot interactions. *Language Resources and Evaluation*, 57(3):985–1009.

Göran Kjellmer. 2009. Where do we backchannel?: On the use of mm, mhm, uh huh and such like. *International Journal of Corpus Linguistics*, 14(1):81–112.

Dong Won Lee, Yubin Kim, Denison Guvenoz, Sooyeon Jeong, Parker Malachowsky, Louis-Philippe Morency, Cynthia Breazeal, and Hae Won Park. 2025. The human robot social interaction (hsri) dataset: Benchmarking foundational models' social reasoning. *arXiv preprint arXiv:2504.13898*.

Mart Lubbers and Francisco Torreira. 2018. pympiling: a python module for processing elans eaf and praats textgrid annotation files. *Computer software]. https://pypi. python. org/pypi/pympiling*.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. 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.

Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. 2020. Engagement in human-agent interaction: An overview. *Frontiers in Robotics and AI*, 7:92.

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*, 5:1062342.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Carol A Prutting and Diane M Kittchner. 1987. A clinical appraisal of the pragmatic aspects of language. *Journal of Speech and hearing Disorders*, 52(2):105–119.

Birgit Rauchbauer, Youssef Hmamouche, Brigitte Bigi, Laurent Prévot, Magalie Ochs, and Thierry Chaminade. 2020. Multimodal corpus of bidirectional conversation of human-human and human-robot interaction during fmri scanning. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 668–675.

Justine Reverdy, Sam O'Connor Russell, Louise Duquenne, Diego Garaialde, Benjamin R Cowan, and Naomi Harte. 2022. Roomreader: A multimodal corpus of online multiparty conversational interactions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2517–2527.

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)*, pages 1–8. IEEE.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735.

Pritam Sarkar and Ali Etemad. 2020. Self-supervised ecg representation learning for emotion recognition. *IEEE Transactions on Affective Computing*, 13(3):1541–1554.

Emanuel A Schegloff. 1982. Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing discourse: Text and talk*, 71(93).

Emanuel A Schegloff. 2007. *Sequence organization in interaction: A primer in conversation analysis I*, volume 1. Cambridge university press.

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

Silero Team. 2024. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad.

Alakhsimar Singh, Nischay Verma, Kanav Goyal, Amritpal Singh, Puneet Kumar, and Xiaobai Li. 2024. Visiophysioenet: Multimodal engagement detection using visual and physiological signals. *arXiv preprint arXiv:2409.16126*.

Alessandra Sorrentino, Laura Fiorini, and Filippo Cavallo. 2024. From the definition to the automatic assessment of engagement in human–robot interaction: A systematic review. *International Journal of Social Robotics*, 16(7):1641–1663.

Kalin Stefanov and Jonas Beskow. 2016. A multi-party multi-modal dataset for focus of visual attention in human-human and human-robot interaction. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4440–4444.

Mingfei Sun, Zhenjie Zhao, and Xiaojuan Ma. 2017. Sensing and handling engagement dynamics in human-robot interaction involving peripheral computing devices. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 556–567.

Ekaterina Torubarova, Caroline Arvidsson, Jonathan Berrebi, Julia Uddén, and Andre Pereira. 2025. Neuroengage: A multimodal dataset integrating fmri for analyzing conversational engagement in human-human and human-robot interactions. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 849–858. IEEE.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. Elan: A professional framework for multimodality research. In *5th international conference on language resources and evaluation (LREC 2006)*, pages 1556–1559.