Can Machines Resonate with Humans? Evaluating the Emotional and Empathic Comprehension of LMs

Muhammad Arslan Manzoor[1*] **Yuxia Wang**[1*] **Minghan Wang**[2] **Preslav Nakov**[1]
[1]MBZUAI, Abu Dhabi, UAE [2]University of Monash, Melbourne, Australia
{muhammad.arslan, yuxia.wang, preslav.nakov}@mbzuai.ac.ae

Abstract

Empathy plays a pivotal role in fostering prosocial behavior, often triggered by the sharing of personal experiences through narratives. However, modeling empathy using NLP approaches remains challenging due to its deep interconnection with human interaction dynamics. Previous approaches, which involve fine-tuning language models (LMs) on human-annotated empathic datasets, have had limited success. In our pursuit of improving empathy understanding in LMs, we propose several strategies, including contrastive learning with masked LMs and supervised fine-tuning with large language models. While these methods show improvements over previous methods, the overall results remain unsatisfactory. To better understand this trend, we performed an analysis which reveals a low agreement among annotators. This lack of consensus hinders training and highlights the subjective nature of the task. We also explore the cultural impact on annotations. To study this, we meticulously collected story pairs in Urdu language and find that subjectivity in interpreting empathy among annotators appears to be independent of cultural background. Our systematic exploration of LMs' understanding of empathy reveals substantial opportunities for further investigation in both task formulation and modeling.

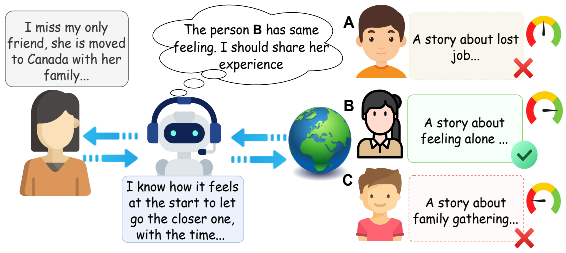

Figure 1: An *ideal* interaction between users and a system. A chatbot can resonate with a human, and a search engine can retrieve stories of similar experience.

1 Introduction

With Large Language Models (LLMs) demonstrating impressive capabilities in generating naturally sounding answers over a broad range of human inquiries, more individuals turn to seek solutions and emotional comfort by interacting with LLM-supported chatbots (OpenAI, 2023; Chang et al., 2024). They express thoughts, feelings and share their experiences, expecting deep understanding and sympathetic responses from a chatbot that can resonate with them, as shown in Figure 1 (Berridge et al., 2023). This requires LLMs to first fully understand the event, the emotions, and the empathy in the narratives, and then to respond appropriately (Lin et al., 2023).

Shen et al. (2023) proposed the task of measuring empathic similarity which assesses the similarity between two narratives describing personal experiences across four aspects: event, emotion, moral, and empathy, using a numerical score ranging from 1 to 4 (see more in Section 3.1). However, both fine-tuned LMs (BERT, RoBERTa) and few-shot prompted LLMs achieved low correlation with human annotations. Error analysis on Shen et al. (2023) shows that models can recognize dissimilar story pairs with scores in range of 1-2, but struggle to distinguish fine-grained differences between similar story pairs in range of 2.5-4 (see Section 3). This suggests that the nuanced patterns and relationships in similar pairs are not captured by current methods (Wang et al., 2022a).

We hypothesize that the LMs used in these methods were primarily trained for semantic understanding tasks, rather than for emotion and empathy. They may capture the similarity between events, but they struggle to understand the complex social and emotional signals in human narratives (Reimers and Gurevych, 2019), this is further discussed in following sections. Moreover, naïve LLM prompts may not fully empower LLMs'

[*]Equal contribution.

Findings of the Association for Computational Linguistics: EACL 2024, pages 14683–14701
November 12-16, 2024 ©2024 Association for Computational Linguistics

reasoning ability to make correct judgements. To this end, we explore multiple strategies to improve the accuracy of empathic similarity, including contrastive learning of LMs, LLM reasoning and fine-tuning with and without Chain-of-Thought (CoT) (Wei et al., 2022). We achieve around 5–10% improvement in the Pearson and Spearman correlations. However, accuracy appears to be capped and cannot be improved beyond a certain value.

This motivated us to speculate that gold labels might be problematic due to the subjective nature of judging emotion and empathy similarity. To investigate this, we randomly sampled ten story pairs from EmpathicStories and asked eight annotators with different backgrounds to annotate the similarity in terms of event, emotion, moral and empathy (defined in Section 3.1). Overall, low human agreement for all aspects was observed, especially moral, followed by empathy, emotion; the highest agreement was for the event. Interestingly, annotators from the same country/culture, particularly who are friends, had much higher agreement than others. We further collected a new Urdu dataset to explore the cultural impact on empathy similarity labeling, which revealed similar findings.

In sum, this work presents three contributions:

- We explore various approaches to improve empathic similarity estimation, including the use of contrastive loss in LMs, reasoning and fine-tuning of LLMs.
- We analyze the upper bound of correlation between model predictions and human gold labels. By gathering collective human opinions and measuring inter-annotator agreement, we reveal high disagreements in empathic labeling, highlighting the subjective nature of judging empathic similarity.
- We collect a new Urdu dataset to investigate the impact of language and culture on empathic labeling. Our analysis shows that subjectivity in the labeling is independent of the cultural background of annotators.[1]

2 Related Work

Emotion and Empathy in LMs: Many prior work focus on tasks of recognizing (mixed) emotions, scoring intensity, inferring and explaining person's emotional reactions (Liu et al., 2024b; Chen et al., 2024). In general domains, recent study illustrates that current LLMs are far better than human in generation, but fall short of understanding (West et al., 2023). However, for empathy-demanding scenarios, understanding and resonating with support-seekers are even more important than outputting opinions (i.e., generation) (Buecker et al., 2021). In this work, we aim to examine language models capability in understanding and identifying nuanced difference of emotion and empathy, based on the task of *empathic similarity* (Shen et al., 2023). We aim to improve the estimation accuracy by forcing models to discern and learn the underlying reasons for similarities between pairs of stories, via (1) enhancing LM-based sentence embeddings, and (2) empowering LLM reasoning capability.

Sentence Embedding Enhancement: Many previous methods improve semantic embeddings by adjusting loss functions in training (Khan et al., 2022). SimCSE leverages contrastive loss (Gao et al., 2021), and ESimCSE applies momentum contrast strategy to increase the number of negative pairs involved in the loss calculation, showing notable improvements across multiple semantic textual similarity benchmarks (Wu et al., 2021). Hubert et al. (2024) improved the knowledge graph embeddings by integrating semantic awareness into traditional loss functions like hinge loss. Huang et al. (2024) use a multi-task hybrid loss, incorporating both skeletal and semantic embedding into loss functions for micro-gesture classification. Therefore, we investigate a variety of loss functions such as ContrastiveLoss, CosineSimilarityLoss, CoSENTLoss and AnglELoss, based on LMs including MPNet (Song et al., 2020), RoBERTa (base, large) (Liu et al., 2019), DeBERTa (small, large) (He et al., 2020) to improve sentence embedding in representing empathic features.

LLM Reasoning and Fine-tuning: Naive LLM prompts may fail to fully leverage LLMs' reasoning capabilities, leading to poor accuracy on EmpathicStories dataset (Shen et al., 2023). Recent advancements in zero- and few-shot prompting techniques can boost performance such as Chain-of-Thought prompting (CoT) (Wei et al., 2022), Least-to-Most prompting (Zhou et al., 2022), and search-based approaches like Tree-of-Thought (ToT) (Yao et al., 2024). We prompt LLMs by CoT and also fine-tune LLMs with reasoning trajectories, encouraging LLMs to think

[1] Urdu dataset with individual similarity scores and modeling code are available at https://github.com/yuxiaw/Empathic-Similarity

over before making the final decision.

3 Dataset and Error Analysis

In this section, we first introduce the dataset used in this work and then we perform an error analysis for some baseline methods.

3.1 Task and Dataset

In this work, we focus on the task of measuring the *empathic similarity* between two narratives by a numerical score in the range of 1 to 4, with 1 representing totally dissimilar and 4 indicating extremely similar. Empathic similarity assesses how much the narrators of a pair of stories would empathize with one another, in which main event, emotion, and moral similarities of two stories are core features influencing the empathic similarity.

Event highlights the similarity of main events in two experiences, as people empathize more with experiences that are similar to their own. **Emotional** reaction refers to how people emotionally respond to the experience. Individuals may have different feelings to the same experience, e.g., alone vs. happy for *staying at home on weekend*. **Moral** in this context emphasizes a higher level meaning abstracted by readers from a story, i.e., takeaway.

We used the EmpathicStories dataset, which was created by Shen et al. (2023). It consists of 1,500 unique stories and 2,000 story pairs, split into 1,500 pairs for training, 100 for development, and 400 for testing. Appendix Figure E shows the sample from EmpathicStories that consists of pair of stories and similarity scores on four labels. Table 1 shows the distribution of train, development, and test sets across different similarity ranges. Each story has two versions, full and summary, in which the annotators assign labels based on the summary from four perspectives including event, emotion, moral, and overall empathy. Figure 2 shows the correlation between the four similarity scores. We can see that the moral similarity has the highest correlation with empathy, followed by event and emotion.

Evaluation Measurements: Pearson correlation (r) and Spearman rank correlation (ρ) are used to measure the performance of systems for predicting emotional similarity. This measures the linear correlation between the model outputs, the human annotations and the degree of monotonicity under ranking, respectively. Mean square error (MSE) assesses the models' ability to get close to the gold standard. In a discrete setting, following Shen et al. (2023), both the predicted scores and the gold labels are binned into two classes thresholding by 2.5 — label = 1 if score > 2.5; otherwise, it is 0. Then accuracy, precision, recall and macro-F1 are used.

Splits→ **Label**↓	**Train** (1500)	**Dev.** (100)	**Test.** (400)	**SBERT** #Dev	#Test	**BART** #Dev	#Test
1	85	2	15	0	0	0	1
1.5	125	16	29	0	0	2	3
2	310	24	75	1	1	11	19
2.5	288	20	76	4	17	15	33
3	344	21	103	13	49	11	55
3.5	262	15	84	12	70	10	57
4	86	2	18	2	18	2	15

Table 1: EmpathicStories dataset distribution and the number of incorrect predictions on the development / test set by fine-tuned SBERT and BART baselines.

3.2 Error Analysis

Baseline: We reproduced the results of Shen et al. (2023) by fine-tuning SBERT (Reimers and Gurevych, 2019) and BART-base (Lewis et al., 2019) as shown in Table 9, given that prompting *davinci-text-003* and GPT-3.5-turbo had lower accuracy. We obtained a Spearman correlation ρ similar to Shen et al. (2023), but with a lower F1.

Results and Analysis: We regard a predicted score as a severely incorrect estimation when the absolute difference between the predicted score and the gold labels is greater than 1.0,[2] the number of incorrect predictions over the development and the test sets is shown in Table 1.

We found that the model excels at identifying dissimilar story pairs, but struggles with similar pairs with nuanced differences (not immediately obvious). Specifically, the model exhibits lower error rates in the score range of 1–2, and higher error rates in the range of 2.5–4. Particularly for 4, both models have a 100% error rate on the test set. We could attribute this to exposure bias during training. However, the number of training examples scoring between 2.5 and 4 is actually much larger than the number of dissimilar examples. This suggests that it is difficult for models to discern the subtle differences between similar story pairs, underlining a critical area for model improvement.

Hypothesis: We hypothesize that SBERT and BART are primarily pre-trained and fine-tuned to

[2]The cosine similarity scores based on SBERT and BART is in the range of [0,1], we scale them up by multiplying 4 to match the annotation range of 1-4.

learn semantic features, in which emotional and empathetic features are under-represented. Despite the importance of understanding the semantics of the stories, this does not necessarily account for capturing the deep, empathetic connections between narratives, especially when these connections are subtler and more implicit than straightforward lexical or thematic commonalities. The failure to identify the connections between similar narratives in terms of empathy leads to poor performance on similar story pairs ranging from 2.5 to 4, tending to be recognized as dissimilar with a score between 1-2. To alleviate these issues and to enhance the model performance on the empathetic similarity task, we propose a variety of strategies in Section 4.

4 Enhancing Empathy Representation

In order to improve the model's ability to recognize the connections between narratives that are semantically diverse yet empathically similar, we apply constrative losses in LM fine-tuning, and incorporate reasoning in LLM inference and fine-tuning.

Contrastive Learning of LMs: We use a contrastive loss to enhance representation by bringing the embeddings of similar examples closer, while pushing such of dissimilar samples apart in the hidden space (Gao et al., 2021). This involves first grouping them into positive and negative pairs, and then applying a contrastive loss to learn patterns. In this task, we use annotated similarity scores to determine positive and negative pairs, setting a threshold of 2.5 as the boundary following the binning approach in Shen et al. (2023).

We explored various contrastive losses, including the ContrastiveLoss (Hadsell et al., 2006), AnglELoss (Li and Li, 2024) and CoSENTLoss[3]. We also re-implemented the approach using cosine similarity loss as a baseline. We used masked language models (MLMs) including MPNet (Song et al., 2020), RoBERTa (base, large) (Liu et al., 2019) as text encoders leveraging their bidirectional encoding capabilities and computational efficiency.

LLM Reasoning and Fine-tuning: We explored the potential of LLMs for the task of empathic similarity in both inference and fine-tuning, with two prompting strategies:

- **Score-Only**: With annotation guidelines, the LLM is prompted to predict only a score without any additional content.

[3]www.sbert.net

Train Label↓	Model	Loss	r	ρ	F1-macro
`EmpathicStories Baselines` (Summary)					
Empathy	SBERT	MSE	0.359	0.352	0.713
Empathy	BART	MSE	0.342	0.344	0.701
Empathy	GPT-3.5-turbo	NA/5-shot	0.278	0.281	0.696
`Contrastive Loss of LMs` (Summary)					
Empathy	RoBerta-base	Cosine	**0.404**	0.388	0.608
Empathy	Multi-qa-MPNet	Cosine	0.400	**0.395**	0.647
Event	RoBerta-base	Contrastive	0.318	0.309	0.634
Event	Multi-qa-MPNet	Contrastive	0.370	0.364	0.624
Emotion	RoBerta-base	Cosine	0.377	0.371	**0.650**
Emotion	RoBerta-large	Cosine	0.393	0.388	0.611
Moral	RoBerta-base	AnglELoss	0.326	0.323	0.649
Moral	Multi-qa-MPNet	Cosine	0.387	0.374	0.604

Table 2: Test set results for LM fine-tuned over annotations of **event**, **emotion**, **moral**, and overall **empathy** score and tested on **empathy** vs. baseline results obtained from Shen et al. (2023).

- **Chain-of-Thought (CoT)**: The LLM is instructed to first provide an explanation and then to predict the scores.

For the CoT explanation, even though the human-annotated explanations were provided in EmpathicStories, they were collected based on the story summary by GPT-3.5-Turbo, which can be biased and inaccurate. Moreover, the distribution of human-written text is distinct from the distribution of LLM-generated reasons, which makes the reasons provided by the humans less useful for lowering the perplexity in LLM fine-tuning (Liu et al., 2024a). To address this, we prompted an LLM to generate explanations guided by gold labels, leveraging LLM capability of causal reasoning. That is, to explain why two narratives have the similarity of *[scores placeholder]* from the aspects of event, emotion, moral, and empathy.

Gold Label-Guided Explanation: Based on the pair of full stories with the ground truth similarity scores, we asked the LLM to analyze the story pair from dimensions such as theme, emotional content, characters, narrative structures, and overall empathy. The generated analysis served as the reasoning content in supervised fine-tuning (SFT). We explored both full parameter and PEFT (Hu et al., 2021) fine-tuning. We used Llama3-70B-instruct to generate explanations by prompts in Figure 4 of Appendix B.1.

5 Experiments

We experimented with discriminative LMs and generative LLMs with zero-shot and fine-tuning.

Testbed→	Summary							Full						
Test Label↓	r	ρ	MSE↓	Acc	Prec	Recall	F1-macro	r	ρ	MSE↓	Acc	Prec	Recall	F1-macro
`OpenAI-text-embedding-3-large`														
Empathy	0.336	0.329	1.510	0.505	0.633	0.517	0.376	0.362	0.363	1.440	0.507	0.624	0.519	0.384
Event	**0.485**	**0.465**	**0.620**	**0.780**	**0.738**	**0.542**	**0.522**	**0.488**	**0.469**	**0.590**	**0.782**	**0.737**	**0.551**	**0.538**
Emotion	0.392	0.388	1.310	0.550	0.582	0.510	0.392	0.393	0.386	1.260	0.568	0.685	0.529	0.421
Moral	0.366	0.356	1.210	0.620	**0.692**	**0.525**	**0.437**	0.395	0.403	1.140	0.618	0.651	0.524	0.440
`Llama-3-8B`														
Empathy	**0.325**	0.322	**0.620**	**0.595**	0.596	0.593	**0.591**	0.324	0.308	**0.520**	**0.590**	**0.595**	**0.592**	0.588
Event	**0.315**	**0.306**	1.240	0.525	0.574	0.601	0.509	**0.342**	**0.312**	0.900	**0.660**	0.617	**0.659**	0.611
Emotion	0.270	0.265	0.780	0.555	0.564	0.563	0.554	**0.317**	0.294	0.600	0.595	0.590	0.588	**0.588**
Moral	0.319	**0.323**	0.830	**0.600**	**0.622**	**0.623**	**0.600**	0.331	**0.329**	0.640	0.650	**0.636**	0.638	**0.637**
`Llama-3-70B`														
Empathy	0.405	0.403	**0.620**	**0.635**	0.661	0.630	**0.614**	0.304	0.295	**0.970**	**0.545**	0.628	0.534	**0.436**
Event	**0.427**	**0.431**	**1.280**	**0.480**	0.623	**0.639**	0.479	**0.337**	**0.357**	**1.980**	**0.305**	**0.625**	**0.547**	**0.287**
Emotion	0.387	0.374	0.770	0.565	0.605	0.585	0.550	0.305	0.312	1.180	0.495	0.617	0.532	0.407
Moral	0.412	0.415	0.840	0.585	**0.663**	**0.636**	**0.579**	0.305	0.320	1.320	0.455	**0.659**	0.545	0.391
`GPT-4o`														
Empathy	0.442	0.441	0.620	0.652	0.660	0.655	0.650	0.350	0.373	0.650	0.640	0.640	0.640	0.640
Event	**0.492**	**0.488**	**0.560**	**0.703**	0.660	**0.716**	0.659	**0.414**	**0.424**	**0.710**	**0.605**	**0.615**	**0.660**	**0.579**
Emotion	0.466	0.452	0.580	0.647	0.645	0.641	0.641	0.360	0.371	0.660	0.620	0.622	0.622	0.620
Moral	0.476	0.481	**0.560**	0.698	**0.685**	**0.687**	**0.686**	0.396	**0.424**	**0.630**	**0.685**	**0.689**	**0.697**	**0.683**

Table 3: Zero-shot results of discriminative and generative models on the test set using summary and full story over four types of gold similarity scores: **empathy**, **event**, **emotion** and **moral**. Cosine similarity is scaled 1-4 by $\times 4$ for discriminative models. Classification gold labels are binned by score > 2.5.

5.1 Discriminative LMs

We refer to discriminative models as smaller pre-trained LMs like RoBERTa, and sentence embedding models such as SBERT. In both zero-shot and fine-tuning settings, we calculated cosine similarity using the embeddings of two stories generated by either pre-trained or fine-tuned sentence encoders.

Each story pair in EmpathicStories has four similarity scores: event, emotion, moral, and overall empathy. Shen et al. (2023) only focused on fine-tuning and evaluating over empathy scores, under-investigating the other three aspects and their impact on the empathy similarity estimations. In contrast, we experimented across all four labels, providing a comprehensive understanding of the model performance.

5.1.1 Zero-Shot Evaluation

We performed these experiments on both the full stories and on the summaries. We calculated the cosine similarity based on the embeddings of a range of pre-trained text encoders including open-sourced `MiniLM` (Wang et al., 2020) and `MPNet` (Song et al., 2020), as well as close-sourced `OpenAI-text-embedding-3-large` (Neelakantan et al., 2022), and then we evaluated their predictions against the gold labels from four perspectives. Table 10 shows that (*i*) the close-sourced model outperformed open-sourced models on both full and summary across all aspects in correlations, and (*ii*) the cosine similarity scores for all models had the highest correlation with the event labels, indicating that semantics was the dominant feature captured in story representations. They were significantly notable than emotional and moral signals.

5.1.2 Fine-Tuning with Contrastive Loss

We examined the effectiveness of contrastive losses presented in Section 4 based on three LMs. We fine-tuned them on the four types of gold labels separately and evaluated them using the overall empathy score as ground-truth based on pairs of story summary.[4] This enables us to assess the impact of each attribute on modeling empathic similarity.

Table 2 showcases the best two results over multiple LM×Loss settings given the training label. We found that training on empathy similarity yields the highest correlation, followed by emotion, moral, and event across all settings. This suggests that empathy is more closely related to emotion and moral than to events. In terms of loss functions, CosineSimilarity loss consistently outperformed the rest, except for event as training labels, where Contrastive loss was the best. This highlights the robustness

[4] The full story always exceeds the maximum input length of these three LMs, resulting in poor performance.

of CosineSimilarity loss across different scores. Among the three LMs, `Roberta-base` was always in the top-2 list, demonstrating robustness over all settings. `Roberta-large` and `Multi-qa-MPNet` outperformed in correlations when training with event, emotion, and moral.

These results suggest that, while some models and loss functions excel in specific metrics or training labels, `Roberta-base` with CosineSimilarity loss appears to be the most robust and effective combination for capturing the nuances of empathy. See the full results in Table 11.

5.2 Generative LLMs

Unlike discriminative models to predict or measure a similarity score, LLMs explicitly generate scores in natural language. We first evaluated LLMs in zero-shot setting to assess the inherent ability of LLMs to comprehend emotion and empathy, and then we fine-tuned the same LLMs to learn and to reason with the empathy in the stories.

5.2.1 Zero-Shot Evaluation

We optimized the prompt used in Shen et al. (2023) by specifying the differences between the four concepts, and by highlighting this in the system prompt instead of the user input, guiding the model to predict empathic similarity scores. Observing that the order of the stories affected the predicted scores, we generated two scores for each pair of stories by swapping the positions of stories A and B and then averaging the two scores as the final prediction.

Table 3 shows two major findings: (*i*) Zero-shot LLMs have clear advantages over cosine similarity based on sentence embeddings. r=0.442 for GPT-4o vs. r=0.336 for `OpenAI-text-embedding` on story summary. The larger the model, the higher the correlations with the gold empathy scores. This suggests that LLMs benefit from large-scale pre-training that empowers the distinguishability of nuanced empathic differences in narratives. (*ii*) Over all models and over the four types of scores, using the summaries is superior to using the full stories, which aligns with the fact that the gold labels are annotated based on the summary.

5.2.2 Is SFT Helpful?

To further investigate the potential of LLMs in this task, we performed supervised fine-tuning (SFT) with LLaMA-3-8B using the story summaries and the corresponding empathic similarity scores from the training set. In our experiments, we explored two different prompting strategies as introduced in Section 4. For the **Score-Only** strategy, we implemented both full-parameter and LoRA (Hu et al., 2021) SFT. For the **CoT** strategy, we only performed LoRA SFT due to the memory pressure of the longer sequence length. For all strategies, we tuned the model for two epochs to avoid overfitting.

Table 4, shows that, compared to the zero-shot setting, SFT does not enhance the performance of LLaMA-3-8B regardless of the strategy. In some cases, fine-tuning even worsens the results. Moreover, the full-parameter SFT setting yields the poorest results, which is counter-intuitive. This motivated us to analyze why LLMs could not improve the performance of empathic similarity to r>0.5 at least, and why SFT even worsened the results.

5.3 Understanding the Bottleneck

To understand what LLMs learned during fine-tuning and why they struggle with this task, we analyzed the predictions made by the fully fine-tuned model. We selected this model because it is specifically adapted to the task of empathic similarity estimation, minimizing interference from its pre-trained knowledge.

Ideally, if the model learned how to predict the empathic similarity based on the input story pairs, its probability over four similarity classes [1,2,3,4] should be close to [1, 0, 0, 0] when the gold label of the input pair is 1, and to [0, 1, 0, 0] when the gold label of the input pair is 2, and so forth. To observe the real predictive probability over each similarity class, we first grouped the training story pairs X into four sets by their gold empathy labels y (continuous scores are rounded to the nearest integer), denoted as $X_{y_{gold}=i}$ with $i \in [1, 2, 3, 4]$. Next, we computed the probability for each pair by applying the *softmax* function to the logits of $\{1, 2, 3, 4\}$ of the first predicted token, and then we averaged the probabilities across all pairs within each class, as shown in Table 5.

Regardless of the similarity class, the predicted probability kept the same as the empirical distribution of training data over four classes, i.e., $P(Y)$ = (0.140, 0.399, 0.404, 0.057), which is calculated by counting the percentage of story pairs falling into each class. That is, after fine-tuning, the model learned nothing about how to map the input $x \rightarrow y$, but merely the distribution of $P(Y)$. During inference, the model could not estimate the corresponding score conditioned on the input story pair, but

Testbed→ SFT Strategy	Development Set r	ρ	MSE	Acc	Prec	Recall	F1-macro	Test Set r	ρ	MSE	Acc	Prec	Recall	F1-macro
LoRA-SFT-Score	**0.470**	**0.476**	**0.497**	**0.790**	**0.759**	**0.786**	**0.768**	**0.307**	**0.320**	**0.643**	0.600	0.602	0.603	0.599
LoRA-SFT-COT	0.342	0.325	0.618	0.690	0.698	0.709	0.687	0.289	0.299	0.651	**0.613**	**0.619**	**0.609**	**0.603**
Full-SFT-Score	0.209	0.189	0.627	0.630	0.640	0.632	0.625	0.028	0.039	0.773	0.507	0.506	0.506	0.504

Table 4: LLaMA-3-8B SFT by three paradigms. LoRA-SFT-Score: score-only strategy tuned with LoRA. LoRA-SFT-COT: CoT strategy with LoRA. Full-SFT-Score tunes the model with full parameters on empathy scores.

Gold→	y=1	y=2	y=3	y=4
$P(Y\|X_{y_{gold}=1})$	0.142	0.392	0.409	0.058
$P(Y\|X_{y_{gold}=2})$	0.140	0.386	0.418	0.056
$P(Y\|X_{y_{gold}=3})$	0.129	0.381	0.434	0.055
$P(Y\|X_{y_{gold}=4})$	0.132	0.378	0.428	0.062
$P(Y)$	0.140	0.399	0.404	0.057

Table 5: Probability estimated by fine-tuned LLaMA-3-8B over four groups binned by gold label y of story pairs. $X_{y_{gold}=1}$ is the group of examples whose gold empathy similarity is 1. $P(Y)$ represents empirical distribution of training labels. All values are computed on the training set.

sampled a similarity class based on the distribution of $P(Y)$ whatever the input pair was, leading to randomness on the development and the test sets (see confusion matrix of training set in Figure 5).

Based on these observations and previous findings in semantic textual similarity that models would be confused when training with ambiguous and subjective labels (Wang et al., 2022b, 2023, 2024), we speculate that the annotated empathic similarity scores are substantially subjective, i.e., that labels of high human disagreements hinder the model to learn distinguishable patterns. This could also explain why LMs and LLMs could not improve the performance by a large margin despite our extensive efforts. The struggle of the models to overcome the bottleneck can be attributed to the complexity and the variability in the human interpretations of these abstract concepts. This underscores the need for a critical analysis of the dataset's quality and the inherent subjectivity of its labels.

6 Collective Human Opinions

To analyze the subjectivity in empathic similarity labeling, we first collected eight annotations for each pair of English stories. We then explored the impact of language and culture on empathy labeling by collecting and annotating Urdu story pairs.

6.1 English Story Pair Annotation

We sampled 10 story pairs from the development sets of EmpathicStories, and we invited eight annotators to assign labels in three settings.

Annotation Setup: Shen et al. (2023) employed Amazon MTurk workers to annotate similarity scores based on the story summaries considering the heavy cognition load of understanding the full story. Upon performing qualitative analysis of both full stories and their summaries, we observed that summaries generated by GPT-3.5-turbo would dismiss many narratives presenting emotional changes (*from depressed, to sad, finally turn to happy*), inner monologue (*I feel alone after Mary left*) and details about other roles, only keeping the major events and the tone of the full stories. Yet, these details are very important to identify subtle feelings and they can affect empathy. Model predictions based on the full stories also significantly differ from such obtained from the summaries as shown in Section 5. To this end, we collected annotations under three settings:

- **Continuous score on the summary:** The annotators assigned similarity scores for event, emotion, moral, and empathy as values ranging from 1 to 4, based on the story summaries.
- **Continuous score on the full story:** Same as above, but based on the full stories.
- **Discrete class-label on full story:** Considering that continuous scores provide larger labeling space, which may exacerbate subjectivity and inner-annotator disagreement, the annotators are asked to rate event and emotion similarity using class labels: very similar (V), moderately similar (M), and not similar (N). Here, we did not rate moral and empathy given that the concepts were too abstract to perceive and rate.

Based on the annotation results, we aim to measure and to analyze the subjectivity of the human opinions in different annotation setups, i.e., how the representation of the stories and the annotation scale influence inter-annotator agreement (IAA).

IAA Metric↓	S_eve	S_emo	S_mor	S_emp	F_eve	F_emo	F_mor	F_emp	Comb_eve	Comb_emo	S_eve	S_emo	S_mor	S_emp	F_eve	F_emo	F_mor	F_emp	Comb_eve	Comb_emo
	All annotators - 8										English / Chinese Speakers (friends- 2)									
Pearson	0.421	0.078	0.189	0.191	0.442	0.098	0.110	-0.015	0.176	0.141	0.735	0.322	0.192	0.148	**0.848**	0.311	-0.086	-0.241	-0.302	-0.153
Spearman	0.395	0.063	0.187	0.190	0.438	0.088	0.138	0.004	0.158	0.129	0.701	0.252	0.143	0.090	**0.868**	0.313	-0.109	-0.108	-0.437	-0.183
KA	0.349	0.059	0.168	0.114	0.398	0.092	0.105	0.027	0.095	0.107	0.401	0.086	0.174	-0.152	0.761	0.233	-0.065	-0.349	-0.295	-0.128
Cohen Kappa	0.124	-0.015	0.067	0.047	0.157	-0.024	0.030	0.026	0.059	0.049	0.143	-0.045	0.024	-0.013	0.359	-0.053	0.067	-0.039	-0.364	-0.061
	English / Urdu Speakers (colleagues - 4)										English / Urdu speakers (sisters - 2)									
Pearson	0.266	0.037	0.089	0.093	0.322	-0.017	0.234	-0.039	0.137	0.071	**0.836**	0.728	0.698	0.771	0.683	0.697	0.278	0.504	0.625	0.238
Spearman	0.235	0.028	0.101	0.078	0.295	-0.013	0.194	-0.075	0.133	0.065	**0.754**	0.686	0.695	0.787	0.523	0.577	0.301	0.455	0.577	0.221
KA	0.235	0.051	0.050	0.026	0.311	-0.008	0.187	0.057	0.098	0.030	0.724	0.476	0.709	0.773	0.659	0.560	0.273	0.486	0.317	0.195
Cohen Kappa	0.113	0.053	-0.030	0.007	0.114	-0.040	-0.006	-0.070	0.015	-0.016	**0.429**	0.167	0.375	0.412	**0.412**	-0.045	0.143	0.403	0.138	-0.014

Table 6: Agreement scores for four groups. The summary and the full story are abbreviated as S and F, Comb = S+F. Event, emotion, moral, and empathy are shortened as eve, emo, mor, and emp. KA refers to Krippendorff's Alpha.

Pearson and Spearman correlation, Krippendorff's Alpha, and Cohen Kappa (convert to discrete) are used to measure IAA. We first calculated the agreement between each pair of annotators and then we averaged them.

Annotator Background: We had eight annotators, aged between 20 and 35 years, from diverse cultural and ethnic backgrounds, including two native Chinese speakers and six native Urdu speakers, with a balanced gender distribution. All annotators are proficient in English, some majored in psychology with bachelor's degree and some are PhD and postdoc students in NLP. Training was performed before the formal annotations, which instructed the annotators how to rate the similarity score and highlighted the differences between event, emotion, moral, and empathy (guidelines in Figure 6).

Results and Analysis: Table 6 shows the inter-annotator agreements. We can see that the agreement for event is the highest, followed by emotion, empathy, and moral. This suggests that it is easier for the annotators to reach agreement on more concrete aspects such as event and emotion compared to more abstract concepts of moral and empathy.

The closer the relationship between two annotators, the higher the agreement for more subtle aspects. Two Chinese friends (one male and one female) exhibited the highest correlation on judging event similarity based on the full story, but lower correlation on all other settings than the two sisters from Pakistan. They even achieved r/ρ >0.7 for moral and empathy when the average was less than 0.2 based on story summary. Four Urdu speaker colleagues showed extremely low agreement with each other, with correlations around 0.3 for event and r/ρ<0.1 for other aspects.

This indicates that individuals who have closer interpersonal relationships with each other reach better agreement in interpretations of the stories, particular for subtle and abstract concepts. Moreover, varying the agreement scores across different groups suggests that individual background and culture significantly influence how stories are perceived and annotated. This subjectivity poses challenges for models to learn and to replicate human judgements, especially for moral and empathy. The agreement of collective human opinions in similarity score annotations also presents an upper bound for a model.

Comparing the three settings, we can see lower agreement when labeling with three classes than with continuous scores. Moreover, using full stories yields higher correlations compared to using summaries. This guides the setting of future annotations for empathic similarity, providing annotators with full stories and annotating concrete aspects like event and emotion using continuous labels.

6.2 Urdu Dataset Construction

Given that none of the annotators are native English speakers nor have grown up in a Western culture, bias might be introduced in labeling English stories pairs. We thus collected Urdu stories and asked Urdu native speakers to annotate them. This aims to eliminate potential biases, and most importantly to investigate the impact of languages and culture on the empathy similarity labeling. Urdu is widely spoken in many South Asian countries, including Pakistan, India, Bangladesh, and Nepal. Roman Urdu/Hindi, which uses the English alphabet to write Urdu or Hindi, is commonly used in these regions for communication, discussions, and sharing feelings on media platforms. Consequently, we collected stories written in Roman Urdu to capture the authentic expression and nuances of this widely used form of communication.

Synthetic Story Pairs Generation: After manually checking the quality of the stories generated by GPT-4o (e.g., whether they are emotion-rich stories matching local culture), we collected 300 story pairs by instructing GPT-4o to generate story pairs with varying similarities given a diverse range

IAA Metric↓	event	emotion	empathy	overall
	All annotators (4)			GPT-4o
Pearson	0.308	0.381	0.422	0.425
Spearman	0.307	0.362	0.392	0.403
KA	0.214	0.316	0.392	0.392
Cohen Kappa	0.133	0.169	0.190	0.075

Table 7: Agreement scores for Urdu annotators on the Roman Urdu dataset. On the right, the empathy scores of 4 annotators are averaged correspondingly and compared to the overall GPT score.

Gold Label↓	r	ρ	MSE↓	Acc	Prec	Recall	F1
	OpenAI-text-embedding-3-large						
Empathy	0.486	0.527	1.610	0.296	0.609	0.562	0.289
Event	0.539	0.548	1.040	0.399	0.651	0.564	0.372
Emotion	0.493	0.529	1.600	0.292	0.607	0.562	0.286
	Llama-3-8B						
Empathy	0.410	0.375	0.640	0.759	0.772	0.629	0.637
Event	0.445	0.371	0.450	0.832	0.733	0.651	0.675
Emotion	0.428	0.412	0.480	0.808	0.680	0.615	0.632
	GPT-4o						
Empathy	0.715	0.725	0.330	0.797	0.762	0.757	0.759
Event	0.772	0.762	0.310	0.842	0.760	0.848	0.786
Emotion	0.774	0.774	0.310	0.818	0.732	0.807	0.754

Table 8: **Performance on the Urdu dataset** over three types of gold similarity scores: event, emotion, and empathy using cosine similarity (v_a, v_b) with OpenAI embedding and zero-shot LLMs.

of domains and topics (see prompts in Figure 7). Each pair consisted of theme, content of two stories, an overall similarity score *s* in [1,4], and a brief explaination of the score.

Human Annotations: Four Urdu native speakers were trained to annotate similarity scores for event, emotion and empathy, ranging from 1 to 4 with increments of 0.5; we excluded moral due to high ambiguity. The IAA between the four annotators is shown in Table 7. The last column *overall* represents the agreement between the averaged empathy scores of the four raters with the overall similarity score *s* provided by GPT-4o during story generation. Interestingly, empathy achieved the highest IAA, followed by emotion and then event, which differs with the findings in English data: event>emotion>empathy.

Agreement scores can be significantly influenced by shared interpersonal traits among annotators. Cultural factors also play a role, with closer interpersonal relationships leading to higher agreement scores. As demonstrated in the results, when annotators are more closely related, resonate with each other and have the same perception about the world, their agreement score is higher which may enhance the LMs capacity to learn cultural specific empathy. This suggests that, enhancing the machines' understanding of empathy, stories must be collected and annotated within targeted countries or regions where there is a deep understanding of local demographic norms. The correlations between the averaged empathy score and GPT-4o similarity are closer to those between annotators, suggesting the human-like judgement of GPT-4o.

LLM Estimations: Given the superior zero-shot performance obtained by OpenAI-text-embedding-3-large, LLaMA-3-8B, and GPT-4o on English dataset, we applied them on the Urdu story pairs in Table 8. GPT-4o outperforms the other two models by a sizable margin. All of them have the highest scores on event, and then emotion and empathy, which is similar to the results for English, while the correlations of all aspects are much higher than for English across the three models.

7 Conclusion and Future Work

We proposed a variety of strategies to enhance model performance on empathic similarity, including the use of contrastive losses on LMs, LLM reasoning and fine-tuning. Our experiments demonstrated 5–10% improvements compared to baselines. However, our analysis revealed the subjective nature of empathic similarity between narratives. Collective human annotations on both English and Urdu story pairs illustrated the low human agreement in empathy labeling, highlighting the inherent challenges of this task and indicating an upper bound that a model can reach. The annotation of Urdu stories further exposed the cultural impact on empathic labeling.

We find that many factors impact the inner annotator agreement in empathic similarity labeling, including the native language of annotators, the similarity of their background, experience, the closeness between them, and training process. Considering these factors that affect the correlations between annotators in a unified framework would be an interesting direction to explore in future work, especially the studies across multiple languages. Furthermore, we will explore task reformulations to reduce the variability and subjectivity, and we will try more robust approaches to effectively model empathy in narratives.

Limitation

The overall task inherently carries subjectivity. The gold labels are not standardized and vary based on individual backgrounds, demographics, perspectives, experiences, and surroundings. Cultural differences also play a significant role. Emotions are complex and varied, and each emotion can be expressed in multiple ways and at different intensities. Thus, gold labels are subjective, and stories themselves contain many nuances that may not be fully empathized by the annotators. Each emotion has a broad spectrum of intensity, rather than a binary state like happy or sad.

Ethical Statement

We adhere to ethical standards in data collection, annotation, and analysis. All human annotators were well informed about the task and provided their consent. We ensured diverse representation among the annotators to account for various cultural and demographic perspectives, aiming to minimize biases in empathic and emotional labeling. The datasets used, including those in Urdu, were collected and processed with respect for cultural sensitivity. We acknowledge the subjective nature of empathy and emotion analysis and have taken steps to highlight and to address these challenges in our study. Our work is committed to advancing understanding while maintaining ethical integrity and respect for the individuals whose data and annotations were used.

References

Clara Berridge, Yuanjin Zhou, Julie M. Robillard, and Jeffrey Kaye. 2023. Companion robots to mitigate loneliness among older adults: Perceptions of benefit and possible deception. *Frontiers in Psychology*, 14.

Susanne Buecker, Marcus Mund, Sandy Chwastek, Melina Sostmann, and Maike Luhmann. 2021. Is loneliness in emerging adults increasing over time? a preregistered cross-temporal meta-analysis and systematic review. *Psychological bulletin*, 147 8:787–805.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. Tombench: Benchmarking theory of mind in large language models. *CoRR*, abs/2402.15052.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 1735–1742. IEEE Computer Society.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Junqin Huang, Zhongjie Hu, Zihao Jing, Mengya Gao, and Yichao Wu. 2024. Piccolo2: General text embedding with multi-task hybrid loss training. *arXiv preprint arXiv:2405.06932*.

Nicolas Hubert, Pierre Monnin, Armelle Brun, and Davy Monticolo. 2024. Treat different negatives differently: Enriching loss functions with domain and range constraints for link prediction. In *European Semantic Web Conference*, pages 22–40. Springer.

Adnan Khan, Sarah AlBarri, and Muhammad Arslan Manzoor. 2022. Contrastive self-supervised learning: a survey on different architectures. In *2022 2nd International Conference on Artificial Intelligence (ICAI)*, pages 1–6. IEEE.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdel rahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*.

Xianming Li and Jing Li. 2024. Angle-optimized text embeddings.

Tzu-Mi Lin, Jung-Ying Chang, and Lung-Hao Lee. 2023. NCUEE-NLP at WASSA 2023 shared task 1: Empathy and emotion prediction using sentiment-enhanced RoBERTa transformers. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 548–552, Toronto, Canada. Association for Computational Linguistics.

Qin Liu, Fei Wang, Nan Xu, Tianyi Yan, Tao Meng, and Muhao Chen. 2024a. Monotonic paraphrasing improves generalization of language model prompting. *CoRR*, abs/2403.16038.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Zhiwei Liu, Kailai Yang, Tianlin Zhang, Qianqian Xie, Zeping Yu, and Sophia Ananiadou. 2024b. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. *CoRR*, abs/2401.08508.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. Text and code embeddings by contrastive pre-training.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

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.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Yuxia Wang, Daniel Beck, Timothy Baldwin, and Karin Verspoor. 2022a. Uncertainty estimation and reduction of pre-trained models for text regression. *Transactions of the Association for Computational Linguistics*, 10:680–696.

Yuxia Wang, Shimin Tao, Ning Xie, Hao Yang, Timothy Baldwin, and Karin Verspoor. 2023. Collective human opinions in semantic textual similarity. *Transactions of the Association for Computational Linguistics*, 11:997–1013.

Yuxia Wang, Minghan Wang, Yimeng Chen, Shimin Tao, Jiaxin Guo, Chang Su, Min Zhang, and Hao Yang. 2022b. Capture human disagreement distributions by calibrated networks for natural language inference. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1524–1535, Dublin, Ireland. Association for Computational Linguistics.

Yuxia Wang, Minghan Wang, and Preslav Nakov. 2024. Rethinking STS and NLI in large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 965–982, St. Julian's, Malta. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian R. Fisher, Abhilasha Ravichander, Khyathi Raghavi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. 2023. The generative ai paradox: "what it can create, it may not understand". *ArXiv*, abs/2311.00059.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2021. Esimcse: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. In *International Conference on Computational Linguistics*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Appendix

A Error Analysis

Figure 2 shows the correlation between the four similarity scores. Table 9 exhibits the results authors mentioned in (Shen et al., 2023) and our reproduction results on EmpathicStories. Figure 3 show the prediction count with SBERT and BART on test and development set.

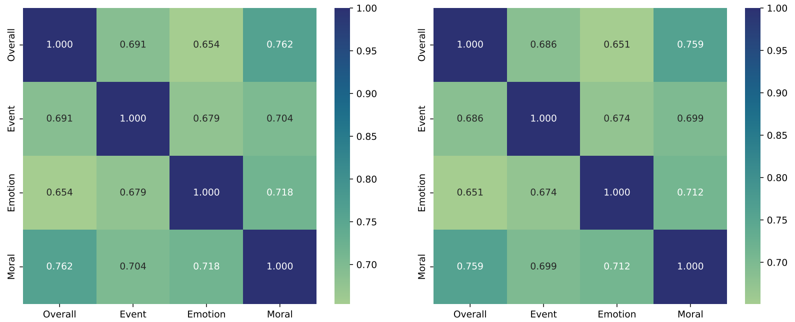

Figure 2: Pearson and Spearsman correlation between overall empathic similarity, event, emotion and moral similarity. Moral similarity has the highest correlation with the empathic, followed by event and emotion.

Model	Macro-F1↑	MSE↓	ρ ↑
SBERT (our rerun)	0.560	0.224	0.317
SBERT (theirs)	0.712	–	0.352
BART (our rerun)	0.579	0.240	0.389
BART (theirs)	0.706	–	0.344

Table 9: Reproduced results vs. results presented in Shen et al. (2023).

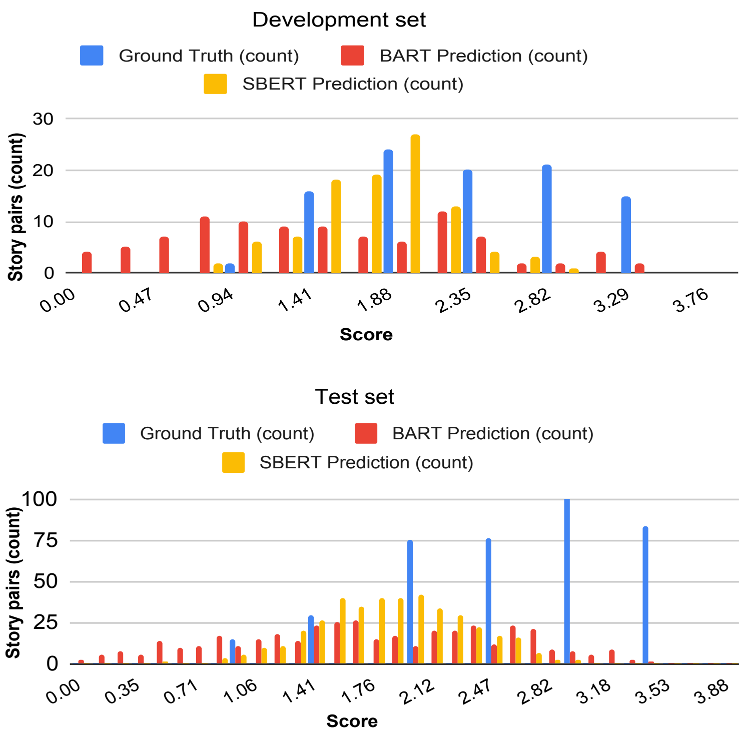

Figure 3: Dev/Test set empathic similarity distribution: predictions of BART vs. SBERT vs. ground truth.

B Experiments

B.1 Gold-label Guided Explanation Generation

We used Llama3-70B-instruct with the prompts detailed in Figure 4 to analyze the story pairs across various dimensions, generating analysis that was subsequently used as reasoning content for supervised fine-tuning (SFT).

System_prompt:
You are given two stories and an empathic similarity score. The score is in a range from 1 to 4, representing low to high empathic similarity.
Your task is to give an explanation on why the score is assigned with that value, why the two stories have high or low empathic similarity.
Here are some recommanded perspectives that can be addressed in your analysis:

Thematic Similarities:
- Identify Common Themes: Describe any common themes that both stories share. Consider elements such as setting, plot, character motivations, and moral dilemmas.
- Relevance of Themes to Empathy: Explain how these themes might resonate on an empathetic level with readers or characters within the stories.

Emotional Content:
- Describe Emotional Tone: Analyze the emotional tone of each story. How do the feelings conveyed in each story align?
- Impact of Emotional Similarity on Empathy: Discuss how similar emotional experiences in the stories might contribute to the empathic similarity score.

Character Analysis:
- Compare Main Characters: Consider the challenges, desires, and emotional journeys of the main characters in each story.
- Role of Character Experiences in Empathy: Reflect on how the characters' experiences might foster empathy between them or with the audience.

Narrative Structure:
- Examine Storytelling Techniques: Look at how each story is told. Are there similarities in narrative perspective, pacing, or style?
- Influence of Narrative on Empathy: Suggest how these narrative elements could affect the empathic connection between the stories.

Overall Empathy Evaluation:
- Synthesize Insights: Combine your observations from the above categories to justify the assigned empathic similarity score.
- Score Justification: Provide a detailed explanation of why the given score is appropriate based on your analysis.

By going through this structured analysis, you can provide a comprehensive **explanation of why two stories received their specific empathic similarity score**. This exercise not only aids in understanding the nuances of empathy in storytelling but also enhances the model's ability to perform nuanced literary analysis.

User_prompt:
Narrative A: {story_1}

Narrative B: {story_2}

Empathic Similarity: {score}

Explanation:

Figure 4: Gold-label Guided Explanation Generation Prompt using Llama3-70B-instruct.

B.2 Discriminative Models Results

Testbed→	Development Set							Test Set						
Test Label↓	r	ρ	MSE↓	Acc	Prec	Recall	F1-macro	r	ρ	MSE↓	Acc	Prec	Recall	F1-macro
							Multi-qa-MiniLM (Summary)							
Empathy	0.204	0.201	0.0815	0.63	0.591	0.543	0.513	0.220	0.213	0.095	0.550	0.590	0.557	0.508
Event	**0.331**	**0.263**	**0.061**	**0.84**	**0.711**	**0.670**	**0.686**	**0.311**	**0.294**	**0.056**	**0.740**	**0.623**	**0.609**	**0.614**
Emotion	0.271	0.231	0.0792	0.72	0.643	0.578	0.578	0.226	0.226	0.088	0.560	0.554	0.534	0.500
Moral	0.169	0.133	0.075	0.760	0.600	0.568	0.574	0.216	0.214	0.084	0.610	0.575	0.550	0.533
							Multi-qa-MiniLM (Full)							
Empathy	0.226	0.236	0.078	0.600	0.520	0.509	0.466	0.151	0.13	0.092	0.520	0.543	0.527	0.475
Event	**0.337**	**0.285**	**0.060**	**0.790**	**0.593**	**0.569**	**0.577**	**0.268**	**0.239**	**0.058**	**0.715**	**0.584**	**0.574**	**0.578**
Emotion	0.206	0.208	0.082	0.690	0.572	0.537	0.524	0.226	0.226	0.088	0.560	0.554	0.534	0.500
Moral	0.169	0.139	0.075	0.760	0.600	0.568	0.574	0.216	0.214	0.084	0.610	0.575	0.550	0.533
							Multi-qa-MPNet-base (Summary)							
Empathy	0.170	0.140	**0.066**	**0.440**	**0.554**	**0.528**	**0.405**	0.263	0.238	**0.058**	**0.540**	0.581	0.530	0.447
Event	**0.298**	**0.280**	0.133	0.270	0.515	0.513	0.269	**0.368**	**0.347**	0.102	0.315	0.575	0.538	0.303
Emotion	0.224	0.204	0.086	0.370	0.544	0.525	0.355	0.318	0.297	0.067	0.515	**0.632**	**0.549**	**0.445**
Moral	0.206	0.198	0.093	0.310	0.539	0.528	0.306	0.274	0.256	0.075	0.445	0.571	0.527	0.392
							Multi-qa-MPNet-base (Full)							
Empathy	0.145	0.127	**0.076**	**0.440**	0.6380	0.543	**0.384**	0.299	0.288	**0.065**	**0.530**	**0.645**	**0.518**	0.389
Event	**0.291**	**0.303**	0.158	0.230	0.524	0.512	0.223	0.349	**0.346**	0.132	0.250	0.500	0.500	0.218
Emotion	0.218	0.228	0.100	0.350	0.589	0.532	0.320	**0.350**	0.333	0.078	0.465	0.537	0.504	**0.347**
Moral	0.231	0.227	0.108	0.290	0.614	**0.550**	0.277	0.301	0.282	0.089	0.420	0.625	0.516	0.330
							OpenAI-text-embedding-3-large (Summary)							
Empathy	0.335	0.315	1.280	0.630	0.813	0.513	0.411	0.336	0.329	1.510	0.505	0.633	0.517	0.376
Event	**0.437**	**0.411**	**0.600**	**0.820**	0.414	0.494	0.451	**0.485**	**0.465**	**0.620**	**0.780**	**0.738**	**0.542**	**0.522**
Emotion	0.394	0.350	1.130	0.720	0.859	0.517	0.451	0.392	0.388	1.310	0.550	0.582	0.510	0.392
Moral	0.359	0.309	0.960	0.800	**0.899**	**0.524**	**0.489**	0.366	0.356	1.210	0.620	0.692	0.525	0.437
							OpenAI-text-embedding-3-large (Full)							
Empathy	0.303	0.309	1.400	0.620	0.561	0.505	0.406	0.362	0.363	1.440	0.507	0.624	0.519	0.384
Event	**0.385**	**0.409**	**0.680**	**0.810**	0.413	0.488	0.448	**0.488**	**0.469**	**0.590**	**0.782**	**0.737**	**0.551**	**0.538**
Emotion	0.345	0.361	1.260	0.710	0.607	0.510	0.446	0.393	0.386	1.260	0.568	0.685	0.529	0.421
Moral	0.337	0.325	1.050	0.790	**0.648**	**0.517**	**0.484**	0.395	0.403	1.140	0.618	0.651	0.524	0.440

Table 10: **Cosine similarity(v_a, v_b) across three sentence embedding models.** Results on dev and test sets over four type of gold similarity scores: **empathy**, **event**, **emotion** and **moral** based on full story and summary. Cosine similarity is normalized to the scale 1-4 by $\times 4$. Classification gold labels are binned by score > 2.5.

Model	Loss	r	ρ	Acc	Prec	Recall	F1-macro	r	ρ	Acc	Prec	Recall	F1-macro
					Empathy						Event		
RoBerta-base	**AnglELoss**	0.355	0.351	0.640	0.642	0.641	0.640	0.254	0.248	0.593	0.608	0.597	0.583
RoBerta-base	**Cosine**	**0.404**	0.389	0.620	0.629	0.616	0.609	0.337	0.315	0.565	0.622	0.573	0.520
RoBerta-base	**Contrastive**	0.331	0.319	0.637	0.639	0.636	0.634	0.318	0.309	**0.635**	0.637	**0.636**	**0.635**
RoBerta-large	**ConSENT**	0.291	0.298	0.603	0.611	0.605	0.599	0.281	0.299	0.590	0.602	0.594	0.583
RoBerta-large	**Cosine**	0.373	0.346	0.608	0.612	0.604	0.599	**0.353**	0.350	0.593	**0.653**	0.600	0.557
Multi-qa-MPNet	**Cosine**	0.400	**0.396**	**0.647**	**0.648**	**0.648**	**0.647**	0.343	0.346	0.573	0.643	0.581	0.525
Multi-qa-MPNet	**Contrastive**	0.358	0.347	0.615	0.615	0.614	0.613	0.371	**0.365**	0.625	0.629	0.627	0.624
					Emotion						Moral		
RoBerta-base	**AnglE**	0.292	0.279	0.580	0.581	0.581	0.580	0.326	0.324	**0.650**	0.650	0.650	**0.650**
RoBerta-base	**Cosine**	0.378	0.371	**0.652**	**0.654**	**0.651**	**0.650**	0.339	0.334	0.645	0.645	0.645	0.645
RoBerta-base	**Contrastive**	0.335	0.321	0.618	0.626	0.614	0.607	0.346	0.331	0.642	0.645	0.644	0.642
RoBerta-large	**CoSENT**	0.303	0.287	0.585	0.585	0.585	0.585	0.287	0.311	**0.650**	**0.651**	**0.651**	0.650
RoBerta-large	**Cosine**	**0.394**	**0.388**	0.620	0.626	0.617	0.611	**0.387**	0.373	0.640	0.640	0.640	0.640
Multi-qa-MPNet	**Cosine**	0.367	0.359	0.613	0.612	0.611	0.611	**0.387**	**0.374**	0.608	0.615	0.610	0.605
Multi-qa-MPNet	**Contrastive**	0.339	0.325	0.610	0.610	0.609	0.608	0.368	0.362	0.608	0.616	0.610	0.604

Table 11: Performance of LMs fine-tuned based on annotations of **event**, **emotion** and **moral** & overall **empathy** similarity scores respectively. Note that we consistently evaluate against empathy similarity score (test gold labels) though training with labels from four aspects.

B.3 Generative LLMs Results

Zero-shot Generative LLMs. Table 3 demonstrates the results of zero-shot generative LLMs including Llama-3-8B, Llama-3-70B and GPT-4o on test set, Table 12 provides the additional results over development set.

Testbed→	Development Set							Test Set						
Test Label↓	r	ρ	MSE	Acc	Prec	Recall	F1-macro	r	ρ	MSE	Acc	Prec	Recall	F1-macro
`Llama-3-8B` (Summary)														
Empathy	**0.494**	**0.500**	**0.550**	**0.740**	**0.751**	**0.765**	**0.738**	**0.325**	0.322	**0.620**	0.595	0.596	0.593	0.591
Event	0.468	0.465	1.340	0.570	0.597	0.671	0.531	0.315	0.306	1.240	0.525	0.574	0.601	0.509
Emotion	0.459	0.466	0.780	0.630	0.648	0.678	0.619	0.270	0.265	0.780	0.555	0.564	0.563	0.554
Moral	0.411	0.400	0.900	0.590	0.614	0.671	0.563	0.319	**0.323**	0.830	**0.600**	**0.622**	**0.623**	**0.600**
`Llama-3-8B` (Full)														
Empathy	**0.326**	**0.351**	**0.540**	0.670	**0.646**	**0.637**	**0.640**	0.324	0.308	**0.520**	0.590	0.595	0.592	0.588
Event	0.208	0.212	1.220	**0.680**	0.577	0.620	0.573	**0.342**	0.312	0.900	**0.660**	0.617	**0.659**	0.611
Emotion	0.235	0.233	0.780	0.660	0.600	0.608	0.603	0.317	0.294	0.600	0.595	0.590	0.588	0.588
Moral	0.165	0.147	0.860	0.600	0.502	0.502	0.493	0.331	**0.329**	0.640	0.650	**0.636**	0.638	**0.637**
`Llama-3-70B` (Summary)														
Empathy	**0.450**	**0.459**	**0.610**	**0.700**	**0.705**	**0.717**	**0.697**	0.405	0.403	**0.620**	**0.635**	0.661	0.630	**0.614**
Event	0.420	0.403	1.400	0.570	0.583	0.647	0.524	**0.427**	**0.431**	1.280	0.480	0.623	**0.639**	0.479
Emotion	0.416	0.435	0.840	0.630	0.639	0.668	0.616	0.387	0.374	0.770	0.565	0.605	0.585	0.550
Moral	0.369	0.369	0.960	0.610	0.622	0.683	0.579	0.412	0.415	0.840	0.585	**0.663**	0.636	0.579
`Llama-3-70B` (Full)														
Empathy	0.264	0.284	**1.110**	**0.460**	**0.621**	**0.554**	**0.421**	0.304	0.295	**0.970**	**0.545**	0.628	0.534	**0.436**
Event	**0.380**	**0.402**	2.220	0.270	0.549	0.537	0.268	**0.337**	**0.357**	1.980	0.305	0.625	**0.547**	0.287
Emotion	0.344	0.333	1.380	0.390	0.617	0.560	0.372	0.305	0.312	1.180	0.495	0.617	0.532	0.407
Moral	0.288	0.262	1.580	0.330	0.619	0.576	0.325	0.305	0.320	1.320	0.455	**0.659**	0.545	0.391
`GPT-4o` (Summary)														
Empathy	0.482	0.479	0.620	0.740	**0.726**	0.709	**0.714**	0.442	0.441	0.620	0.652	0.660	0.655	0.650
Event	0.474	0.457	0.660	0.710	0.605	0.662	0.608	**0.492**	**0.488**	**0.560**	**0.703**	0.660	**0.716**	0.659
Emotion	**0.518**	**0.514**	**0.590**	**0.750**	0.700	**0.712**	0.705	0.466	0.452	0.580	0.647	0.645	0.641	0.641
Moral	0.423	0.416	0.650	0.710	0.621	0.659	0.628	0.476	0.481	**0.560**	0.698	**0.685**	0.687	**0.686**
`GPT-4o` (Full)														
Empathy	0.351	0.351	**0.630**	**0.710**	**0.697**	**0.705**	**0.699**	0.350	0.373	0.650	0.640	0.640	0.640	0.640
Event	0.385	0.367	0.850	0.660	0.616	0.702	0.595	**0.414**	**0.424**	0.710	0.605	0.615	0.660	0.579
Emotion	**0.397**	**0.385**	0.670	0.680	0.654	0.683	0.653	0.360	0.371	0.660	0.620	0.622	0.622	0.620
Moral	0.271	0.236	0.780	0.660	0.622	0.680	0.609	0.396	**0.424**	**0.630**	**0.685**	**0.689**	**0.697**	**0.683**

Table 12: Zero-shot Generative LLMs results using the full story vs. story summary, over four types of gold similarity scores: empathy, event, emotion and moral. Classification gold labels are binned by score > 2.5.

Understanding Bottleneck of Fine-tuned LLMs. Figure 5 shows the confusion matrix predicted by fine-tuned Llama-3-8B on training set. The model learned nothing but statistical $P(Y)$ of training set, leading to random predictions conditioned on input story pairs. See more in Section 5.3.

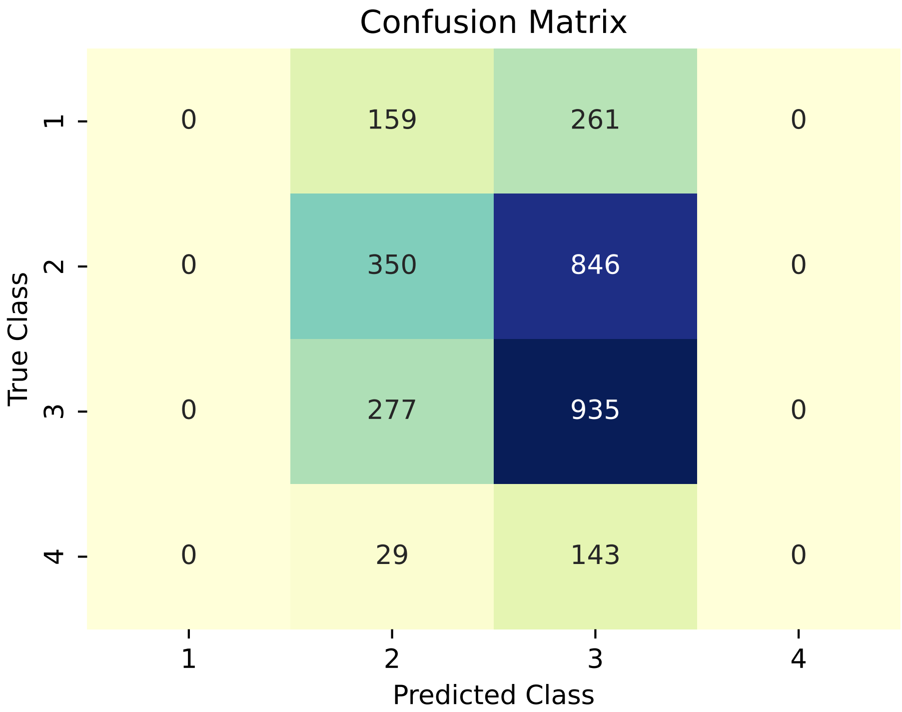

Figure 5: Confusion matrix of fine-tuned Llama-3-8B on training set. The model could not estimate the corresponding score conditioned on the input story pair, but sampled a similarity class based on *the gold class distribution of training data* $P(Y)$ = *(0.140, 0.399, 0.404, 0.057)* whatever the input pair was, leading to randomness on seen and unseen cases prediction. It learned nothing but statistical $P(Y)$ of training set. See more in Section 5.3.

C Annotation Guidelines

This figure shows the guidelines provided to annotators for obtaining annotations on the Roman Urdu dataset. Annotators were first familiarized with established tasks in the English language before being introduced to the Urdu dataset, which is their native language.

Dear Annotator,

Thank you for helping us with this task! Your job is to read pairs of stories and give them scores from 1 to 4 based on how similar they are. You will score each pair in three different ways: event, emotion, and empathy.

Here's what each type of similarity means:

Event: This is about what is happening in the stories. Are the characters going through similar situations? For example, if one story is about someone starting a new job and another is about someone moving to a new city, both are dealing with new environments.

Emotion: This is about how the characters feel. Are their feelings similar? For instance, if one story is about someone feeling lonely in a new place and another is about someone missing their friends, both characters might share feelings of loneliness.

Empathy: This is about how much you think the experiences of the characters resonate with each other. Do you think one character could understand or relate to what the other character is going through? For example, if one story is about someone who is struggling to make new friends and another is about someone feeling isolated at work, they might have a similar empathetic experience.

How to Rate the Stories:
1. **Read** both stories in the pair carefully.
2. **Think** about how similar the stories are in terms of the events happening, the emotions felt by the characters, and the empathy you feel between the two stories.
3. **Use** the button after each pair to select a score from 1 to 4 for each category:
   - 1 means "not similar at all"
   - 4 means "very similar"

Figure 6: Annotation guidelines provided to annotators.

Recruitment And Payment Four Urdu native speakers were employed to annotate Urdu dataset. We pay 1 AED per story pair for each annotator, in total of 300 * 4 = 1200 AED.

D Urdu Story Pair Generation

Figure 7 displays the prompt we prepared to synthetically generate the Roman Urdu story data. To maintain consistency between the English and Urdu datasets, we incorporated all the themes mentioned in (Shen et al., 2023). Figure 8 shows one of the 300 story pairs produced by GPT-4o.

Generate some short stories in Roman Urdu inspired by **Pakistani culture**, each around 200-250 words, infused with deep emotion and empathy. These stories should be created in pairs, labeled as Story_A, Story_B, with a similarity score ranging from **1 to 4** (where 1 means least similar and 4 means very similar).

The stories should span various **themes**, including but not limited to:
1. Romantic Relationships (keywords: relationships, divorced, passion)
2. Positive Life Events (keywords: opportunities, wedding, cruise)
3. Depression (keywords: depression, therapy, psych)
4. Family (keywords: families, parents, relatives)
5. Substance Use (keywords: recovery, drugs, addiction)
6. Encouragement (keywords: encouragement, caring, distress)
7. College and School (keywords: students, classes, college)
8. Loneliness (keywords: loneliness, relationships, haircut)
9. Youth (keywords: teenage, childhood, twenties)
10. Life Changes (keywords: goodbyes, retired, graduating)
11. Work (keywords: mundane, coworkers, volunteering)
12. Trauma (keywords: abused, traumas, therapist)

Generate four story pairs with varying similarity ranging from 1 to 4 to reflect different levels of thematic and emotional connection. Ensure that the stories feel authentic and relatable, avoiding vulgar language while maintaining a genuine and heartfelt tone.

OUTPUT four pairs as a list, with each story pair in json format as follows:
{
"**Story_A**": a story,
"**Story_B**": a story,
"**Similarity_score**": similarity between these two stories,
"**Theme**_story_**A**": theme or topic of story A,
"**Theme**_story_**B**": theme or topic of story B,
"**Reason**": briefly analyze why they have such a similarity,
}

Figure 7: Prompt used for Urdu story pair generation based on GPT-4o.

Roman Urdu story pair produced by GPT-4o.

{
"**Story_A**":"Aqsa ko nayi naukri mili thi Islamabad ke ek badi company mein. Yahaan usay opportunities mila karti thi jo us ne pehle kabhi socha bhi nahi tha. Ek din company ne usay Europe mein ek seminar attend karne ka bulaawa diya. Uska pehla international trip tha aur itni excitement ke baawajood woh apni maa se door jaane par thodi ghabra gayi thi. Lekin maa ne usay encouragement di aur kaha: 'Tujhe apne sapnon ke piche bhagna chahiye'. The trip was a huge success, aur Aqsa ko badi promotion mili. Ab woh aur uski maa khush hain aur har chutti pe ghar diya karti hai kahaniyan sunane ke liye.",

"**Story_B**":"Yasir ne apne career ke doran buhot mushkilein dekhi thi lekin usay kuch farak nahi padta tha kyunki uska passion usay aage barhne par majboor karta tha. Arsa guzar gaya aur aaj uska retirement ka din tha. Company ne ek badi farewell party arrange ki thi, aur uske har coworker ne apna pyar aur respect exams mein bayaan kiya. Har nayi, muskurahat le aayi unke chehre pe kyunki sab ne usay bata ya ke unka encouragement kitna zaroori tha. Usne waqt nikal kar apne coworkers ke saath gappe mari, unko shukriya kaha aur apni retirement ka safar khushi se shuru kiya.",

"**Similarity_score**":2,

"**Theme_story_A**":"Positive Life Events",

"**Theme_story_B**":"Life Changes",

"**Reason**":"Dono kahaniyon main positive life changes dikhayi gayi hain; ek mein pehli naukri ki excitement aur dosray mein retirement ka joy. Difference mainly transition phases aur personal milestone types mein hai."
}

Figure 8: Example of Urdu story pair.

E English Story Pair Example

The illustrative example form the EmpathicStories dataset (Shen et al., 2023) shows a sample pair of stories along with their summarized versions and scores in four aspects: empathy, event, emotion, and moral.

Example of the story pair with its summarized version and gold labels.

Empathy score: 4.0, Event score: 2.0, Emotion score: 2.5, Moral score: 2.5

Story_A: I have very few friends in the city I live in. I moved here a few years ago for work and I'm an introvert but somehow ended up making a few girl friends and I was really close to 2 of them. One moved to Canada last year and now the other is moving to another state to be with her boyfriend (soon to be husband). She was originally going to leave in December but some unavoidable stuff happened and she has to leave next week. My friend is social and popular so I don't think it affects her that much and I don't begrudge that. But I'm feeling very sad and heart broken. I'm pregnant so I think it's affecting me way worse. We had planned a bunch of activities in October (my baby shower, her bachelorette and her birthday) and those plans all need to modified now. She's going to miss the baby shower and I'm going to miss her bday. Attempting now to to throw a last minute bachelorette this weekend. I feel pretty down and blue.",

Story_B: I'm from Spain and my country went into lockdown around March 14th. My college lectures got canceled on the 11th since there were a couple of corona cases on campus. I haven't seen my best friend since March 6th or so. However, since our lockdown has been just eased, we're now able to go outside and ""meet"" others: as long as you're both wearing gloves and masks and keep a certain distance. It'll be weird because I won't be able to hug her, but I'm very happy that I get to see her again. I bought her favourite chocolates, and I can't wait to see her reaction.

Story_A_summary: The writer moved to a new city a few years ago for work and made a few close girl friends, but one moved to Canada last year and now the other is moving to another state to be with her boyfriend. The writer is feeling sad and heartbroken, especially because they had planned a bunch of activities in October, including a baby shower and a bachelorette party. The writer is attempting to throw a last-minute bachelorette party, but feels down and blue.

Story_B_summary: The author is from Spain and experienced a lockdown due to COVID-19. Their college lectures were canceled and they haven't seen their best friend since early March. The lockdown has been eased and they are now able to meet with gloves, masks, and distance. They bought their friend's favorite chocolates and are excited to see her again.",1,0.0," the two stories have completely different contexts and emotions, making it difficult for the narrators to empathize with each other.

Figure 9: Example of English story pair.