Empathy Prediction from Diverse Perspectives

Francine Chen, Scott Carter, Tatiana Lau, Nayeli Bravo
Sumanta Bhattacharrya[0], Kate Sieck, Charlene Wu
Toyota Research Institute
Los Altos, CA
[scott.carter,tatiana.lau]@tri.global

Abstract

A person's perspective on a topic can influence their empathy towards a story. To investigate the use of personal perspective in empathy prediction, we collected a dataset, EmpathyFromPerspectives, where a user rates their empathy towards a story by a person with a different perspective on a prompted topic. We observed in the dataset that user perspective can be important for empathy prediction and developed a model, PPEP, that uses a rater's perspective as context for predicting the rater's empathy towards a story. Experiments comparing PPEP with baseline models show that use of personal perspective significantly improves performance. A user study indicated that human empathy ratings of stories generally agreed with PPEP's relative empathy rankings.

1 Introduction

Empathy is "the ability to recognize, understand, and share the thoughts and feelings of another person, animal, or fictional character."[1] The ability to predict the degree to which a person will empathize with something can be important when trying to reduce inter-group and inter-personal conflict. Most proposed methods for empathy prediction assume that empathy can be predicted without consideration of individuals' perspectives. Often, an aggregate of empathy ratings are used. However, people will react with different levels of empathy to a given text passage, based in part on their background and current context. In addition, stories by others with diverging views about an issue may be ignored or brushed aside because they are so different from the reader's own perspective. Thus, individual perspective can be important for predicting a person's empathy.

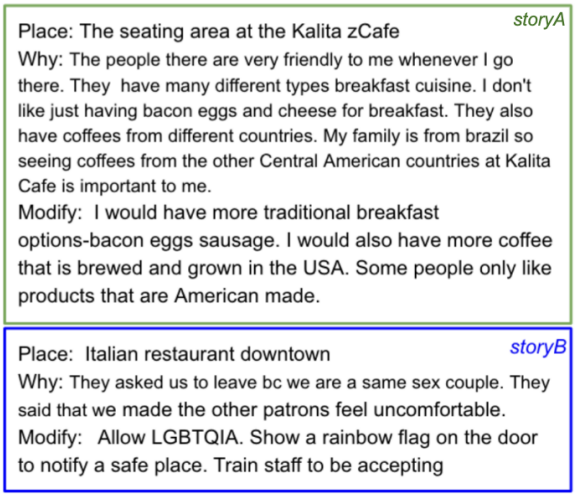

Figure 1: An example story pair. The green box contains a story where personA feels welcome. The blue box contains a story about a similar type of place where personB feels excluded.

Models for empathy prediction have been developed using data labeled by third parties (e.g., Barriere et al. (2023)). For some situations, prediction of the average empathy felt by multiple people towards a story, such as a political event, may be desirable. In other situations, however, predicting individual empathic reactions may be important, e.g., when negotiating a compromise with a person (Cohen, 2010) or finding ways to tell a story to maximize empathy (Gueorguieva et al., 2023). In these cases, people may have different levels of empathy to a given text passage. Our work takes a perspectivist approach (Abercrombie et al., 2024) in which data is not aggregated when labeled, but instead recognizes individual perspectives.

In this paper, we investigate predicting empathy from different perspectives. For this, we:

- created the Empathy from Perspectives (**EFP**) dataset for empathy prediction modeling and for examining the effect of individual perspective on empathy,

[0]Author is currently at the University of Illinois, Chicago
[1]https://www.psychologytoday.com/us/basics/empathy, also e.g., Hasson et al., 2022

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8959–8974

- proposed, to the best of our knowledge, a new task of using user perspective for first-person empathy prediction of a *divergent* perspective, and proposed a model for Personal Perspective Empathy Prediction (**PPEP**),
- conducted experiments (i) comparing PPEP, which uses story-as-perspective, to two previously proposed approaches to empathy prediction (ii) examining how different inputs, such as empathic similarity data, demographics, or knowing a person's rating for a different story, influence performance, and (iii) comparing the use of LLMs instead, and
- conducted a human subject study evaluating PPEP.

While we observed that prediction of absolute empathy (i.e., a person's empathy rating without regard to other empathy values) is poor when the rater perspective is new (unseen in training), our human subject study showed that PPEP's predicted **relative empathy** ratings, (i.e., the difference between a person's empathy ratings of two stories), generally agreed with human relative empathy ratings.

2 Related Work

Perspectivist approaches to NLP in general have been presented at NLPerspectives workshops (https://nlperspectives.di.unito.it/), and have been applied to prediction. For example, Plepi et al. (2022) predicted binary labels (e.g., whether the poster of a Reddit post is a "wrong doer") by an annotator towards a Reddit post using demographics and other Reddit posts as context. Here, we focus on empathy prediction of a rater who has written a story about a topic towards another person's divergent perspective story on a similar topic.

2.1 Empathy Prediction

There has been much work related to third-person empathy prediction (e.g., Shetty et al., 2024; Lahnala et al., 2022; Sharma et al., 2020; Lee et al., 2021). A number of researchers have participated in the WASSA shared empathy prediction tasks (summaries in Barriere et al., 2023, 2022; Tafreshi et al., 2021), which centers around predicting the "perceived empathy" of a person towards news stories. In Barriere et al. (2022), the empathy prediction task was based on a participant's written reaction to a news article. Barriere et al. (2023) collected both self-reported and third-party empathy annotations. Buechel et al. (2018) modeled empathy and distress together for self-reported annotation of news stories, comparing the use of Ridge regression, Feed-Forward, and CNN models, finding that CNN models had the best Pearson's correlation performance on their data. Srinivas et al. (2023) and Lu et al. (2023) embedded the input text and used a multi-layer-perception (MLP) for empathy prediction. Their models placed 3rd and 5th, respectively, in the WASSA 2023 empathy track.

Guda et al. (2021) proposed EmpathBERT and found that demographics improve empathy prediction as a binary task. In our work, we examine empathy prediction as a continuous variable, which allows for ranking of a set of stories. We examine the effect of demographics in our continuous rating value context. Hasan et al. (2024) examined demographics in empathy prediction of news articles using a BERT-based language model. They found that demographics (i.e., gender, education level, race, age, and income) improved Pearson r by 0.013 to 0.049 depending on whether demographics is given as text or both text and a number.

These empathy models predict the empathy of a text annotated as the *average* of several annotators (third-person) or self-reported on their own text. In contrast, we examine empathy prediction from each rater's point of view towards someone else's opposing view. To capture the rater's perspective towards another view, the empathy ratings in our EFP dataset are in first person and the model input includes a story written by the rater.

2.2 Empathic Similarity Prediction

Shen et al. (2023) proposed the task of predicting the empathic *similarity* of a *pair* of stories and created the EmpathicStories dataset, which we used in some of our experiments. Their model computed the cosine similarity of two story embeddings, either BART or SBERT. It was also used for empathy prediction on the EmpathicStories++ dataset (Shen et al., 2024), with low correlation performance.

In Shen et al. (2023) story pairs were collected from online sources and annotated for empathic similarity by annotators. In contrast, we examine how empathy prediction is influenced by the perspective of the person rating their empathy towards a story. Due to privacy concerns, the request form

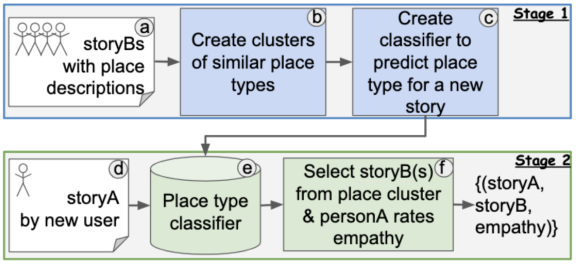

Figure 2: Steps for collecting the EmpathyFromPerspectives dataset. storyAs are positive and storyBs are negative.

for the EmpatheticStories++ dataset[2] states: "The login details will only be emailed to the given academic institute email address". We did not consider this dataset because of its limited availability.

3 EmpathyFromPerspectives Dataset

Our EmpathyFromPerspectives (EFP) dataset[3] contains pairs consisting of a positive story, **storyA** by **personA**, and a negative story, **storyB** by **personB**, plus personA's empathy rating for storyB. The data was collected as part of a larger study (Lau et al., 2024) focused on how people could create more inclusive public spaces, which are typically places where one may feel welcomed (i.e., places where you feel mentally at ease) or safe (i.e., physically at ease). The larger study's focus stems from current debates in the U.S. about how to design public spaces often being contentious, and empathy may be helpful when finding solutions.

3.1 Data Collection

The data was collected in two stages as shown in Figure 2. In **Stage 1** (top row), **storyB**s were collected from subjects recruited from dscout[4] (Fig. 2a). Each subject, **personB**, was asked to describe a *place* that made them feel either *safe*, *welcome*, *less safe* or *excluded* (four total). They were then asked *why* they felt that way and how they would *modify* a safe or less safe place to be safer, or a welcoming or excluding place to be more welcoming. The exact prompts are given in Appendix A.2.

To create story pairs about a similar type of place, place types were defined by clustering the storyB place descriptions using agglomerative clustering, so that each cluster represents a type of place (Fig. 2b). The safe and less safe places were clustered together and the welcoming and excluding places were clustered together. The number of safe/less safe clusters and the number of welcoming/excluding clusters were each set to 14. (See Appendix A.4 for a list of cluster/place types.) Then a place type classifier was developed (Fig. 2c) to identify which of the 14 clusters a new positive (i.e., welcoming or safe) story was most similar to.

In **Stage 2** (bottom row), a second set of participants were recruited from Prolific[5]. The new participants, **personA**, were asked to write a story, **storyA**, about a place that felt either *welcoming* or *safe* to them, using the same prompts used to collect the storyBs (Fig. 2d). Then the place classifier (Fig. 2e) identified the type of place written about and selected one or more negative stories from the place-type cluster of storyBs for personA to rate (Fig. 2f). Welcoming stories were matched to excluding stories, and safe stories were matched to less safe stories. PersonA then rated their empathy for one, three or five storyBs, with an equal number of personA's assigned to each condition. (See Appendix A.5 for details.) PersonA was also asked if they identified with storyB. The data was collected in two parts and we asked PersonAs in the second part to also provide demographics, i.e., gender, age, ethnicity, education, people they live with, and income. See Appendix A.1 for prompts and Appendix A for collection details.

3.2 Dataset Statistics

Table 1 shows statistics of the story pairs and empathy ratings for the EFP dataset and different subsets of EFP. The dataset was split 75/5/20 into train, dev and test sets and fixed for all experiments. Also shown is the EmpathicStories dataset (**ES**) (Shen et al., 2023) which was used in some of our experiments.

The EFP$_B$ dataset is the EFP dataset with storyA removed. We did not remove duplicate storyB's because for a given storyB, different personAs may assign very different empathy values (see Fig. 3). Thus in Table 1, the number of storyB empathy ratings is the same as that of EFP.

The EFP$_{dem}$ dataset is the subset of the EFP dataset for which demographics were collected. We used it to evaluate the usefulness of demographics.

[2] https://forms.gle/JfoLfiyeto7Zt9V86 accessed Sept 19, 2024

[3] available at https://osf.io/4szk5/

[4] https://dscout.com/

[5] https://www.prolific.com/

dataset	# ratings	# personA	# storyA	# storyB	place	why	modify	total
EFP	3234	1157	1336	301	9,6,8	48,43,29	37,32,23	94,87,48
EFP_B	3234	1157	0	301	8,6,8	42,38,26	37,34,22	88,81,47
EFP_{dem}	1922	792	948	268	8,6,7	48,43,28	36,32,23	92,85,46
ES	2000	–	–	–	–	–	–	235,–,–

Table 1: Datasets used in experiments. We collected the EmpathyFromPerspectives (EFP) dataset. EFP_B is EFP where storyA has been removed (i.e., storyB only). EFP_{dem} is the subset of EFP for which demographic information was collected. The ES dataset (EmpatheticStories) is from (Shen et al., 2023) which used 3rd-person annotation of empathic similarity. The EFP ratings are first-person empathy ratings by personA for personB's story. The word count statistics of mean, median and std dev are given for the three story parts—type of place written about (*place*), why the writer feels the way they do about a place (*why*), and how would they modify the place to make it more welcoming/safer (*modify*)—and the combined (*total*) story.

3.3 Empathy Values for a Story Vary by Rater

Figure 3 (left) shows the distribution of empathy values that personAs assigned to the set of storyB's in our training dataset when a story has been rated at least twice. These values varied from 1 to 100, which spans the full rating range. The mean value of empathy assigned to a story was 55, while the mean of the standard deviation of empathy values per story was 23. The range of values assigned to each storyB varied from about 1 to 100, with a mean range of 67. This relatively large range indicates that for the stories in our dataset, people often assign very different empathy values to the same story, and a single empathy value may not be very meaningful.

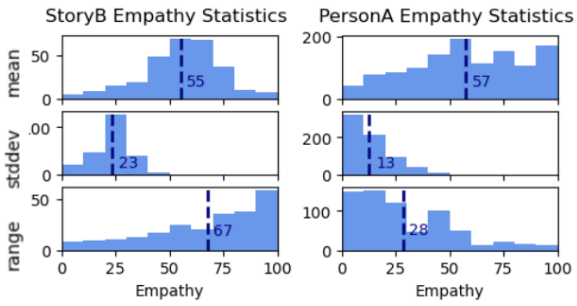

Figure 3: Histograms of storyB (left) and personA (right) empathy ratings statistics. The mean, stddev and range are computed per storyB from the assigned empathy ratings (left) or per person from their empathy ratings (right). The mean of each statistic is in dark blue.

3.4 Empathy Values Assigned by a Person

The distribution of empathy values by each personA who provided empathy ratings for at least two stories is shown in Figure 3 (right) for the training set. We note that while the mean empathy value assigned by a person who rated at least two stories varied over the full range, the mean of the standard deviations of personA ratings was only 13. Thus, the variation of empathy scores assigned by a person tends to vary less than the variation of scores assigned to a storyB. This is also reflected in the larger range of empathy values assigned to storyB than the range of empathy values individuals (personAs) assigned to stories. This suggests that information about personA can be helpful when predicting personA's empathy for a story.

4 Models for Empathy Prediction from Diverse Viewpoints

In this section we describe our Personal Perspective Empathy Prediction (PPEP) model, which uses the rater's perspective, and other models used in our experiments. For embedding text in all models, we used the pretrained SBERT sentence transformer (multi-qa-mpnet-base-dot-v1) (Reimers and Gurevych, 2019), which had the highest correlation performance in (Shen et al., 2023).

PPEP: Prediction with Rater's Viewpoint Our PPEP model is designed to predict the empathy of the author of storyA towards storyB. We used an MLP as a classifier that can capture storyA as context for storyB. The PPEP model is shown in Figure 4a. Each of the paired positive and negative stories are separately embedded using SBERT, which has an embedding size of 768, and concatenated. The combined embedding is input to an MLP classifier with three hidden layers, and the final output layer fed to a sigmoid for prediction. The number of nodes in each layer of the MLP was 768*2, 768, 384, and 192, with a ReLU between each linear layer. See Appendix E for details.

The following are baseline models and alternative representations examined in our experiments described in Section 5.

Prediction based on Similarity People tend to be more empathetic towards stories that are similar

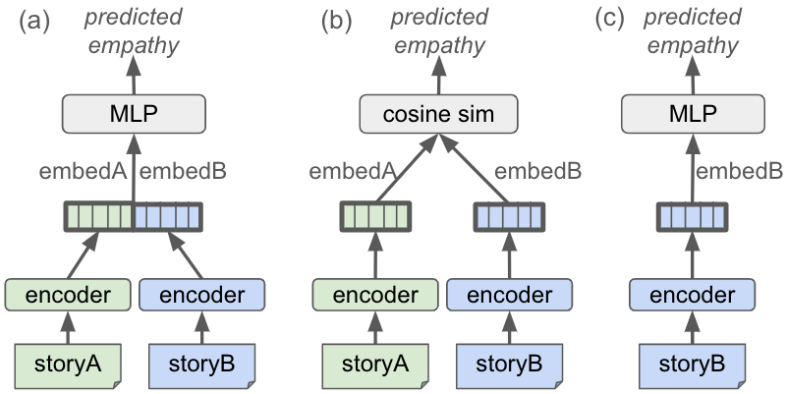

Figure 4: Empathy prediction models (a) PPEP model for computing perspective empathy (b) cosSim model for computing empathic similarity from Shen et al. (2023) (c) PPEP$_{noA}$ model where only storyB is input. (b) and (c) are baseline models. 'embedX' refers to the embedding of storyX.

to their own, which is modeled by the empathic similarity model by Shen et al. (2023). This model uses the cosine similarity of the SBERT embedded representations of storyA and storyB for prediction (Figure 4b). In contrast, we use storyA as context for prediction.

Prediction Given StoryB Only We examine empathy prediction without the rater's (opposing) perspective. For this, we removed the storyA branch of the model, embedding only storyB, the text to be rated (Figure 4c). The embedding is then fed into a two-layer MLP, since the embedding for one story is half the size of the concatenated embeddings for two stories. Thus the size of the layers is 768, 384 and 192, with a ReLU between each layer.

Prediction with Story Parts We compared the use of the full story (i.e., the responses to the three data collection prompts about place, why, and desired modifications) with use of modifications (mods) only. The responses/story parts were combined into one input text by either: using a separator '[SEP]' between story parts or preceding each story part with a prompt, e.g., "I feel safe here because". The prompts are given in Appendix C.

Prediction using Demographics The effect of adding demographics was compared to no demographics. We added demographics either as a categorical text value, each separated by [SEP] or with the subject demographics formatted as sentences. The template for formatting the demographics is: "I am a <gender>, "My age is <age-range>, "My education level is <education range>, "My household income is <income range>, "My ethnicity is <ethnicity>, "I live with <type of people in household>". Demographic statistics are shown in Appendix B.

Prediction using LLMs We asked Claude 3 Opus, GPT-4o and Claude 3.5 Sonnet to predict the empathy of the author of storyA towards storyB in 0-shot and 5-shot prompting. The prompt used (see Appendix D) is from Shen et al. (2024) for prompting GPT-4. We also added the phrase "where 0 is not empathetic and 1 is extremely empathetic" to the prompt to inform the LLMs of the desired ratings range, which the trained models could learn from the labeled data.

5 Model Evaluation

In our experiments, we (a) compared PPEP against baselines models on the EFP dataset, including using the cosineSimilarity of storyA and storyB (Shen et al., 2023, 2024), using only storyB and ignoring personA's perspective, and using the larger e5[6] embedding; and (b) examined different inputs, including adding the EmpathicStories dataset to EFP, using demographics, and adding the place and why story parts to the modifications story part. Finally, we also compared the performance of three LLMs for empathy prediction.

Our model is a modification of the Empathic Similarity code[7]. We used the same parameter values except for the number of training epochs, and the empathy labels were normalized to range from 0.0 to 1.0. MSE loss between the predicted empathy and human empathy ratings was used, and validation loss was computed at the end of each epoch. Early stopping on the validation data was used when training.

We did not binarize values when computing Spearman correlation (ρ) and Pearson correlation

[6] https://huggingface.co/intfloat/e5-large-v2
[7] https://github.com/mitmedialab/empathic-stories

	epochs	train	test	model	ρ(%)	r(%)	acc(%)	F1(%)	P(%)	R(%)	MSE
A	400*	EFP_B	EFP_B	$PPEP_{noA}$	1.95	2.73	51.99	54.44	62.13	48.45	0.1181
B	50	EFP	EFP	cosSim	30.51	29.50	63.96	71.09	67.68	**74.87**	**0.0836**
	100*	EFP	EFP	cosSim	34.18	33.52	66.26	**72.15**	70.54	73.83	0.0869
	400	EFP	EFP	cosSim	30.79	30.32	66.10	71.78	70.78	72.80	0.0960
C	100	EFP	EFP	PPEP	39.77	39.53	66.41	71.74	71.47	72.02	0.0857
	400*	EFP	EFP	PPEP	**40.12**	**40.04**	**66.87**	72.02	**72.02**	72.02	0.0898
D	400*	EFP	safe	PPEP	43.44	43.50	68.60	74.31	74.87	73.76	0.0767
	400*	EFP	welcome	PPEP	37.70	37.13	65.12	69.54	68.98	70.11	0.1031
	400*	safe	safe	PPEP	45.05	45.48	65.55	69.46	63.75	76.29	0.0858
	400*	welcome	welcome	PPEP	23.76	26.61	62.12	69.41	66.01	73.18	0.1089
E	100*	EFP+SEP	EFP+SEP	PPEP	38.64	38.10	66.26	71.50	71.50	71.50	0.0866
	100*	EFP+PRT	EFP+PRT	PPEP	38.32	37.87	67.33	72.73	71.90	73.58	0.0873
F	100*	EFP	EFP	e5+cosSim	26.97	27.23	63.34	69.86	68.06	71.76	0.0982
	100*	EFP	EFP	e5+PPEP	38.81	39.15	65.03	70.47	70.47	70.47	0.0876

Table 2: Empathy prediction performance when trained on the EmpathyFromPerspectives (EFP) dataset. PPEP is our MLP-based empathy prediction model. A: $PPEP_{noA}$ model, which ignores storyA perspective, takes only storyB input. B: Cosine similarity model. C: PPEP model. D: Performance comparison of safe and welcome subsets when trained together on EFP vs. training on a safe or welcome subset alone. E: Effect of including the place and why story parts with separators (SEP) or prompts (PRT). F: Comparison of cosine similarity and PPEP models using e5-large-v2 embedding. Best values when models are tested on EFP are shown in **bold**. * indicates the best number of training epochs. 'train' and 'test' refer to the (fixed) train or test split, respectively, of the named dataset.

(r), but did for accuracy (acc), F1, precision (P), and recall (R). The specific model parameters are given in Appendix E. All models were trained for 100 and 400 epochs (with early stopping), and only the best of either 100 or 400 epochs is shown for each condition in Table 2 for all models except (B) and (C), the primary models compared, where both are shown. CosSim also was trained on 50 epochs, in keeping with Shen et al. (2023). The models that include demographics (Table 4) were not finished training (i.e., did not reach early stopping) at 400 epochs and so were trained to 800 epochs. Only the mods story part was used in the other models because the experiments indicated slightly worse performance when place and why were included with mods.

5.1 Evaluation Metrics

The metrics used to evaluate empathy prediction are: Spearman correlation (ρ), Pearson correlation (r), accuracy(acc), F1, precision (P), recall (R) and mean-squared error (MSE). When computing correlations, the predicted value was compared to the labeled value, where the labeled value was normalized to range from 0 to 1. For computation of accuracy, F1, precision and recall, the prediction value was binarized, with a threshold of 0.5. Since we are interested in ranking a person's empathy towards stories, we used Spearman correlation values for all significance tests, testing at the 0.05 level of significance.

5.2 Experiments and Results

We examine the effect of different ways of using perspective in empathy prediction models, the use of different input data (i.e., story parts, demographics, data labeled with similarity) and comparison with several LLMs. The results of our empathy prediction experiments are shown in Tables 2–5. We describe the results by preceding each with a research question, followed by discussion of the relevant results from the tables.

Does incorporating an individual's story enhance prediction of how they will rate their empathy toward another story? Table 2 shows empathy prediction performance on the EFP dataset. Training and testing when only storyB is input (A) had the lowest performance among the different models and was significantly lower than PPEP ($p = 0.0$), indicating that some knowledge about the rater can be useful to an empathy prediction model.

How does the story-as-perspective approach compare to the empathic similarity approach explored in an earlier work? The use of cosine similarity (B) did not perform as well as PPEP (C), with a Spearman ρ almost 6 percentage points less, although the difference was not significant ($p = 0.1079$).

Is it more effective to train separate models for the safe and welcome topics? In section (D), performance for the safe and welcome subsets of the test data in section (C) are shown in the top two lines. Training separate classifiers for safe or wel-

	epochs	train	test	model	ρ(%)	r(%)	acc(%)	F1(%)	P(%)	R(%)	MSE
G	100*	EFP+ES	EFP	PPEP	**47.59**	**47.24**	**69.02**	**74.23**	**73.12**	**75.39**	**0.0768**
	100*	EFP+ES	safe	PPEP	44.80	44.08	68.29	74.26	74.26	74.26	0.0740
	100*	EFP+ES	welcome	PPEP	50.30	50.00	69.75	74.21	71.94	76.63	0.0796
H	50*	EFP+ES	EFP$_{newA}$	cosSim	8.06	6.86	65.03	74.49	73.74	75.26	0.1084
	50*	EFP+ES	EFP$_{seenA}$	cosSim	36.08	35.77	63.65	70.02	65.85	74.74	0.0766
	100*	EFP+ES	EFP$_{newA}$	PPEP	12.70	11.44	63.64	74.51	71.03	78.35	0.1166
	100*	EFP+ES	EFP$_{seenA}$	PPEP	55.44	55.28	70.53	74.14	73.88	74.39	0.0656

Table 3: Performance when trained jointly on the EFP dataset and the EmpathicStories (ES) datasets. G: Test on EFP, safe and welcome individually. H: Although the pairs of stories are all unique, a storyA may occur in both train and test. EFP$_{newA}$ are storyAs that did not occur in the training data, and EFP$_{seenA}$ are storyAs that did occur in the training data. Best values when tested on EFP with SBERT embeddings across Tables 2 and 3 are in **bold**.

epochs	train	test	demog	model	ρ(%)	r(%)	acc(%)	F1(%)	P(%)	R(%)	MSE
100*	EFP$_{dem}$	EFP$_{dem}$	none	PPEP	33.03	31.23	63.14	70.52	70.08	70.95	0.1084
800*	EFP$_{dem}$	EFP$_{dem}$	categ.	PPEP	32.83	32.76	62.11	76.63	62.11	100.00	0.09044
800*	EFP$_{dem}$	EFP$_{dem}$	sent.	PPEP	33.12	33.70	62.11	76.637	62.11	100.00	0.09038

Table 4: The effect of including the demographics of PersonA and PersonB is small.

come showed performance on welcome decreased noticeably, supporting the use of a model trained on the combined safe and welcome data in the EFP dataset to learn generalizations across datasets.

Does knowing the background of place and why in addition to the suggested modifications improve empathy prediction? In (E), place and why were included in the stories as described in Section 4. There was no significant difference from using only modifications, with p-values of 0.38 and 0.35 for SEP and PROMPT separators, respectively, vs. not using place and why. Since the modifications tend to be the most controversial part of the stories, e.g., adding surveillance cameras or speed bumps, they may be the largest factor influencing empathy.

Do the results generalize to another embedding model? To examine whether PPEP's better performance holds for another embedding model, we trained a model using the e5 embedding, which was the top performing embedding for similarity on the Hugging Face MTEB leaderboard when we checked in July 2024. As with SBERT embedding, with e5 embedding (F), PPEP performed significantly better ($p = 0.0085$) than the cosSim model. The e5 performance was somewhat lower for both cosSim and PPEP models, so SBERT embeddings were used for our experiments.

Is empathic similarity helpful to empathy prediction? Since empathic story similarity is relevant to empathy, the Empathetic Stories dataset (ES), labeled with story similarity, was added to the PPEP training set (Table 3G). PPEP trained on EFP+ES was significantly better than other models when tested on EFP; the p-values comparing training PPEP on EFP+ES vs. training PPEP, cosSim and PPEP$_{noA}$ on EFP only were $0.0321, 0.0006$ and 0.0, respectively, indicating similarity labels help empathy prediction. It also improved performance on welcome.

Can PPEP effectively use a person's rating for one storyB to better predict their empathy towards other stories? In Section (H), the results from (G) are split into whether the model has been trained on personA (i.e., their storyA and rating for a storyB) to examine the effect of knowing at least one rating by personA. (Recall that subjects rated 1, 3 or 5 stories.) Note the much lower performance when personA is new, EFP$_{newA}$, than when personA has rated at least one storyB, EFP$_{seenA}$. While each story pair in the EFP training set is unique, the better EFP$_{seenA}$ performance may be due to the model learning to predict relative empathy instead of only absolute empathy from story pairs with the same personA.

Does demographic information improve performance? Table 4 indicates that adding demographic features had a small, insignificant effect ($p = 0.48, 0.49$) on empathy prediction performance, consistent with (Hasan et al., 2024).

How does LLM performance compare to models trained on labeled data? LLM performance was examined under two types of conditions: (a) input is either only the desired modifications (mods) or also place and why (full) and (b) use of either 0-shot or 5-shot training. In Table 5, the LLMs performed better when the full text is given, but being given five examples usually did not help. LLM per-

epochs	train	test	model	ρ(%)	r(%)	acc(%)	F1(%)	P(%)	R(%)	MSE
-	-	EFP	Claude 3 Opus mods 0-shot	-7.62	-8.04	45.40	50.00	54.60	46.11	0.1510
-	-	EFP	Claude 3 Opus mods 5-shot	-6.78	-5.79	44.33	37.31	55.96	27.98	0.1836
-	-	EFP	Claude 3 Opus full 0-shot	**8.63**	**8.14**	**57.67**	**68.56**	61.18	**77.98**	**0.1158**
-	-	EFP	Claude 3 Opus full-5shot	2.05	2.47	51.99	55.48	**61.51**	50.52	0.1740
-	-	EFP	GPT-4o mods 0-shot	-7.25	-6.90	43.87	37.76	54.95	28.76	0.2289
-	-	EFP	GPT-4o full 0-shot	-2.43	-1.60	49.08	51.88	58.88	46.37	0.1475
-	-	EFP	GPT-4o full 5-shot	-3.01	-1.34	52.91	60.08	60.31	59.84	0.6010
-	-	EFP	Claude 3.5 Sonnet full 0-shot	-1.72	-1.56	54.75	65.66	59.62	73.06	0.1162
-	-	EFP	Claude 3.5 Sonnet full 5-shot	1.84	2.60	49.39	50.60	59.93	43.78	0.1441

Table 5: Performance of LLM when given modifications only (mods), full story (full), and 0 or 5 examples.

formance was most similar to PPEP trained without storyA (Table 2A), indicating that LLMs did not effectively use storyA, except for Opus full story performance, which was similar to new storyA performance (Table 3H). The low performance may be due to LLMs being trained on web text, with little exposure to perspective-based empathy ratings. Others have noted that LLMs perform much better on academic questions than on ones related to social intelligence (Xu et al., 2024; Sap, 2024)

6 Human Evaluation of Empathy Ratings

We compared PPEP's empathy ratings to human ratings using a new set of participants. Specifically, PPEP's predicted absolute and relative empathy ratings were compared to human ratings using new storyAs by new personAs. The **relative** empathy, $e = e_t - e_1$, is the difference between empathy for a test storyB, e_t, and the empathy for storyB that a subject is predicted to be most empathetic towards, e_1. The **absolute** empathy is e_t only. Relative empathy is a within-subjects measure that we hypothesize may be more robustly predicted than absolute empathy (across-subjects), which is very person dependent, as observed in Section 3.3. We structured our study to allow within-subjects comparison examining the relative empathy.

The collection of new storyAs and empathy ratings for our human evaluation was performed using the approach outlined in Stage 2 of Fig. 2 to collect storyAs and empathy ratings for our EFP dataset. Steps 2d and 2e were identical: in response to a storyA welcome or safe prompt, a subject wrote a story, storyA$_h$, that describes a place where they feel welcome or safe, why they feel welcome or safe there, and suggested modifications to make the place more welcoming or safer; then the place/topic classifier identified the place type/topic cluster that best matches storyA$_h$. For the original rating task in Fig. 2f, 1, 3 or 5 stories were rated. For this task, personA$_h$ rated their empathy for two stories, e_t, a selected storyB, and e_1, the storyB they were predicted to be most empathetic towards; and then the relative empathy, $e = e_t - e_1$, was computed. By rating only stories e_t and e_1, the possible confound of personA's empathy being influenced by other stories they read and rated (see Appendix F) was removed.

The selected test storyB, storyB$_t$, was a selected percentage down the ranked list of stories on the topic, where the ranking is based on PPEP's prediction of the empathy that personA$_h$ would have for each story in the best matching topic cluster. We chose stories at four ranks in the ranked list of storyBs in the topic cluster: 20%, 50%, 75% and 100% (the farthest). Subjects were also asked whether they identified with each story. Each subject rated one pair of "safe" and one pair of "welcome" stories. We used Prolific to recruit 50-55 new participants per percentage rank, or over 200 people not used in earlier studies.

People often feel more empathy for stories by people with whom they identify or share greater similarity (Gutsell and Inzlicht, 2012; Stevens et al., 2021). To remove this potential confound, empathy ratings when the human did not identify with either story were considered separately from when humans identified with both stories. After this filtering, there were an average of 21 ratings for the four no-identify ranks and an average of 10.5 ratings for the four identify ranks, respectively. The mean absolute empathy for the predicted most empathic story should be the same across the ranks if raters use values that are consistent with each other, but instead ranged from 53.2 to 70.4 for identify and from 56.6 to 73.25 for no identify, illustrating variability in mean absolute empathy value, due in part to rater differences noted in Section 3.3.

Given this variability, we focused our analysis on the human-rated empathy of the second storyB *relative* to the first, shown in Fig 5 for the four rank percentages to the target (i.e., the storyB that

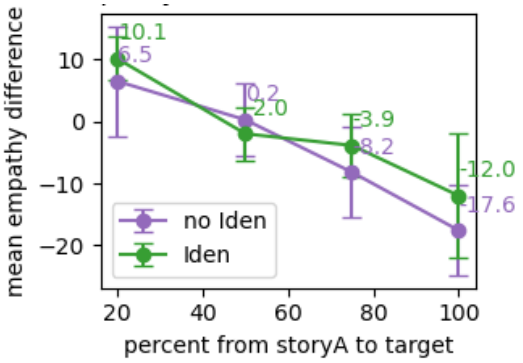

Figure 5: Relative human empathy ratings of stories for different predicted empathy rank, where 100% is the story a rater is predicted to assign the lowest empathy score among a set of stories (target). Bars indicate standard error.

personA is predicted to be least empathic towards). The average relative empathy for stories that a human identifies with (iden) had only about 10 ratings per rank and may be noisy, but tend to be greater than ones not identified with (no iden), as expected. Note that empathy for a story decreases as stories are closer to the the target, i.e., the predicted lowest empathy rated story, indicating that PPEP's empathy predictions generally follow human ranking of empathy for stories. Empathy at 20% increased by 10.1 and 6.5, and decreased at 100% by 12.0 and 17.6, similar to the psychological observation that people are more likely to accept events close to their current beliefs and to view highly surprising events negatively (Filipowicz et al., 2018).

Although the correlations were relatively low when PPEP was tested on new personAs (Table 3H), PPEP's relative within-subject empathy ratings generally follow trend in the human subject ratings. By training on multiple empathy ratings for some of the personAs and on empathic similarity, the results indicate that PPEP may have generally learned relative empathy between stories.

7 Discussion

Experimental results show that having at least one story from a rater (storyA) and their empathy rating for another story (storyB) can greatly improve rater empathy prediction on new storyBs. The almost 40 percentage points empathy prediction improvement with use of one storyA is much larger than the roughly five percentage point improvement with use of demographics noted by others (e.g., Hasan et al., 2024; Guda et al., 2021). PPEP's better performance over cosine similarity indicates that modeling a rater's story as context is more informative than similarity. However, including EmpathicStories to the training data to provide similarity examples also improved performance, indicating that empathetic similarity is complementary and relevant to perspective-based empathy prediction. The poor LLM empathy predictions indicates that LLMs currently do not make good use of user perspective.

Our human evaluation examined the within-subject change in human empathy ratings between the story personA is predicted to be most empathic towards and another story which personA is predicted to be less empathetic towards. The study showed that subjects tended to be less empathetic to stories that are predicted to be farther from the most empathic story, indicating that PPEP's empathy predictions are in general agreement with human empathy ratings. This indicates that PPEP may have learned relative empathy from storyBs paired with the same storyA as well as from the empathic similarity labels in the EmpathicStories data.

8 Summary and Conclusions

Analysis of our EmpathicPerspectives dataset indicated that user perspective is important for empathy prediction when there are diverse perspectives. Experiments with our PPEP model showed that modeling user perspective (storyA) as context had better absolute empathy prediction performance than the baseline models, and adding empathic similarity data to training further improved performance. Our human subject studies showed that while prediction of absolute empathy is difficult, relative empathy can be more robustly predicted by PEPP. Its relative empathy rating of stories generally agreed with human empathy ratings, including the tendency towards greater empathy ratings when personA identifies with storyB.

9 Limitations

This work is an initial investigation of the use of perspective in empathy prediction from different viewpoints. The data was collected in English from participants in the US under a scenario where people were asked to tell stories about two types of places. Thus, the model has been trained on only those two types of places and we have not tested on topics other than safe/less safe and welcome/excluding places. The collection scenario

allowed for asynchronous data collection, but is artificial. A future direction is to predict empathy for another's perspective when two people have diverse perspectives in the wild.

Because clustering was used to group place descriptions, several place types may be grouped into one cluster, such as "park" and "lake", often due to place descriptions being more similar on some aspect to place descriptions in the assigned cluster than to other clusters. While many of the desired modifications for the different types are relevant, e.g., for "park" and "lake", some common modifications include fencing, paths, and security cameras for safety, there will be modifications that aren't relevant. Since the expressed viewpoints of stories in the cluster are different and many of the issues are common to both, all stories in a cluster were considered when matching and ranking stories. Finer granularity of place types would allow for better matching of place type and desired modifications, but would require collection of a larger number of storyBs to insure close matches.

The inclusion of place and why text from the stories along with the proposed modifications of a place slightly lowered performance but not significantly. However, the LLMs performed better when the full text was included (although overall performance was low). Further investigation of a model that better uses the full information was left for future work, and for the current paper we focused on using only modifications.

Our experiments with demographics produced similar findings to earlier works of small performance differences (e.g., Hasan et al., 2024; Guda et al., 2021). Other methods for including demographics could be employed, but we have left further exploration as future work. Another possible interesting area for future work is to include involuntary physiological responses to provide real-time capture of emotional and cognitive states.

10 Ethical Concerns

Upon publication we released the story pairs and empathy values from the EmpathyFromPerspectives dataset, with the names of places anonymized. We will not release the demographic part of the data because it contains Personally Identifying Information, which our review board does not allow to be released.

Our goal in this work is to increase a person's empathy for other people's perspectives, in order to increase people's acceptance of others. However, an ethical concern is that the ability to predict another person's empathy could be misused to increase empathy in undesirable ways, such as for marketing purposes or conspiracy theories. As mentioned by other works in this area, developing ways for detecting misuse of empathy prediction may be needed.

References

Gavin Abercrombie, Valerio Basile, Davide Bernadi, Shiran Dudy, Simona Frenda, Lucy Havens, and Sara Tonelli, editors. 2024. *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*. ELRA and ICCL, Torino, Italia.

Valentin Barriere, João Sedoc, Shabnam Tafreshi, and Salvatore Giorgi. 2023. Findings of WASSA 2023 shared task on empathy, emotion and personality detection in conversation and reactions to news articles. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 511–525, Toronto, Canada. Association for Computational Linguistics.

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan Alqahtani. 2022. WASSA 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 214–227, Dublin, Ireland. Association for Computational Linguistics.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Mina Cikara, Emile Bruneau, Jay J Van Bavel, and Rebecca Saxe. 2014. Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. *Journal of experimental social psychology*, 55:110–125.

Taya R Cohen. 2010. Moral emotions and unethical bargaining: The differential effects of empathy and perspective taking in deterring deceitful negotiation. *Journal of Business Ethics*, 94:569–579.

Alex Filipowicz, Derick Valadao, Britt Anderson, and James Danckert. 2018. Rejecting outliers: Surprising changes do not always improve belief updating. *Decision*, 5(3):165.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. Empathbert: A bert-based framework for demographic-aware empathy prediction. In *Proceedings of the 16th Conference of the*

European Chapter of the Association for Computational Linguistics: Main Volume, pages 3072–3079.

Emma Gueorguieva, Tatiana Lau, Eliana Hadjiandreou, and Desmond Ong. 2023. The language of an empathy-inducing narrative. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Jennifer N Gutsell and Michael Inzlicht. 2012. Intergroup differences in the sharing of emotive states: neural evidence of an empathy gap. *Social cognitive and affective neuroscience*, 7(5):596–603.

Md Rakibul Hasan, Md Zakir Hossain, Tom Gedeon, and Shafin Rahman. 2024. LLM-GEm: Large language model-guided prediction of people's empathy levels towards newspaper article. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2215–2231, St. Julian's, Malta. Association for Computational Linguistics.

Yossi Hasson, Einat Amir, Danit Sobol-Sarag, Maya Tamir, and Eran Halperin. 2022. Using performance art to promote intergroup prosociality by cultivating the belief that empathy is unlimited. *Nature Communications*, 13(1):7786.

Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022. A critical reflection and forward perspective on empathy and natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2139–2158, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tatiana Lau, Scott Carter, Francine Chen, Brandon Huynh, Everlyne Kimani, Matthew L Lee, and Kate A Sieck. 2024. Democratizing design through generative ai. In *Companion Publication of the 2024 ACM Designing Interactive Systems Conference*, pages 239–244.

Yoon Kyung Lee, Inju Lee, Jae Eun Park, Yoonwon Jung, Jiwon Kim, and Sowon Hahn. 2021. A computational approach to measure empathy and theory-of-mind from written texts. *arXiv preprint arXiv:2108.11810*.

Xin Lu, Zhuojun Li, Yanpeng Tong, Yanyan Zhao, and Bing Qin. 2023. HIT-SCIR at WASSA 2023: Empathy and emotion analysis at the utterance-level and the essay-level. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 574–580, Toronto, Canada. Association for Computational Linguistics.

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Luiza A Santos, Jan G Voelkel, Robb Willer, and Jamil Zaki. 2022. Belief in the utility of cross-partisan empathy reduces partisan animosity and facilitates political persuasion. *Psychological Science*, 33(9):1557–1573.

Maarten Sap. 2024. Artificial social intelligence? on the challenges of socially aware and ethically informed large language models. *Winter Bridge on Frontiers of Engineering*, 54.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Jocelyn Shen, Yubin Kim, Mohit Hulse, Wazeer Zulfikar, Sharifa Alghowinem, Cynthia Breazeal, and Hae Park. 2024. EmpathicStories++: A multimodal dataset for empathy towards personal experiences. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4525–4536, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jocelyn Shen, Maarten Sap, Pedro Colon-Hernandez, Hae Park, and Cynthia Breazeal. 2023. Modeling empathic similarity in personal narratives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6237–6252, Singapore. Association for Computational Linguistics.

Vishal Anand Shetty, Shauna Durbin, Meghan S Weyrich, Airín Denise Martínez, Jing Qian, and David L Chin. 2024. A scoping review of empathy recognition in text using natural language processing. *Journal of the American Medical Informatics Association*, 31(3):762–775.

Addepalli Sai Srinivas, Nabarun Barua, and Santanu Pal. 2023. Team_Hawk at WASSA 2023 empathy, emotion, and personality shared task: Multi-tasking multi-encoder based transformers for empathy and emotion prediction in conversations. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 542–547, Toronto, Canada. Association for Computational Linguistics.

Samantha M Stevens, Carl P Jago, Katarzyna Jasko, and Gail D Heyman. 2021. Trustworthiness and ideological similarity (but not ideology) promote em-

pathy. *Personality and Social Psychology Bulletin*, 47(10):1452–1465.

Shabnam Tafreshi, Orphee De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. WASSA 2021 shared task: Predicting empathy and emotion in reaction to news stories. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–104, Online. Association for Computational Linguistics.

Erika Weisz, Desmond C Ong, Ryan W Carlson, and Jamil Zaki. 2021. Building empathy through motivation-based interventions. *Emotion*, 21(5):990.

Ruoxi Xu, Hongyu Lin, Xianpei Han, Le Sun, and Yingfei Sun. 2024. Academically intelligent llms are not necessarily socially intelligent. *Preprint*, arXiv:2403.06591.

Jamil Zaki, Niall Bolger, and Kevin Ochsner. 2008. It takes two: The interpersonal nature of empathic accuracy. *Psychological science*, 19(4):399–404.

A Collection Details of the EmpathyFromPerspectives Dataset

Human subject data was collected for storyAs, storyBs, and empathy ratings for human evaluation of PPEP. The collection of the storyAs and storyBs was part of a larger study. The larger study focused on how people could co-design with one another in their communities to create more inclusive public spaces, which are typically places where one may feel welcomed (i.e., places where you feel mentally at ease) or safe (i.e., physically at ease). In particular, this larger study focused on interventions for increasing a person's empathy for someone else (in particular when their initial empathy for this other person may be low). Our focus on places stems from the current situation in the U.S., where debates about how to design public spaces remain contentious and empathy may be necessary to find solutions and compromises.

Subjects who wrote storyBs were recruited from dscout. The first set and second set of subjects who wrote storyAs and the human evaluation subjects were recruited from Prolific, which was a better fit for the tasks. All participants were located in the US, speak English, and were at least 18 years of age. They were paid at least $17/hour. Each person A participant was asked to write one story about a place where they felt safe and one story about a place where they felt welcome.

After filtering for missing empathy ratings, the number of storyA-storyB pairs is 1609 for welcome and 1625 for safe. There were 365 personA participants without demographics and 792 participants with demographics . Each participant without demographics was asked to write a storyA for welcome and a storyA for safe, in random order, and rated 1, 3 or 5 storyBs for each storyA. PersonA participants with demographics wrote either a safe or welcome storyA, with a smaller number writing both safe and welcome storyAs. For each storyA, a participant rated 1 or 3 storyBs. There were 239 personB participants who contributed one story each for welcome, safe, excluded and less safe. Each story was used for clustering places, but only excluded and less safe were used as storyB's.

The human evaluation data collection was performed independently of the larger study and it used only the subset of questions from the larger study that are relevant to PEPP evaluation. Specifically, we used the answers to prompts 2–4 in Section A.1, which were part of all studies. The exact prompts used in this work are shown below. The studies and prompts were reviewed by an Internal Review Board.

A.1 Prompts for storyA and human evaluation

Prompt 1 is shown for context. In this work, we used the responses to prompts 2-4, corresponding to "place", "why" and "modifications".

The prompts for **safe** are:

1. Share a photo of a local place where you feel **safe**. This should be a place that is seen as publicly accessible to all.
2. Where or what did you take a photo of?
3. Help us and others who aren't from your community understand this place. In more than 2 sentences, what about this place makes you feel **safe**? Tell us a story about why it makes you feel this way.
4. Think about someone who, unlike you, might feel like the safety in this space could be improved. In more than 2 sentences, if you could, how would you modify this space to make it **more safe for them**?

The prompts for **welcome** are:

1. Share a photo of a local place where you feel **welcomed**. This should be a place that is seen as publicly accessible to all.
2. Where or what did you take a photo of?

3. Help us and others who aren't from your community understand this place. In more than 2 sentences, what about this place makes you feel **welcomed for who you are**? Tell us a story about why it makes you feel this way.

4. Think about someone who, unlike you, might feel **excluded** in this space. In more than 2 sentences, if you could, how would you modify this space to make it **more welcoming for them**?

A.2 Prompts for storyB

As in the prompts in A.1, Prompt 1 is shown for context. In this work, we used the responses to prompts 2-4, corresponding to "place", "why" and "modifications".

The prompts for storyB for safe and welcome are the same as the prompts for storyA, except for prompt 1, as shown:

Prompt 1 for **safe** is:

1. Share a photo of a place in your community where you feel **safe**.

Prompt 1 for **welcome** is:

1. Share a photo of a place in your community where you feel **welcomed by others for who you are**.

The prompts for **safety could be improved** are:

1. Share a photo of a place in your community where you feel like **safety could be improved.**

2. Where or what did you take a photo of?

3. Help us and others who aren't from your community understand this place. In more than 2 sentences, what about this place makes you feel like **safety could be improved**? Tell us a story about why it makes you feel this way.

4. In more than 2 sentences, if you could, how would you modify this space to make it **more safe**?

The prompts for **excluded** are:

1. Share a photo of a place in your community where you feel **excluded by others for who you are**.

2. Where or what did you take a photo of?

3. Help us and others who aren't from your community understand this place. In more than 2 sentences, what about this place makes you feel **excluded because of who you are**? Tell us a story about why it makes you feel this way.

4. In more than 2 sentences, if you could, how would you modify this space to make it **less excluding** and **more welcoming**?

A.3 Clustering of places into types

To identify a small number of types of places in the collected place descriptions by personBs so that matching of positive stories (storyA) about a place could be matched to negative stories (storyB) about a similar type of place, the place description in storyBs were clustered. The place descriptions for welcome and excluded were clustered together. Separately, the place descriptions for safe and less safe were clustered together. By clustering place descriptions from positive and negative stories together, places are grouped irrespective of whether the story is positive or negative. Each cluster can be roughly characterized as a type of place, and the place description from a positive storyA can be matched against the negative storyBs from the matching place type cluster.

To cluster the place descriptions, each description was embedded using SBERT and then the set of embeddings were clustered using agglomerative clustering with Ward linkage (https://scikit-learn.org/0.15/modules/generated/sklearn.cluster.AgglomerativeClustering.html). The stopping criteria was set to 14 clusters for welcome/exclude and for safe/less safe after manual examination for cluster purity and redundancy. All clusters contained multiple negative stories, except for the 'library' cluster in which libraries are always described as welcoming. If a storyA was matched to the library cluster, the participant was matched to a random cluster. The place-type classifier, which matched a storyA to a cluster was a k-NN classifier where k=3 (https://scikit-learn.org/1.5/modules/generated/sklearn.neighbors.KNeighborsClassifier.html). A manual characterization of the cluster place types is listed in the next section. We did not remove "outlier" stories since they might be stories that participants would be least empathetic towards.

A.4 Manual characterization of cluster place types

Manual examination of the clusters showed that a rough categorization could be assigned to each cluster. The place types roughly corresponded to:

welcome/excluded:

- local shops, mostly grocery
- church
- sports spaces and clubhouses, e.g., golf, pools, baseball
- my neighborhood
- coffee shops and restaurants
- elementary and middle schools
- library
- park, may be near a body of water
- misc
- gym, figness center
- nearby park
- bars
- multi-unit housing
- misc

safe/less safe:

- neighborhood
- streets and crosswalks
- local businesses
- parks, boardwalks and ponds and rivers
- trails, parks and nature
- stores and fast food shops
- local park
- community spaces
- train and subway stations
- playgrounds, school playgrounds, rec areas
- gym
- parking lot
- bus depots, highways

A.5 Empathy ratings

For each pair of opposing viewpoints, the empathy of PersonA towards storyB is collected. Several recent psychology studies have used a single question to assess state (vs. trait) empathy (e.g., Zaki et al. (2008); Cikara et al. (2014); Weisz et al. (2021); Santos et al. (2022)). We employed this approach to assess a participant's state empathy immediately after they read each story. The specific question asked was: "To what extent did you feel empathy, compassion, or sympathy toward the person while reading the narrative?".

The data was collected on two different presentation platforms; one had a 7-point Likert scale setup, while the other allowed for a 100-pt sliding scale setup. We do not believe that the differences should affect analysis. For the first platform, users rated on a Likert scale from 1 (no empathy) to 7 (very empathetic), while for the second platform, users used a slider to select a rating from 1 (no empathy) to 100 (very empathetic). The values from the first collection were normalized to range from 1 to 100 for combining the two datasets for analysis.

The ratings of the EFP data set were split 75/5/20 for train, validation and test. The variants of EFP dataset, e.g., EFP$_B$ and EFP$_{newA}$, were derived the original 75/5/20 splits, e.g., the test data split is from the same original test split in all cases. For the EFP+ES dataset, each combined split was composed of the corresponding EFP and ES splits.

B Demographic Data Collected

In our modeling, six demographic characteristics were collected from the participants in the second data collection on the second platform using the slider scale. They include the five types used in the WASSA (Barriere et al., 2023) shared tasks, plus "people I live with".

The demographic distribution for a subset of storyA participants for which we collected demographics from Prolific are shown below.

- Gender: Man 354; Woman 378; Prefer to self-describe 5; Genderqueer or non-binary 12; Agender 2; Prefer not to answer 3
- Age: mean 37; median 39; min 18; max 76; std dev 13
- Ethnicity: White 510; American Indian or Alaskan Native,White 11; Middle Eastern or North African 1; Black or African American

103; Asian 65; Prefer not to say 9; American Indian or Alaskan Native 5; Asian ,Native Hawaiian or Other Pacific Islander 1; Prefer to self-describe 10; Native Hawaiian or Other Pacific Islander 1; Black or African American,Middle Eastern or North African 1; Asian ,White 11; Black or African American ,White 9; Middle Eastern or North African 2; Native Hawaiian or Other Pacific Islander ,White 1; Black or African American ,White ,Prefer to self-describe 1

- Education: Bachelor's degree 300; High School or equivalent 70; Master's degree in the Arts and Sciences 77; Some college but no degree 155; Associate's degree 76; Professional Master's degree 29; Ph.D. 17; Some High School 3; Professional degree (Ex; JD, LLM, SJD, MD, DO, DDS, DVM) 10; Trade School 4; Other Doctoral degree (e.g., EdD, DDiv, DrPH, DBA, etc.) 5; Prefer not to answer 2

- People I live with: With partner/spouse + child 205; With partner/spouse 207; With parent(s) 101; Solo 125; With roommate(s) 53; With child under 18 29; Other 28

- Income: Less than $25,000 96; $25,000 to $49,999 132; $50,000 to $74,999 148; $75,000 to $99,999 119; $100,000 to $124,999 94; $125,000 to $149,999 65; Over $150,000 79; I prefer not to respond 15

The demographic distribution for storyB participants collected from dscout is shown below.

- Gender: woman 74; man 41; non_binary 1

- Age: mean 38.9; median 38; min 24; max 74; std dev 10.9

- Ethnicity: Asian 10; Hispanic or Latinx 3; White 78; Black or African American 13; White, Native Hawaiian or other Pacific Islander, I , Asian, Hispanic or Latinx, Black or African American 1; Black or African American, White 2; Middle Eastern or North African, White 1; Prefer not to say 1; Hispanic or Latinx, White 3; Asian, White 1; White, Asian 2; American Indian or Alaska Native 1

- Education: Some high school 1; High school graduate 4; Some college 21; College graduate 55; Post graduate coursework 35

- People I live with: With partner/spouse 31; With child under 18 8; With partner/spouse + child 46; With roommate(s) 7; Solo 19; With parent(s) 2; Other 2

- Income: Less than $25,000 5; $25,000 to $49,999 15; $75,000 to $99,999 21; $50,000 to $74,999 23; $100,000 to $124,999 24; $125,000 to $149,999 7; Over $150,000 20; I prefer not to respond 1

C EFP Dataset Prompts when Place and Why are Used

When the place and why story parts are included in the EFP dataset with prompts, prompts were used to introduce place, why and modifications. The prompts depend on whether the text was collected for safe, less safe, welcome or excluded. The exact prompts are:

Prompts for safe "I am writing about this place:", "I feel safe here because" and "Some ways this place could be modified to be more safe are:"

Prompts for less safe "I am writing about this place:", "I feel less safe here because" and "Some ways this place could be modified to be more safe are:"

Prompts for welcome "I am writing about this place:", "I feel welcome here because" and "Some ways this place could be modified to be more welcoming are:"

Prompts for exclude "I am writing about this place:", "I feel excluded here because" and "Some ways this place could be modified to be more welcoming are:"

D LLM prompt

This is the prompt based used in our LLM experiments. It was used by Shen et al. (2024) for prompting GPT-4:

- 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. You will receive two stories, one from person A and the other from person B. Please predict, on a scale from 0 to 1, where 0 is not empathetic and 1 is extremely empathetic, how much person A would empathize with story B. Return just the number, no other text.

E PPEP Model Settings

The models in this paper are based on the bi-encoder empathic similarity model at https://github.com/mitmedialab/empathic-stories. For PPEP, the cosine-similarity is replaced with a 3-layer MLP with an input dimension of 768*2, and layer dimensions of 768, 384, 192, followed by a sigmoid. The target only (storyB) classifier is a 2-layer MLP with input dimension of 768 and layer dimensions of 384, 192, followed by a sigmoid.

The classifier used for studies when a second embedding is used for additional data, e.g., demographics or place and why, is a 4-layer MLP with layer dimensions of 1536, 768, 384, 192 followed by a sigmoid. For experiments with e5, which has an embedding size of 1024, the MLP input dimension was 1024*2 and the layer dimensions were 1024, 512, 192, followed by a sigmoid.

We used the hyperparameter values used in the empathic similarity model, except we reduced the number of workers to 16 and increased the number of epochs to 100 and 400 epochs for the models and datasets tested. Additional epochs tested for some conditions are specified in Section 5. As in the empathic similarity model, SBERT was not frozen so the embeddings could better model empathy.

The Adam optimizer with a learning rate of 1e-6 was used. Early stopping on the validation data was used when training and a checkpoint of the current best model during training was saved. All models were run on a 2-GPU Puget with NVIDIA RTX 6000 ADA 48 GB GPUs. All training conditions took approximately an hour or less per 100 epochs to train. MSE loss between the predicted empathy and human empathy ratings, both ranging from 0 to 1, was used. Validation loss was used to select the best-performing model.

F Effect of multiple stories on empathy

We conducted a variation of the human evaluation described in Section 6 to examine the influence of multiple stories on empathy. Instead of presenting personA with only two stories, either three stories or five stories were presented. Our results indicate that presenting more stories can increase empathy for the storyB that the rater is predicted to be least empathetic towards.

We examined three cases, where in addition to presenting the story that personA is predicted to be most empathetic towards, storyB$_1$, the other stories presented are:

1. storyBs at rank 50% and 100%
2. three randomly selected storyBs and then the storyB at rank 100%
3. storyBs at rank 20%, 50%, 75%, 100%.

storyB (%) ranks presented	ave. relative empathy for rank 100%
1, 50, 100	-22.0
1, three random, 100	-13.2
1, 20, 50, 75, 100	-8.0

Table 6: Effect on empathy of reading multiple storyBs . The left column is the ranks of the sequence of storyBs presented. The right column is the empathy for storyB at rank 100%, relative to empathy for storyB at rank 1%. The rank of a storyB was described in Section 6. Greater relative empathy is better.

The average relative empathy was computed as $e_{100} - e_1$ where e_{100} is the empathy for the storyB that personA is predicted to be least empathic towards (storyB at rank 100%) and e_1 is the empathy for the storyB that personA is predicted to be most empathic towards (storyB at rank 1%). The results shown in Table 6 indicate that people are, on average, more empathetic towards storyB$_{rank100\%}$ when they have read more stories (i.e., five stories vs three stories). The results also indicate that presenting stories in rank order sampled though the range of ranks is more effective than presenting a random selection of the same number of stories.