# EmpathicStories++: A Multimodal Dataset for Empathy towards Personal Experiences

**Jocelyn Shen*** **Yubin Kim*** **Mohit Hulse** **Wazeer Zulfikar**
**Sharifa Alghowinem** **Cynthia Breazeal** **Hae Won Park**
Massachusetts Institute of Technology, Cambridge, MA, USA
{joceshen, ybkim95, mhulse, wazeer, sharifah, cynthiab, haewon}@mit.edu

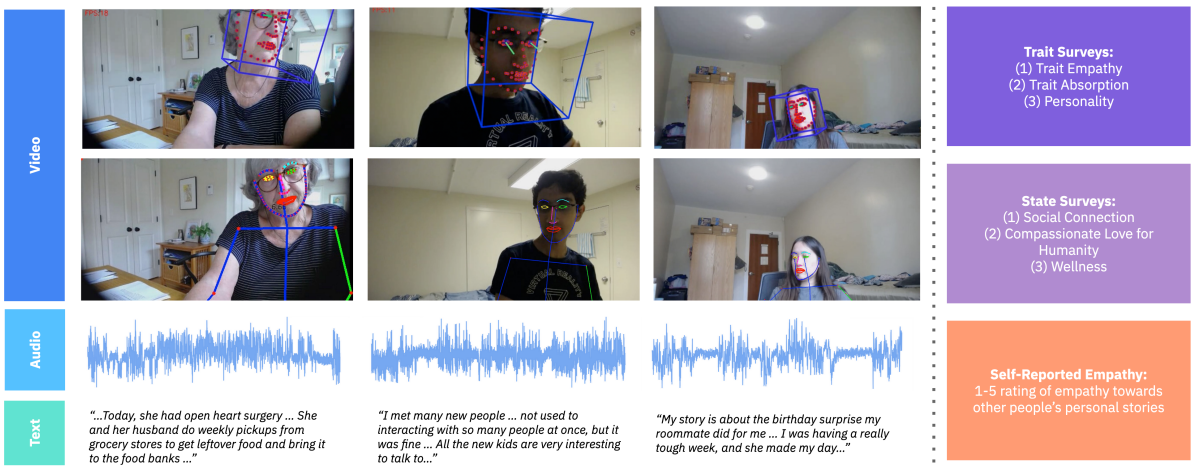


Figure 1: **The EMPATHICSTORIES++ dataset** is collected from a month-long in-the-wild deployment of 41 participants (across 269 sessions with 53 hours of data) telling personal stories and reading other people's personal stories with an AI agent. We publicly release all video, audio, and text data in addition to psychometric surveys in order to advance computational empathy research, and more broadly, social-emotional reasoning in AI.

## Abstract

Modeling empathy is a complex endeavor that is rooted in interpersonal and experiential dimensions of human interaction, and remains an open problem within AI. Existing empathy datasets fall short in capturing the richness of empathy responses, often being confined to in-lab or acted scenarios, lacking longitudinal data, and missing self-reported labels. We introduce a new multimodal dataset for empathy during personal experience sharing: the EMPATHICSTORIES++ dataset[1] containing 53 hours of video, audio, and text data of 41 participants sharing vulnerable experiences and reading empathically resonant stories with an AI agent. EMPATHICSTORIES++ is the first longitudinal dataset on empathy, collected over a month-long deployment of social robots in participants' homes, as participants engage in natural, empathic storytelling interactions with AI agents. We then introduce a novel task of predicting individuals' empathy toward others' stories based on their personal experiences, evaluated in two contexts: participants' own personal shared story context and their reflections on stories they read. We benchmark this task using state-of-the-art models to pave the way for future improvements in contextualized and longitudinal empathy modeling. Our work provides a valuable resource for further research in developing empathetic AI systems and understanding the intricacies of human empathy within genuine, real-world settings.

## 1 Introduction

Empathy is a fundamental pillar of interpersonal human interactions ranging from prosocial behavior to enhancing human connection (Morelli et al., 2015). Modeling and understanding empathy is a complex task, due to its inherently interpersonal and experiential nature: empathy is tied to neurological synchronizations between representations of self and other (Decety and Lamm), and is dependent on a person's past experiences (Hodges et al., 2010). Interest in empathy within AI communities has grown in recent years, as systems advance in context-awareness, naturalness, and fluency, although they typically fall short in social reasoning (Sap et al., 2022). Few prior works present datasets that are sufficient to capture the richness

[1] https://mitmedialab.github.io/empathic-stories-multimodal/



*Findings of the Association for Computational Linguistics ACL 2024*, pages 4525–4536

of human empathy responses during personal experience sharing. These datasets are limited in the following ways: (1) They are not captured in-the-wild. Existing multimodal empathy datasets are sourced from in-lab, online, or acted settings, which may differ greatly from empathy expressed in natural conversations. (2) They are not longitudinal, capturing only one-shot interaction settings, despite empathy being dependent on a combination of many past experiences. (3) Previous datasets are not self-labeled. While empathy can be inferred by external cues, it is an inherently subjective process, requiring self-reported labels for user-centric or personalized modeling.

In this work, we present the EMPATHICSTORIES++ dataset, a multimodal dataset collected from a month-long deployment of social robots in-the-wild. In this work, participants shared personal stories with the robot, read stories that were empathically similar to their own experience, and then reflected on stories they empathized with (Shen et al., 2023a). Our interaction design allows researchers to explore empathy in the context of personal experience sharing and understand the influence of users' past experiences on their empathy towards others' experiences. We address gaps in previous empathy and emotion recognition datasets through the following attributes: (1) Participant data is captured in their own homes with a social robot. Previous works have shown that users are more comfortable disclosing sensitive information with AI partners than with people (Pickard et al., 2016; Lucas et al., 2014). Participants in our study shared emotionally diverse and vulnerable stories from the comfort of their own homes. (2) Participants interacted with the robot over the course of a month, allowing us to obtain longitudinal data. (3) After participants read other peoples' empathically similar stories, they self-rated their empathy towards the story, resulting in more authentic empathy labels. In addition to providing the raw video, audio, and text data for each interaction, we provide extracted features from all three modalities, as well as self-reported psychometric data (i.e. personality, well-being, etc.) and empathy ratings towards other people's stories. These properties enable AI researchers to capture the complexity of empathy in its contextual, longitudinal, and personal dimensions. In addition to providing the EMPATHICSTORIES++ dataset, we present a new task on predicting a person's empathy towards other's stories based on their own personal experiences. We evaluate this task in two settings: (1) predicting empathy based on the user's own shared story (e.g. *"I do weekly pickups at the local grocery store to bring leftover food to the food banks..."*), and (2) predicting empathy based on a user's reflection on a story they read (e.g. *"I can really relate to the narrator's feeling of wanting to help others..."*). In summary, our contributions are as follows: (1) The first multimodal dataset with in-the-wild, long-term, and self-reported cues on empathy towards other people's experiences, containing video, audio, text, as well as low-level features from each modality, self-reported psychometric data, and empathy ratings towards other people's stories. (2) A novel task for predicting a user's empathy towards another person's story. (3) Benchmarking empathy prediction using state-of-the-art approaches to enable further improvements in contextual and longitudinal empathy modeling. Our work is a valuable resource for future work in developing social-emotional AI systems, improving interpretability of empathy prediction models, and promoting research on understanding cognitive insights of human empathy.

## 2 Related Work

Relevant prior works span two major areas: (1) social psychological theory on the relationship between prior experience and empathy and (2) datasets containing emotion or empathy ratings used for social-emotional reasoning tasks.

### 2.1 Empathy and Memory of Experiences

Empathy towards others is conditioned on situational (similarity between observer and target) and trait factors (personality, learning history) (Davis, 2004; Roshanaei et al., 2019). Furthermore, empathy is tied to important social functions such as prosocial behaviors, social connection, well-being, and psychiatric disorders (Morelli et al., 2015). A person's past experiences and memories play an important role in both situational and trait empathy. This has been shown clearly in prior work on the social neuroscience of representations of self and other: an observer's reaction to a target is elicited by language-based cognitive networks that trigger relevant memories with observer's own feelings (Davis, 2004). Other studies use neuroimaging to show that prosocial behaviors may be due to synchronized representations of self and other (Decety

| Dataset | Modalities | Self-annotated | Longitudinal | Source | Collected in-the-wild | # Subjects | Quantity (video/audio) | Quantity (text) |
|---|---|---|---|---|---|---|---|---|
| MELD (Poria et al., 2019) | V + A + T | ✗ | ✗ | TV | ✗ | 407 | – | 13,708 utterances |
| M3ED (Zhao et al., 2022) | V + A + T | ✗ | ✗ | TV | ✗ | 626 | – | 990 dialogues / 24,449 utterances |
| EmoInt-MD (Singh et al., 2023) | V + A + T | ✗ | ✗ | Movies | ✗ | 4375 | 534 hrs / 32,040 min | 724,756 utterances |
| MEDIC (Zhou'an_Zhu et al., 2023) | V + A + T | ✗ | ✗ | Acted motivational interviews | ✗ | – | 11 hrs / 678 min | 771 utterances |
| OMG-Empathy (Barros et al., 2019) | V + A + T | ✓ | ✗ | In-lab | ✗ | 10 listeners, 2 speakers | 8 hrs / 480 min | – |
| EmpatheticDialogues (Rashkin et al., 2019) | T | ✗ | ✗ | Crowdsourced | ✗ | 810 | – | 24,850 dialogues / 107,247 utterances |
| Empathic Conversations (Omitaomu et al., 2022) | T | ✓ | ✗ | Crowdsourced | ✗ | 92 | – | 5,821 utterances |
| EmpathicStories (Shen et al., 2023a) | T | ✗ | ✗ | Online stories | ✓ | – | – | 1,500 stories |
| Sharma et al. (2020) | T | ✗ | ✗ | Online peer support platforms | ✓ | – | – | 10,143 utterances |
| EDOS (Welivita et al., 2021) | T | ✗ | ✗ | Movie subtitles | ✗ | – | – | 3,488,300 utterances |
| EmpathicStories++ | V + A + T | ✓ | ✓ | Real-world deployment | ✓ | 41 | 53 hrs / 3,180 min | 5,380 utterances |

Table 1: **Comparison of EMPATHICSTORIES++ to related datasets**. In contrast to other datasets, we collect data in-the-wild, over a month-long deployment, and our data is self-annotated with empathy and psychometrics. Since our dataset is interaction-based (we fixed the number of conversation turns per session) and in the real world, we have a limited number of utterances compared to text-only datasets that are crowdsourced from the internet.

and Lamm). Memories of other people's past experiences can modulate empathy, as these memories are used to simulate how one might feel in a new situation (Ciaramelli et al., 2013), and the vividness of memory of others' experiences is tied to prosocial intentions (Gaesser, 2013).

Besides recalling prior experiences of oneself or others, the process of sharing personal experiences is strongly tied to empathy elicitation. Sharing personal memories makes conversations more truthful, engaging and communicates a person's intentions or feelings (Pillemer, 1992; Bluck, 2003). The elicited empathy from experience sharing is even stronger when a listener responds with their own personal memories. In empathetic communication, both verbal (vividness of images, verb tense) and nonverbal (emotional gesturing, prosody) cues play a role in perceived empathy (Pillemer, 1992; Haase and Tepper).

Our dataset addresses all the previous points about empathetic communication: (1) self-reported annotations of situational and trait factors, (2) surveys of relevant social functions including social connection and wellbeing, and (3) video, audio, and transcripts of sessions with participants recalling their own memories and reflecting on others' past experiences over time.

## 2.2 Social-Emotional Datasets

Beyond modeling empathy alone, more broadly, datasets for social and emotional benchmarking have garnered interest in recent years. Datasets such as MELD (Poria et al., 2019), M$^3$ED (Zhao et al., 2022), and EmoInt-MD provide multimodal datasets annotated with emotion in conversations pooled from TV shows or movies. The Social-IQ dataset provides a multimodal benchmark for measuring social intelligence (Zadeh et al., 2019) and the related Social-IQA dataset benchmarks social intelligence with the text modality alone (Sap et al., 2019). There are also datasets that capture the emotions of individuals during story sharing, such as SEND (Ong et al., 2021), emotions of dyads, such as IEMOCAP (Busso et al., 2008) and DAMI-P2C (Chen et al.), as well as datasets of naturalistic conversations, such as the CANDOR dataset (Reece et al., 2023).

Few prior works have provided multimodal datasets for empathy tasks alone, and most prior works in empathy benchmarking are text-only. Table 1 shows a summary of the most relevant datasets compared to our EMPATHICSTORIES++ dataset. One dataset, the OMG-Empathy dataset measures the emotional effect stories have on the listener (Barros et al., 2019), but contains a limited amount of data collected from in-lab settings. Two recent works present more substantial datasets: MEDIC, which contains video clips annotated with 3 labels to describe empathy between counselors and clients in psychotherapy sessions (Zhou'an_Zhu et al., 2023), and a motivational interviewing dataset for assessing therapist empathy (Tran et al., 2023). Prior works also provide datasets related to empathy focusing on single modalities. The EmpatheticDialogues dataset (Rashkin et al., 2019), the EDOS dataset (Welivita et al., 2021), the EmpathicStories dataset (Shen et al., 2023a), the Empathic Conversations dataset (Omitaomu et al., 2022), and Sharma et al. (2020) contain text-only benchmarks for empathetic conversations and stories. A few datasets focus on empathy and emotion in nonverbal contexts only, such as the EyeT4Empathy dataset (Lencastre et al., 2022) and iMiGUE dataset (Liu et al., 2021), which use gaze and gesture respectively.

In contrast to these prior works, our dataset is the first dataset that focuses on empathy in relation to past experiences, and is collected in-the-wild, over a long term deployment with longitudinal survey and interaction data, and contains self-annotated

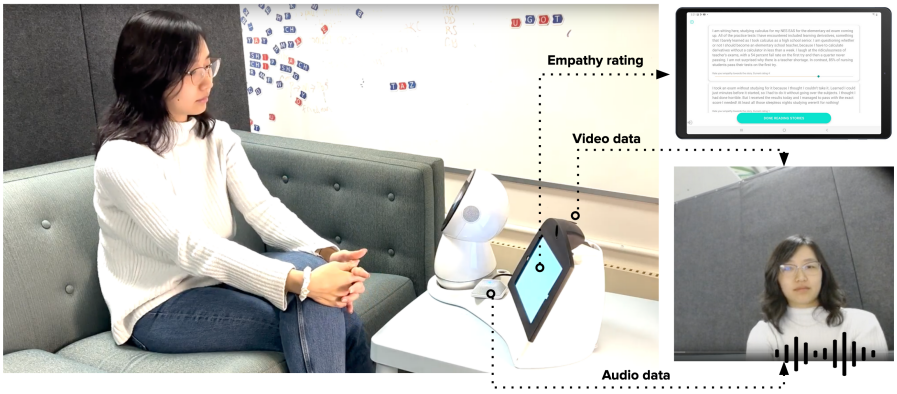


Figure 2: **Data collection setup.** The robot station houses a webcamera and microphone for video/audio data collection. A tablet displays stories read by participants, as well as sliders for self-rating empathy on a scale of 1-5.

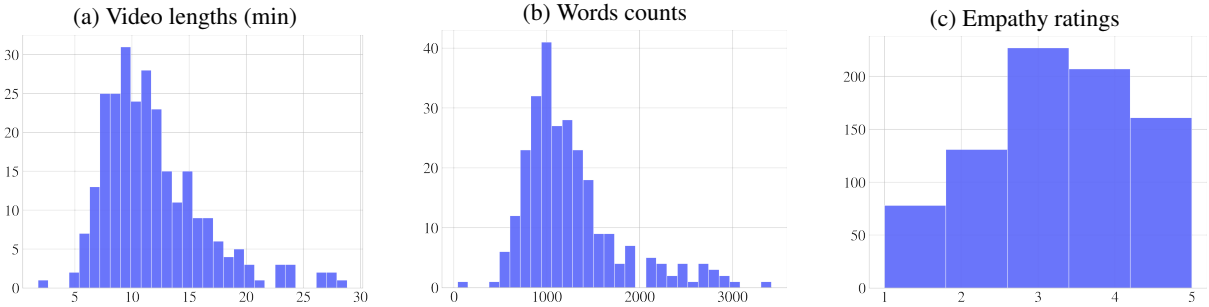


Figure 3: **Basic dataset statistics.** Video length and word count statistics of all participant sessions, as well as the distribution of self-rated empathy labels.

empathy ratings.

## 3 Data Collection

We deployed 46 in-home robots, powered by ChatGPT,[2] to converse with participants and record data. We recruited participants through mailing lists, and participants explicitly consented to data sharing. Our protocol was approved by our institution's ethics review board. Five participants withdrew from data collection for reasons not related to the study protocol. Data collection took place over the course of a month, and participants were asked to complete between 6-12 conversation sessions with the robot (compensated $60 for 12 sessions). Figure 2 shows the robot station in the participants' home and our data collection setup. The use of robots for data collection normalizes speaker-dependent characteristics that could add noise to the data from in-lab, human-human studies or acted scenarios (Wood et al., 2013b,a). While one might hypothesize that the use of a robot would users less expressive, prior work shows that embodied social agents still elicit empathy behaviors similar to that of human-human interaction (Spitale et al., 2022; Wood et al., 2013b). We use the social robot to scaffold the interaction while still allowing for natural conversation. Within each session, participants were guided through a conversation with the agent using the following scheme.

1. **Warm up phase.** At the beginning of each section, the participant warms up to the robot through casual conversation about their day or the previous robot-participant interaction.

2. **Story share phase.** In this phase, the robot prompts the user to share a meaningful story from their journal or on their mind.

3. **Story receive phase.** The robot then addresses the user's shared story by responding empathically, and retrieves 3 stories that the user might empathize with, using the empathic similarity retrieval model from Shen et al. (2023a).

4. **Story reflection phase.** We carefully designed reflection prompts based on narrative therapy approaches and emotion regulation (Gardner and Poole, 2009; White and Epston, 1990; Yoosefi Looyeh et al., 2014). Next, the

[2] https://chat.openai.com/

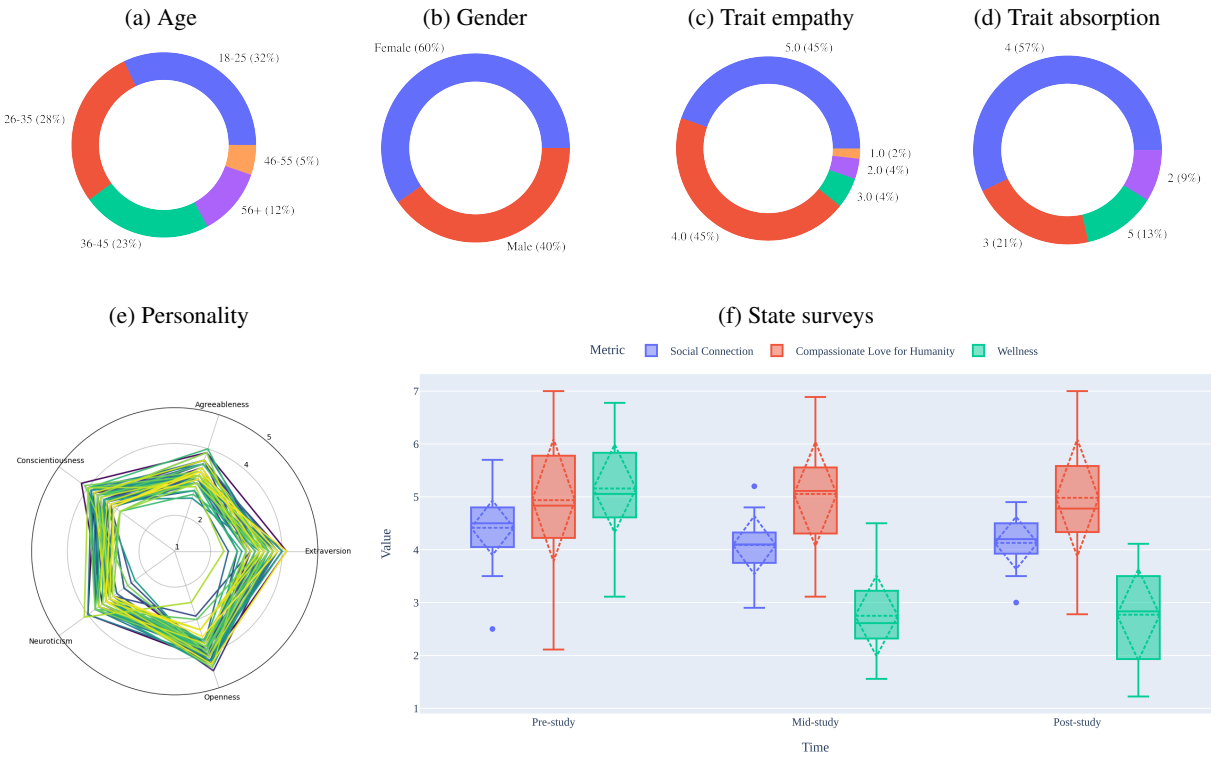


Figure 4: **Trait and state surveys**. Participant demographic information, trait and state survey overviews show diversity across age, gender, personality type, and feelings of social connection and wellbeing over time.

robot asks the participant to reflect on the following four areas: ways in which they related to the narrator, identifying the emotions of the narrator, regulating or comforting the narrator, and high-level takeaways from the story that the participant could apply to their own life.

5. **Cool-down phase.** Finally, the agent summarizes the session and thanks the participant.

**Self-Report Survey Measures** We collected self-reported measurements before the study, two weeks into the study, and at the one-month point. During our pre-study questionnaires, we administered the following *trait* surveys: the Big 5 Personality Test (Goldberg, 1993), the absorption scale dimensions of the Multidimensional Personality Questionnaire (measure ability to absorb into fictional experiences) (Cain et al., 2015), the Single Item Trait Empathy Scale (Konrath et al., 2018), and the following *state* surveys: the Compassionate Love for Humanity Scale (Sprecher and Fehr, 2005), and the UBC State Social Connection Scale (Lok and Dunn, 2022). Note that we use both the Compassionate Love for Humanity Scale and the UBC State Social Connection to measure overall "social connectedness." For the mid-study and post-study questionnaires, all *state* surveys were repeated.

**Interaction Data** In the *Story receive phase*, participants read three personal stories retrieved based on the user's own story. On the tablet, users rated their empathy toward each story using a slider on a scale of 1-5 (low to high).

**Video and Audio Recordings** During the study, each station completely recorded each interaction session, using the station's built in Logitech 1080p webcam and MXL AC-44 USB Boundary microphone to obtain high-quality recordings of the participant's face and voice. Note that we made clear when the system was recording the user in our study onboarding and through the robot's ring light.

**Transcripts** Transcripts of all utterances by the robot and the participant were saved on a Firebase Realtime Database. The transcripts were obtained in real-time using the AssemblyAI streaming ASR.

**Feature Extraction** For each label, we trimmed the associated video clip to fit a context window of 120 frames ($k = 120$). This gives us 8 seconds worth of video context for videos that play at 15 frames per second. To augment the video clips, we've applied a sliding window technique every second. Consequently, this has yielded us a total of 99,357 clips for the *Story share* phase, and 84,705 samples for the *Reflection* phase.

- **Vision.**[3] We use the normalized eye gaze direction, location of the head, location of 3D landmarks, and facial action units extracted from OpenFace (Baltrušaitis et al., 2016). We also extract frame-wise image features from the penultimate layer of ResNet50 (He et al., 2015). The two feature vectors (obtained from OpenFace and ResNet50) are concatenated per timestep to be used as the final visual input (dimension/timestep is $F = 2762$).

- **Audio.** We use openSMILE (Eyben et al., 2010) to extract low level acoustic features (i.e. loudness, alpha ratio, etc., $F = 65$)

- **Language.** We convert video transcripts and story contents into text embeddings via pre-trained Glove (glove.840B.300d) (Pennington et al., 2014) word embedding and Sentence BERT (Reimers and Gurevych, 2019) ($F = 300$, $F = 384$ respectively).

## 4 Dataset Statistics and Properties

The EMPATHICSTORIES++ dataset comprises video, audio, and text data from 269 sessions collected from 41 distinct participants, along with self-reported survey and interaction data. Each video is a .avi file recorded at 15fps, whose cumulative length is 3,180 minutes (53 hours). The total number of utterances is 53,80, or about 20 per session (fixed for each interaction phase), totalling 337,147 words (1,258 per session).

Figure 3a shows the distribution of video lengths across sessions, ranging from 2 to 29 minutes (mean = 12 min, s.d. = 4.5 min). Figure 3b depicts a similar distribution for spoken word counts. These ranged between 40 and 3418 words (mean = 1258 words, s.d. = 531 words). Participants felt varying levels of empathy towards with the stories they received, as the distribution (Figure 3c) of their empathy ratings on 1-5 scale shows (mean = 3.3, s.d. = 1.2).

Figure 4 depicts the demographic information of the participants. Figures 4a-4d show the distributions of age, gender, trait empathy, and trait absorption. Measurements of the Big 5 personality traits are shown in the radar chart Figure 4e. The change in levels of Social Connection, Compassionate Love for Humanity, and Wellness (see Figure 3) across the month-long study are shown in Figure 4f. Participants shared vulnerable and meaningful stories across diverse topics (Appendix A).

Our dataset is notable in that it (1) **is captured in-the-wild**, in participants' homes (2) **contains longitudinal data, with trait and state surveys**, and (3) **is self-annotated**, which is crucial for a subjective psychological process like empathy.

## 5 Experiments

### 5.1 Task Definition

We formulate the multimodal empathy prediction task as follows: At time $t$, where $t$ is the timestep in which we want to predict each participant's empathy levels for the story, we are given the $[t - k/2, ..., t + k/2]$ interval of contextual video information (during the *Story Share* and *Reflection* phases), where $k$ is the number of context frames.

For each clip, we extract features from three modalities: text, audio, and video. Each modality has distinct temporal and feature dimension, denoted as $T_{\{V,A,T\}} \times F_{\{V,A,T\}}$. The corresponding contextual behavior features for each modality can be viewed as $X_T \in \mathbb{R}^{T_T \times F_T}$, $X_A \in \mathbb{R}^{T_A \times F_A}$, and $X_V \in \mathbb{R}^{T_V \times F_V}$, respectively. The comprehensive multimodal feature set is represented as $X = [X_T, X_A, X_V]$. Finally, we train a model $f_\theta(\cdot)$ that takes $X$ as input and outputs a multimodal representation $Z = f_\theta(X)$, which is further used to calculate empathic similarity score $sim(Z, E(S_i))$ where $sim(\cdot)$ is a similarity metric (e.g., cosine similarity), and $E(S_i)$ is the embedding of the $i$th story $S_i$ ($i = 1, 2, 3$). Finally, this similarity score is compared with the empathic label $y$ to calculate the loss.

### 5.2 Models

**Attention-based multimodal Emotion Reasoning model (AMER) (Shen et al., 2020):** AMER is a model designed to facilitate the task of multimodal emotion reasoning in videos. It employs an attention-based approach to model intra- and inter-personal emotion contexts, propagation, and prior knowledge of personalities.

**Tensor Fusion Network (TFN) (Zadeh et al., 2017):** TFN is a representative tensor-based network, initially developed for multimodal sentiment

[3] As illustrated in Figure 1, we additionally provide the whole-body (bodies, hands and faces) 2D/3D poses obtained from DOPE (Weinzaepfel, Philippe and Brégier, Romain and Combaluzier, Hadrien and Leroy, Vincent and Rogez, Grégory) in our dataset.

Table 2: **Model performance for empathy prediction in *Story Share* scenario** across correlation, accuracy, and retrieval metrics. $r$ = Pearson's correlation, $\rho$ = Spearman's correlation, $Acc$ = Accuracy, $F1$ = Binary F1-score, and $MSE$ = Mean Squared Error. Note that all scores are multiplied by 100 for easier comparison. For each column, the best result is **bolded**, and the second best is underlined.

| Model | | $r$ (↑) | $\rho$ (↑) | $Acc$ (↑) | $F1$ (↑) | $MSE$ (↓) |
|---|---|---|---|---|---|---|
| AMER (Shen et al., 2020) | $t$ | 5.500 ± 0.800 | 5.500 ± 0.800 | 53.400 ± 0.100 | 38.900 ± 0.600 | 25.200 ± 0.000 |
| | $v+a$ | 6.300 ± 0.300 | 6.300 ± 0.300 | 52.500 ± 1.600 | 40.000 ± 0.600 | 25.800 ± 0.300 |
| | $v+t$ | 4.000 ± 1.000 | 4.000 ± 1.000 | 51.800 ± 0.200 | 38.400 ± 0.700 | 25.800 ± 0.100 |
| | $a+t$ | 6.800 ± 0.200 | 6.800 ± 0.200 | 54.100 ± 0.200 | 39.600 ± 0.100 | 25.500 ± 0.000 |
| | $v+a+t$ | 10.500 ± 7.000 | 10.500 ± 7.000 | 51.700 ± 0.700 | 43.000 ± 4.900 | 26.400 ± 0.900 |
| TFN (Zadeh et al., 2017) | $t$ | 11.000 ± 2.900 | 11.000 ± 2.900 | 55.100 ± 1.700 | 41.200 ± 1.600 | 24.300 ± 0.100 |
| | $v+a$ | 0.200 ± 6.200 | 0.200 ± 6.200 | 50.700 ± 2.800 | 34.600 ± 3.700 | 24.400 ± 0.200 |
| | $v+t$ | -4.700 ± 10.900 | -4.700 ± 10.900 | 48.100 ± 4.900 | 32.000 ± 6.300 | 24.400 ± 0.500 |
| | $a+t$ | -4.400 ± 9.300 | -4.400 ± 9.300 | 48.600 ± 4.500 | 32.100 ± 5.200 | 24.400 ± 0.200 |
| | $v+a+t$ | -1.900 ± 9.300 | -1.900 ± 9.300 | 50.200 ± 3.000 | 33.100 ± 6.300 | **24.200** ± 1.000 |
| EF-LSTM (Hochreiter and Schmidhuber, 1997) | $t$ | 3.000 ± 2.300 | 3.000 ± 2.300 | 52.500 ± 1.000 | 37.300 ± 1.400 | 25.300 ± 0.200 |
| | $v+a$ | 4.800 ± 3.400 | 4.800 ± 3.400 | 51.300 ± 0.700 | 39.300 ± 2.300 | 26.000 ± 0.200 |
| | $v+t$ | 7.900 ± 1.300 | 7.900 ± 1.300 | 50.200 ± 2.000 | 42.000 ± 1.300 | 26.600 ± 0.700 |
| | $a+t$ | 2.800 ± 2.100 | 2.800 ± 2.100 | 52.300 ± 0.700 | 37.200 ± 1.400 | 25.400 ± 0.100 |
| | $v+a+t$ | 7.400 ± 1.000 | 7.400 ± 1.000 | 51.200 ± 2.700 | 41.400 ± 0.400 | 26.300 ± 0.700 |
| LF-LSTM (Hochreiter and Schmidhuber, 1997) | $t$ | 3.400 ± 0.100 | 3.400 ± 0.100 | 52.600 ± 0.100 | 37.600 ± 0.000 | 25.100 ± 0.000 |
| | $v+a$ | 6.400 ± 1.600 | 6.400 ± 1.600 | 46.000 ± 0.600 | 42.500 ± 0.900 | 27.800 ± 0.200 |
| | $v+t$ | 5.800 ± 4.600 | 5.800 ± 4.600 | 47.200 ± 2.800 | 41.800 ± 2.000 | 27.600 ± 0.600 |
| | $a+t$ | 2.200 ± 1.300 | 2.200 ± 1.300 | 52.300 ± 0.700 | 36.700 ± 0.800 | 25.300 ± 0.000 |
| | $v+a+t$ | 8.200 ± 5.000 | 8.200 ± 5.000 | 48.100 ± 0.800 | 42.800 ± 3.300 | 27.200 ± 0.800 |
| EmpathicStoriesBART (Shen et al., 2023a) | $t$ | 2.400 ± 0.000 | 2.400 ± 0.000 | 80.700 ± 0.000 | 35.500 ± 0.000 | 51.900 ± 0.000 |
| GPT-4 (OpenAI, 2023) | $t$ | **23.200** ± 1.600 | **17.600** ± 1.400 | **82.500** ± 0.000 | **50.600** ± 0.700 | 32.200 ± 0.300 |

analysis. It carries out an outer tensor-product operation on the embeddings of modalities to create a unified multimodal space.

**Late-Fusion LSTM (LF-LSTM) (Hochreiter and Schmidhuber, 1997):** LF-LSTM is a model that separately constructs LSTMs for linguistic, visual, and acoustic inputs. It fuses the final hidden states of these three LSTMs, creating a comprehensive sentence-level multimodal representation.

**Early-Fusion LSTM (EF-LSTM) (Hochreiter and Schmidhuber, 1997):** EF-LSTM assembles linguistic, visual, and acoustic features at each time step, utilizing an LSTM to construct a sentence-level multimodal representation.

**EmpathicStoriesBART (Shen et al., 2023b):** EmpathicStoriesBART is a distinctive model fine-tuned to compute empathic similarity in personal narratives using three key story features. Validated in a user study, it outperforms traditional semantic similarity models, highlighting its potential for our task.

**GPT-4 (OpenAI, 2023):** GPT-4, a state-of-the-art closed-source language model capable of deep contextual understanding and producing highly relevant responses. GPT models have been evaluated for empathetic response generation (Lee et al.).

Implementation details and prompts are included in Appendix B and C.

## 6 Results and Discussion

### 6.1 Automatic Evaluation

To evaluate the quality of empathy predictions, we follow previous work (Shen et al., 2023a) and report Pearson's correlation, Spearman's correlation, accuracy, F1-scores and the mean squared error. For correlations, we calculate the cosine similarity between the multimodal representation and the embedding of the stories and compare these similarity scores with the human-rated empathy labels. For interpretability, we split the scores into binary similar/dissimilar categories and compute the accuracy and $F1$ scores.

Table 2 shows the performance of state-of-the-art multimodal (video, audio and text) models when given the user's *Story Share* context (video and audio) + the story they read (text) as inputs, and their empathy ratings as labels. In the context of *Story Share*, GPT-4 showed the highest Pearson's correlation ($r$ = 0.232) and Spearman's correlation ($\rho$ = 0.176) with $t$-only input. Notably, it also recorded the highest accuracy ($Acc = 0.825$) and $F1$-score ($F1$ = 0.506) which aligns well with the observation that participants in the *Story Share* setting were more focused on conveying their story, rather than on expressive verbal and non-verbal behaviors. Conversly, the performance of models in the context of user *Reflection* (reflections on a read story) is outlined in Table 3. Here, LF-LSTM demonstrated the highest Pearson's correlation ($r$ = 0.560) and Spearman's correlation ($\rho$ = 0.559) with $v$+$t$ inputs. While GPT-4 continued to show the highest accuracy and F1-scores, it's worth noting that among multimodal models, AMER showed comparable

performance ($Acc = 0.688$, $F1 = 0.665$) even with a significantly smaller number of parameters and using only audio with text inputs.

### 6.2 Ablation Studies

Here, we analyze the influence of various input modalities on six models in both *Story Share* and *Reflection* settings, focusing particularly on the impact of text-only inputs.

In the *Story Share* scenario, across different models and input modalities, no significant performance improvements were observed as we add more input modalities to text. Interestingly, using $t$-only input showed the best performance in $Acc$ across all multimodal models. In contrast, in the *Reflection* scenario, where both verbal and non-verbal expressions plays a vital role, AMER showed remarkable performance improvements (26.90% in $Acc$) when adding $a$ to $t$ and 14.02% for EF-LSTM when using $v$+$a$ inputs. Also, by adding $v$ to $t$, all multimodal model showed performance improvements (10.36% for $Acc$ and 26.28% for $F1$ in average). However, EmpathicStoriesBART and GPT-4 model, which solely use $t$-only input, outperforms all other models, achieving an impressive accuracy of 0.737. This significant performance, combined with a high $F1$ score of 0.762 and 0.750, underscores the potential of task and context specificity and the use of key story features to identify moments of empathy. To confirm the robustness of these results, we applied majority voting based on the label distribution, which resulted in an accuracy of 0.750 and an $F1$ score of 0.839.

## 7 Conclusion

This paper presents EMPATHICSTORIES++, the first in-the-wild, long-term, multimodal dataset on empathy towards personal experiences, which can be used to quantitatively evaluate empathy as it relates to one's past experiences. Our dataset is self-annotated with empathy ratings and psychometric surveys. We present and benchmark a task on predicting user empathy from their interaction contexts. We observe that modality selection impacts model performance and is context-dependent. In the *Story Share* phase, where textual context was dominant, GPT-4 with text input performed the best in most metrics. However, in the *Reflection* phase, where introspective verbal and non-verbal expressions are abundant, using $v$+$t$ inputs showed 26.28% improvement in average for $F1$ score, demonstrating their proficiency in extracting meaningful insights from multi-modal inputs. Our work provides a valuable resource for future work in empathetic AI, quantitative exploration of cognitive insights, and empathy modeling. We publicly release our dataset to foster advancements in social-emotional AI.

## Acknowledgments

We would like to thank the participants of our work for contributing to this dataset. We would also like to thank Jon Ferguson and Audrey Lee for their technical contributions in deploying the robot stations for data collection. This work was supported by an NSF GRFP under Grant No. 2141064

## Ethics Statement

Our dataset contains intimate, personal stories and video-audio data of participants necessary for modeling empathetic response. However, this type of naturalistic data is sensitive and private. As such, we made sure participants explicitly consented with data sharing, and our protocol was approved by our institutions ethics review board. Furthermore, in the design of our robot station, we made sure it was clear whenever the robot was listening (through a blue ring light) and that data would only be recorded during sessions, not the entire duration the robot was in a participant's home. We made sure to store videos on a private, lab-hosted server. For transparency we note that 17% of participants mentioned concerns of the robot infringing their privacy/security during our post-study interviews. While our dataset contains intimate information, we believe that such a resource is necessary in advancing science about empathy, which by nature occurs in personal and natural settings. We will ensure that distribution of the dataset is only granted upon ethics review board approval, and that the dataset is only used towards the goal of furthering positive empathy research in the future.

## Limitations and Future Work

The main limitation of our work is the limited sample size afforded by use of a physical robot. However, we believe the use of an embodied agent is essential for our data collection, as embodied agents provide experiences closer to that of natural human interaction than virtual interactions. Future work can use our system to replicate the data collection through a physically embodied robot.

Table 3: **Model performance for empathy prediction in the *Reflection* scenario.** For each column, the best result is **bolded**, and the second best is underlined.

| Model | | $r$ (↑) | $\rho$ (↑) | $Acc$ (↑) | $F1$ (↑) | $MSE$ (↓) |
|---|---|---|---|---|---|---|
| AMER (Shen et al., 2020) | $t$ | 5.400 ± 0.700 | 5.300 ± 0.700 | 53.900 ± 1.400 | 43.500 ± 1.600 | 23.800 ± 0.700 |
| | $v+a$ | 36.500 ± 0.500 | 36.600 ± 0.500 | 68.400 ± 0.100 | 65.400 ± 0.500 | 22.500 ± 0.000 |
| | $v+t$ | 40.000 ± 0.400 | 39.900 ± 0.400 | 67.800 ± 0.200 | 66.300 ± 6.600 | **22.400** ± 0.000 |
| | $a+t$ | 37.300 ± 0.200 | 37.200 ± 0.100 | 68.800 ± 0.000 | 66.500 ± 0.100 | 22.500 ± 0.000 |
| | $v+a+t$ | 36.700 ± 0.500 | 36.800 ± 0.600 | 68.400 ± 0.400 | 65.400 ± 1.100 | 22.500 ± 0.100 |
| TFN (Zadeh et al., 2017) | $t$ | 0.500 ± 4.800 | 0.500 ± 4.200 | 51.100 ± 2.600 | 32.200 ± 3.200 | 23.700 ± 0.100 |
| | $v+a$ | 2.300 ± 4.600 | 2.400 ± 4.500 | 49.800 ± 1.100 | 32.600 ± 2.000 | 23.800 ± 0.300 |
| | $v+t$ | -0.100 ± 5.500 | 0.000 ± 5.400 | 52.000 ± 3.300 | 32.700 ± 4.700 | 23.600 ± 0.200 |
| | $a+t$ | 5.100 ± 4.600 | 5.100 ± 4.600 | 51.500 ± 1.400 | 30.000 ± 1.300 | 23.800 ± 0.200 |
| | $v+a+t$ | 8.400 ± 4.000 | 8.300 ± 4.000 | 49.400 ± 0.800 | 32.800 ± 0.400 | 23.900 ± 0.200 |
| EF-LSTM (Hochreiter and Schmidhuber, 1997) | $t$ | 6.000 ± 0.700 | 5.900 ± 0.700 | 53.500 ± 0.200 | 30.000 ± 0.600 | 23.400 ± 0.100 |
| | $v+a$ | 24.500 ± 1.400 | 24.500 ± 1.400 | 61.100 ± 1.400 | 44.900 ± 0.900 | 23.400 ± 0.600 |
| | $v+t$ | 22.600 ± 4.300 | 22.400 ± 3.400 | 58.700 ± 3.500 | 43.900 ± 2.400 | 24.100 ± 1.100 |
| | $a+t$ | 2.500 ± 4.200 | 2.500 ± 4.200 | 54.900 ± 1.500 | 31.000 ± 2.700 | 23.400 ± 0.000 |
| | $v+a+t$ | 20.500 ±4.100 | 20.500 ±4.100 | 57.300 ±3.900 | 42.800 ±2.200 | 24.300 ±1.000 |
| LF-LSTM (Hochreiter and Schmidhuber, 1997) | $t$ | 3.100 ± 0.500 | 3.100 ± 0.500 | 55.400 ± 0.400 | 31.200 ± 1.200 | 23.200 ± 0.000 |
| | $v+a$ | 1.900 ± 5.300 | 1.900 ± 5.300 | 54.100 ± 2.000 | 30.900 ± 3.300 | 23.500 ± 0.300 |
| | $v+t$ | **56.000** ± 4.900 | **55.900** ± 4.900 | 56.700 ± 1.700 | 32.700 ± 3.200 | 23.100 ± 0.100 |
| | $a+t$ | 2.400 ± 0.500 | 2.400 ± 0.500 | 54.900 ± 0.700 | 30.900 ± 0.100 | 23.300 ± 0.100 |
| | $v+a+t$ | 6.800 ±8.700 | 6.800 ±8.700 | 57.200 ±3.900 | 33.600 ±5.500 | 23.000 ±0.200 |
| EmpathicStoriesBART (Shen et al., 2023a) | $t$ | 32.700 ± 0.000 | 34.000 ± 0.000 | **73.700** ± 0.000 | **76.200** ± 0.000 | 34.900 ± 0.000 |
| GPT-4 (OpenAI, 2023) | $t$ | 50.800 ± 0.000 | 46.000 ± 0.000 | **73.700** ± 0.000 | 75.000 ± 0.000 | 30.000 ± 0.000 |

Another limitation of our experimental results is that we only ablated contribution of modalities, but did not further interpret behavioral cues that might influence model performance. As such, these results are less interpretable due to lack of additional fine-grained annotations. Future work can obtain fine-grained annotations of the video data for empathy-relevant behavioral cues such as arousal, valence, self disclosure, etc.

Our dataset is a valuable resource for furthering research in empathy modeling for AI systems. Novel future directions to explore could include personalized modeling of empathy patterns, using the longitudinal data as well as understanding cognitive insights behind when empathy arises in personal story sharing.

## References


Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10.

Pablo Barros, Nikhil Churamani, Angelica Lim, and Stefan Wermter. 2019. The OMG-Empathy Dataset: Evaluating the Impact of Affective Behavior in Storytelling. pages 1–7. IEEE Computer Society.

Susan Bluck. 2003. Autobiographical memory: Exploring its functions in everyday life. *Memory*, 11(2):113–123. Publisher: Routledge _eprint: https://doi.org/10.1080/741938206.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

N. M. Cain, M. R. Lukowitsky, and A. G. Wright. 2015. Multidimensional Personality Questionnaire (MPQ).

Huili Chen, Sharifa Alghowinem, Soo Jung Jang, Cynthia Breazeal, and Hae Won Park. Dyadic Affect in Parent-child Multi-modal Interaction: Introducing the DAMI-P2C Dataset and its Preliminary Analysis. page 16.

Elisa Ciaramelli, Francesco Bernardi, and Morris Moscovitch. 2013. Individualized Theory of Mind (iToM): When Memory Modulates Empathy. *Frontiers in Psychology*, 4.

Mark Davis. 2004. Empathy: Negotiating the Border Between Self and Other. pages 19–42.

Jean Decety and Claus Lamm. Human Empathy Through the Lens of Social Neuroscience. *The Scientific World Journal*, 6:1146–1163. Publisher: Hindawi.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Brendan Gaesser. 2013. Constructing Memory, Imagination, and Empathy: A Cognitive Neuroscience Perspective. *Frontiers in Psychology*, 3.

Paula J. Gardner and Jennifer M. Poole. 2009. One Story at a Time: Narrative Therapy, Older Adults, and Addictions. *Journal of Applied Gerontology*, 28(5):600–620. Publisher: SAGE Publications Inc.

Lewis R. Goldberg. 1993. The structure of phenotypic personality traits. *American Psychologist*, 48(1):26–34. Place: US Publisher: American Psychological Association.

Richaed F Haase and Donald T Tepper. NONVERBAL COMPONENTS OF EMPATHIC COMMUNICATION.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

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

Sara D. Hodges, Kristi J. Kiel, Adam D. I. Kramer, Darya Veach, and B. Renee Villanueva. 2010. Giving Birth to Empathy: The Effects of Similar Experience on Empathic Accuracy, Empathic Concern, and Perceived Empathy. *Personality and Social Psychology Bulletin*, 36(3):398–409. Publisher: SAGE Publications Inc.

Sara Konrath, Brian P. Meier, and Brad J. Bushman. 2018. Development and validation of the single item trait empathy scale (SITES). *Journal of Research in Personality*, 73:111–122.

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. Does GPT-3 Generate Empathetic Dialogues? A Novel In-Context Example Selection Method and Automatic Evaluation Metric for Empathetic Dialogue Generation. page 15.

Pedro Lencastre, Samip Bhurtel, Anis Yazidi, Gustavo B. M. e Mello, Sergiy Denysov, and Pedro G. Lind. 2022. EyeT4Empathy: Dataset of foraging for visual information, gaze typing and empathy assessment. *Scientific Data*, 9(1):752. Number: 1 Publisher: Nature Publishing Group.

Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. 2021. iMiGUE: An Identity-free Video Dataset for Micro-Gesture Understanding and Emotion Analysis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10626–10637, Nashville, TN, USA. IEEE.

Iris Lok and Elizabeth Dunn. 2022. The UBC State Social Connection Scale: Factor Structure, Reliability, and Validity. *Social Psychological and Personality Science*, page 19485506221132090. Publisher: SAGE Publications Inc.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100.

Sylvia A. Morelli, Matthew D. Lieberman, and Jamil Zaki. 2015. The Emerging Study of Positive Empathy. *Social and Personality Psychology Compass*, 9(2):57–68. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/spc3.12157.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. 2022. Empathic Conversations: A Multi-level Dataset of Contextualized Conversations. ArXiv:2205.12698 [cs].

Desmond C. Ong, Zhengxuan Wu, Tan Zhi-Xuan, Marianne Reddan, Isabella Kahhale, Alison Mattek, and Jamil Zaki. 2021. Modeling Emotion in Complex Stories: The Stanford Emotional Narratives Dataset. *IEEE Transactions on Affective Computing*, 12(3):579–594.

OpenAI. 2023. Gpt-4 technical report.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Pickard, Catherine Roster, and Yixing Chen. 2016. Revealing sensitive information in personal interviews: Is self-disclosure easier with humans or avatars and under what conditions? *Computers in Human Behavior*, 65:23–30.

David B. Pillemer. 1992. Remembering personal circumstances: A functional analysis. In *Affect and accuracy in recall: Studies of "flashbulb" memories*, Emory symposia in cognition, 4., pages 236–264. Cambridge University Press, New York, NY, US.

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. ArXiv:1810.02508 [cs].

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y.-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset. Technical Report arXiv:1811.00207, arXiv. ArXiv:1811.00207 [cs] type: article.

Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. 2023. The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13):eadf3197. Publisher: American Association for the Advancement of Science.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Mahnaz Roshanaei, Christopher Tran, Sylvia Morelli, Cornelia Caragea, and Elena Zheleva. 2019. Paths to Empathy: Heterogeneous Effects of Reading Personal Stories Online. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 570–579, Washington, DC, USA. IEEE.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. ArXiv:2210.13312 [cs].

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. SocialIQA: Commonsense Reasoning about Social Interactions. ArXiv:1904.09728 [cs].

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. ArXiv:2009.08441 [cs].

Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. 2020. Memor: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 493–502.

Jocelyn Shen, Maarten Sap, Pedro Colon-Hernandez, Hae Won Park, and Cynthia Breazeal. 2023a. Modeling Empathic Similarity in Personal Narratives. ArXiv:2305.14246 [cs].

Jocelyn Shen, Maarten Sap, Pedro Colon-Hernandez, Hae Won Park, and Cynthia Breazeal. 2023b. Modeling empathic similarity in personal narratives.

Gopendra Vikram Singh, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2023. EmoInt-Trans: A Multimodal Transformer for Identifying Emotions and Intents in Social Conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:290–300. Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Micol Spitale, Sarah Okamoto, Mahima Gupta, Hao Xi, and Maja J Matarić. 2022. Socially Assistive Robots as Storytellers That Elicit Empathy. *ACM Transactions on Human-Robot Interaction*, page 3538409.

Susan Sprecher and Beverley Fehr. 2005. Compassionate love for close others and humanity. *Journal of Social and Personal Relationships*, 22(5):629–651.

Trang Tran, Yufeng Yin, Leili Tavabi, Joannalyn Delacruz, Brian Borsari, Joshua Woolley, Stefan Scherer, and Mohammad Soleymani. 2023. Multimodal Analysis and Assessment of Therapist Empathy in Motivational Interviews.

year=2020 Weinzaepfel, Philippe and Brégier, Romain and Combaluzier, Hadrien and Leroy, Vincent and Rogez, Grégory, booktitle=ECCV. DOPE: Distillation Of Part Experts for whole-body 3D pose estimation in the wild.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A Large-Scale Dataset for Empathetic Response Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael White and David Epston. 1990. *Narrative Means to Therapeutic Ends*, 1st edition edition. W. W. Norton & Company, New York.

Luke Jai Wood, Kerstin Dautenhahn, Hagen Lehmann, Ben Robins, Austen Rainer, and Dag Sverre Syrdal. 2013a. Robot-Mediated Interviews: Do Robots Possess Advantages over Human Interviewers When Talking to Children with Special Needs? In *Social Robotics*, Lecture Notes in Computer Science, pages 54–63, Cham. Springer International Publishing.

Luke Jai Wood, Kerstin Dautenhahn, Austen Rainer, Ben Robins, Hagen Lehmann, and Dag Sverre Syrdal. 2013b. Robot-Mediated Interviews - How Effective Is a Humanoid Robot as a Tool for Interviewing Young Children? *PLOS ONE*, 8(3):e59448. Publisher: Public Library of Science.

Majid Yoosefi Looyeh, Khosrow Kamali, Amin Ghasemi, and Phuangphet Tonawanik. 2014. Treating social phobia in children through group narrative therapy. *The Arts in Psychotherapy*, 41(1):16–20.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-IQ: A Question Answering Benchmark for Artificial Social Intelligence. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8799–8809, Long Beach, CA, USA. IEEE.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ED: Multi-modal Multi-scene Multi-label Emotional Dialogue Database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5699–5710, Dublin, Ireland. Association for Computational Linguistics.

Zhou'an_Zhu, Xin Li, Jicai Pan, Yufei Xiao, Yanan Chang, Feiyi Zheng, and Shangfei Wang. 2023. MEDIC: A Multimodal Empathy Dataset in Counseling. ArXiv:2305.02842 [cs].

## A Story Topics

The topics in Figure 5 were obtained as follows: ada-v002 embeddings of stories were calculated via the OpenAI API, and a UMAP model was fit on the data to reduce the 1536 dimension vectors to $x$ and $y$ coordinates using cosine similarity as the distance measure and clusters were obtained with K-means.

## B Implementation Details

We train our models on 4 NVIDIA RTX A6000 with a batch size of 64 for 10 epochs. We use the AdamW (Loshchilov and Hutter, 2019) optimizer

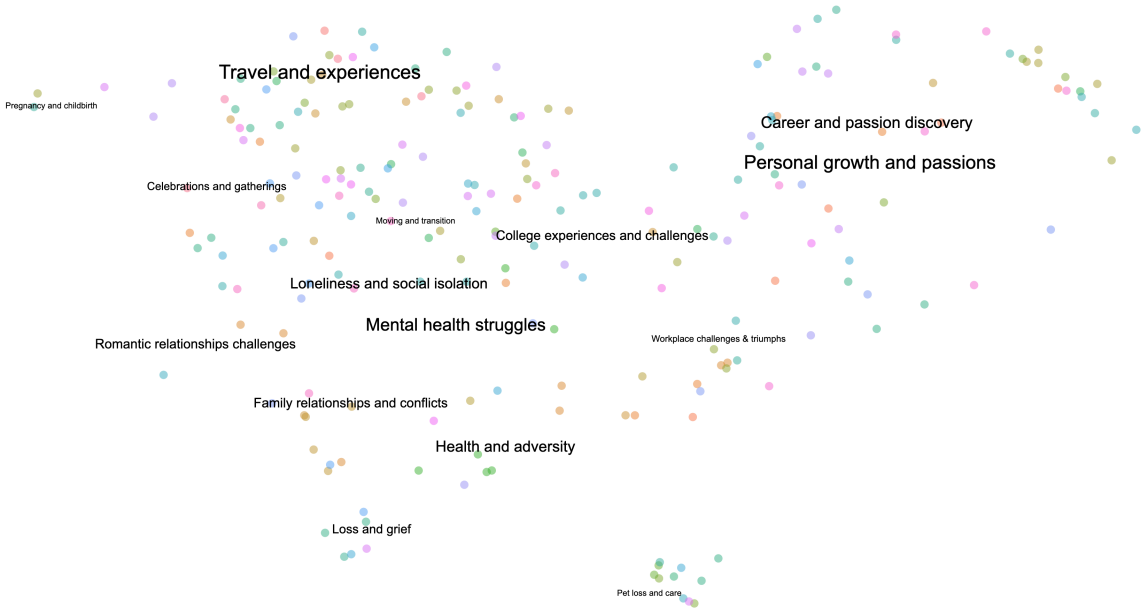


Figure 5: **Story Topics**: We visualize the embeddings (obtained with UMAP of ada-002 embeddings) of Story topics. Our deployment across the United States gives us a diverse set of meaningful personal stories.

with an initial learning rate of 1e-4 with a scheduler StepLR that decays the learning rate by 0.1 ($\gamma$) every 5 epochs (step_size). For the loss function, we use the Mean Squared Error (MSE). For the dataloader, we first conduct oversampling based on the empathy ratings due to its imbalance distribution as shown in Figure 3c. Next, we separate participants into train/valid/test sets in the ratio of 0.7/0.2/0.1 to ensure the model does not see the participants who were in the train sets. All models except for GPT-4 and EmpatheticStoriesBART were re-implemented to output multimodal representations that can be used to calculate similarities of story embeddings. We follow the default model parameters from the original implementations.

## C Prompting

We include prompts for GPT-4 benchmarking below:

**Story Sharing:**

- **System prompt:** *You are a psychologist with expertise in analyzing empathy. You can predict how much people might empathize with each other, based on their past experiences.*
- **User prompt:** *You will receive two stories, one from person A and the other from person B. Please predict, on a scale from 0 to 1, how much person A would empathize with B's story. Return just the number, no other text.*

**Reflection:**

- **System prompt:** *You are a psychologist with expertise in analyzing empathy. You can predict how much people might empathize with each other, based on their past experiences.*
- **User prompt:** *You will receive a story and conversation between person A and person B about person A's reflections about the story. Based on this, please predict, on a scale from 0 to 1, how much person A would empathize with the story. Return just the number, no other text.*