LLM-GEm: Large Language Model-Guided Prediction of People's Empathy Levels towards Newspaper Article

Md Rakibul Hasan[1] **Md Zakir Hossain**[1] **Tom Gedeon**[1] **Shafin Rahman**[2]
[1]Curtin University, Perth WA 6102, Australia
[2]North South University, Dhaka 1229, Bangladesh
{rakibul.hasan, zakir.hossain1, tom.gedeon}@curtin.edu.au
shafin.rahman@northsouth.edu

Abstract

Empathy – encompassing the understanding and supporting others' emotions and perspectives – strengthens various social interactions, including written communication in healthcare, education and journalism. Detecting empathy using AI models by relying on self-assessed ground truth through crowdsourcing is challenging due to the inherent noise in such annotations. To this end, we propose a novel system, named Large Language Model-Guided Empathy *(LLM-GEm)* prediction system. It rectifies annotation errors based on our defined annotation selection threshold and makes the annotations reliable for conventional empathy prediction models, e.g., BERT-based pre-trained language models (PLMs). Previously, demographic information was often integrated numerically into empathy detection models. In contrast, our *LLM-GEm* leverages GPT-3.5 LLM to convert numerical data into semantically meaningful textual sequences, enabling seamless integration into PLMs. We experiment with three `NewsEmpathy` datasets involving people's empathy levels towards newspaper articles and achieve state-of-the-art test performance using a RoBERTa-based PLM. Code and evaluations are publicly available at `https://github.com/hasan-rakibul/LLM-GEm`.

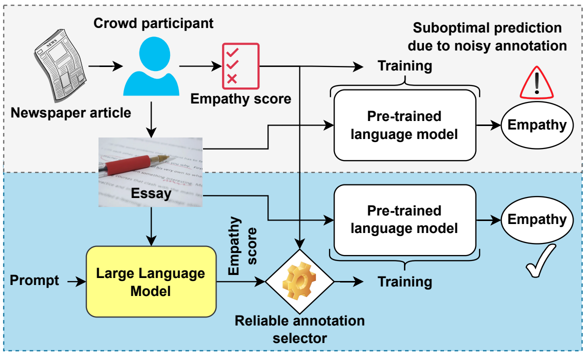

Figure 1: A typical empathy prediction workflow by directly utilising a PLM (Tafreshi et al., 2021; Barriere et al., 2022, 2023) versus our proposed LLM-guided workflow. Because of the noise in crowdsourced data, a typical workflow often results in suboptimal prediction. Our proposed workflow employs LLM to refine or re-define noisy annotations automatically and outperforms the typical approach.

1 Introduction

Empathy refers to an inherent ability to understand and convey suitable emotional responses in reaction to the emotions and viewpoints of others (Decety and Jackson, 2004; Olderbak et al., 2014). Seminal work by Batson et al. (1987) proposed the widely-recognised empathy measurement scale by defining empathy as having six aspects: sympathetic, moved, compassionate, tender, warm and softhearted. Empathic capability is key in cultivating interpersonal relationships and mitigating stress and discontent among individuals in our society in various human-to-human interactions.

Empathic doctors are better equipped to understand their patients' concerns, leading to improved communication and patient outcomes (Jani et al., 2012). This empathic connection is not confined to face-to-face interactions but extends to written communication, such as medical reports and informative articles that convey a compassionate understanding of patients' experiences. In education, especially with the shift towards online learning due to the COVID-19 pandemic, empathy is critical in helping teachers understand their students' emotional states and create a positive learning environment (Aldrup et al., 2022). In addition to verbal communication, empathy in the education sector also surfaces in written communications, such as emails and feedback on assignments, where the tone and language reflect a genuine concern for students' well-being. In examining the role of empathy in written journalism, consider the poignant example of a newspaper article detailing a local family's struggle after a devastating house fire. The journalist's empathic narrative goes beyond factual reporting, weaving a story that not only informs but also connects readers emotionally to the human experiences within the news.

Assessment of empathy levels is crucial in determining interaction quality (Bellet and Maloney, 1991). Empathy deficits often lead to conflicts and miscommunications, which can be resolved by measuring empathy levels as the first step, but such measurement is challenging, even for humans (Lawrence et al., 2004). Research endeavours in computational empathy remain limited (Alam et al., 2018) compared to other domains of affective computing, such as emotion (Kaklauskas et al., 2022), primarily due to the lack of high-quality data.

The aphorism, 'garbage in, garbage out', signifies how inaccurate data results in inaccurate outputs (Geiger et al., 2020). While crowdsourcing platforms (e.g., Amazon Mechanical Turk, CrowdFlower and Prolific) offer a simpler and faster way to get a sizeable participant pool, they suffer from false information (Sheehan, 2018). Such erroneous data result from carelessness and multitasking and threaten the validity of findings relying on such data (Jia et al., 2017; Huang et al., 2012). However, such crowdsourcing with self-assessment annotation is a major source of data collection in computational social science and human behaviour studies, such as empathy (Tafreshi et al., 2021) and emotion (Mohammad and Turney, 2010). Computational empathy using crowdsource annotation, therefore, often provides suboptimal performance (Figure 1).

In addition, the subjective nature of empathy necessitates consideration of people's demographic information, which is normally represented as numbers in the datasets. We, therefore, leverage demographic information into our prediction pipeline and introduce *LLM-GEm*, a Large Language Model (LLM)-guided empathy prediction system. While earlier studies, such as Wang et al. (2021), employed GPT-3 LLM for direct data annotations, to the best of our knowledge, no work has focused on using LLM to refine human annotations. To this end, we leveraged the enhanced capabilities of GPT-3.5 to reduce labelling errors in pre-existing crowdsourced annotations. It will be particularly useful when there is already some noisy crowdsource annotation. We experiment with three publicly available datasets to predict people's empathy levels toward newspaper articles, where our system results in competitive performance by outperforming prior work.

Our major contributions include **(1)** application of GPT-3.5 LLM to convert numerical demographic information to semantically meaningful text in order to seamlessly integrate them with a pre-trained language model (PLM), **(2)** employing GPT-3.5 LLM to reduce annotation errors caused by crowdsourcing, and **(3)** defining *annotation selection threshold* to systematically select between crowdsource annotation and LLM annotation.

2 Related Work

A Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA) has organised a series of competitions on predicting people's empathy towards newspaper articles. In these challenges from 2021 to 2023, several works (Vasava et al., 2022; Kulkarni et al., 2021; Srinivas et al., 2023; Lu et al., 2023) predicted empathy by fine-tuning RoBERTa PLM followed by some Multi-Layer Perceptron (MLP) layers. Apart from these, some studies (Ghosh et al., 2022; Butala et al., 2021; Hasan et al., 2023a) fine-tuned BERT PLM, and some other studies (Mundra et al., 2021; Lin et al., 2023; Chavan et al., 2023) leveraged an ensemble approach with fine-tuning multiple PLMs. Qian et al. (2022) experimented with multi-task learning and reported that a simple fine-tuning of the RoBERTa base model resulted in better performance (0.480 vs 0.508 Pearson correlation coefficient (r)). Fine-tuning PLMs, therefore, has become the conventional approach to predict people's empathy towards newspaper articles. Among different PLMs, RoBERTa has become the most frequently used prediction model in empathy detection studies, as reported in a recent survey covering computation empathy studies from 2013 to 2023 (Hasan et al., 2023b).

Several authors experimented with various approaches to ensure data quality in empathy predictions. As a data augmentation technique, Vasava et al. (2022) translated texts to a random language using Google Translate and then back to English. They combined five demographic features before the final layer of their empathy prediction pipeline. Qian et al. (2022) also harnessed demographic and personality data, which yielded a validation Pearson correlation coefficient (r) of 0.53. Notably, this performance surpassed that achieved without incorporating demographic and personality information. Data augmentation and demographic information, therefore, help to predict empathy levels.

As for data annotation employing LLM, Wang et al. (2021) experimented with GPT-3 in annotating data for several natural language processing tasks, including sentiment analysis, question generation and topic classification. They concluded that GPT-3 is a cost-effective way of annotating data but is not as reliable as human annotations. In this paper, we systematically select annotations between GPT-3.5 LLM and crowdsourcing to minimise existing noise in crowdsource annotations.

3 Method

3.1 Problem Formulation

Consider for the i^{th} data sample, $X = \{x_i^S, x_i^{D_{1,2,\cdots,m}}\}$, where x_i^S is a text sequence, $x_i^{D_{1,2,\cdots,m}}$ are m demographic data represented as real numbers. We aim to build a model $\mathcal{F}$ to predict the degree of empathy $Y^{\text{crowd}} = \{y_i^{\text{crowd}} \in [u, v]\}$, where y_i^{crowd} represents self-assessed continuous empathy score ranging from u to v, collected through crowdsourcing platforms such as Amazon Mechanical Turk. This self-assessed empathy score is referred to as *crowdsource annotation* throughout this paper.

We investigate two important aspects of this problem. **(1)** *Demographics information*: Prior work has experimented with different approaches in integrating numerical demographics information into text-based empathy prediction workflow. For example, Vasava et al. (2022) fused demographic information as numbers in an MLP layer after the PLM. Whereas Chen et al. (2022) used them as fixed sentences and reported a test performance drop from 0.537 to 0.295 Pearson r. On the other hand, Hasan et al. (2023a) used them as fixed sentences, but in a different style than Chen et al. (2022), and reported a validation performance increase from 0.565 to 0.865 Pearson r. Given that most text-based empathy prediction systems use PLMs in predicting empathy levels (Tafreshi et al., 2021; Barriere et al., 2022, 2023), it would be straightforward to integrate the demographic numerical information as text into the pipeline. Instead of sentences with a fixed pattern for all samples, naturally varying sentences may improve the performance. Further, the recent rise of LLMs necessitates making these converted texts meaningful so we can use this semantic information in prompt engineering with LLMs.

(2) *Annotation:* Prior work on empathy prediction suffers from suboptimal performance, especially with crowdsource self-annotation. In a series of empathy prediction challenges participated by several researchers for three years (Tafreshi et al., 2021; Barriere et al., 2022, 2023), a maximum Pearson correlation coefficient of only 0.558 is achieved. In contrast, another empathy prediction challenge in its debut (Barriere et al., 2023) got a 0.708 Pearson correlation. Apart from the actual text data to predict empathy, a major difference between these two challenges is the annotation protocol: self-annotation by all participants (0.558) versus controlled annotation of all samples by three external annotators (0.708). Given that crowdsource annotation is a faster and simpler way of getting data but suffers from false information (Sheehan, 2018), mitigating the annotation noise is clearly a key problem.

It is important to note that the practice of crowdsource annotation for sentiment analysis (Wang et al., 2021) or image analysis (Nowak and Rüger, 2010) differs substantially from annotations in computational social science. Computational social science involves collecting *raw data*, such as people's reactions to newspaper articles, with or without annotations. Consequently, even if the reliability of self-assessment annotations remains debatable, the underlying raw data can be salvaged by mitigating the noise inherent in the annotations.

3.2 Employing LLM in Empathy Prediction

We employ LLM in three scenarios: (1) processing demographic data, (2) annotation, and (3) data augmentation by rephrasing all essays and demographic sentences.

3.2.1 Numerical Demographics to Text Using LLM

The numerical demographic data $X^{D_{1,2,\cdots,m}}$ can be converted to semantically meaningful text using LLM to effectively integrate them into a text-based empathy prediction pipeline. There can be m demographic information such as gender, education level, ethnicity, etc. Demographic information for each sample i can be converted to sentences by first constructing a prompt and feeding it to an LLM:

$$P_i^D = f(x_i^{D_1}, x_i^{D_2}, \cdots, x_i^{D_m}) \quad (1)$$

$$x_i^D = \text{LLM}(P_i^D) \quad (2)$$

The actual text sequence where empathy would be predicted, and the demographic sentence for each data sample can then be concatenated as $x_i =$

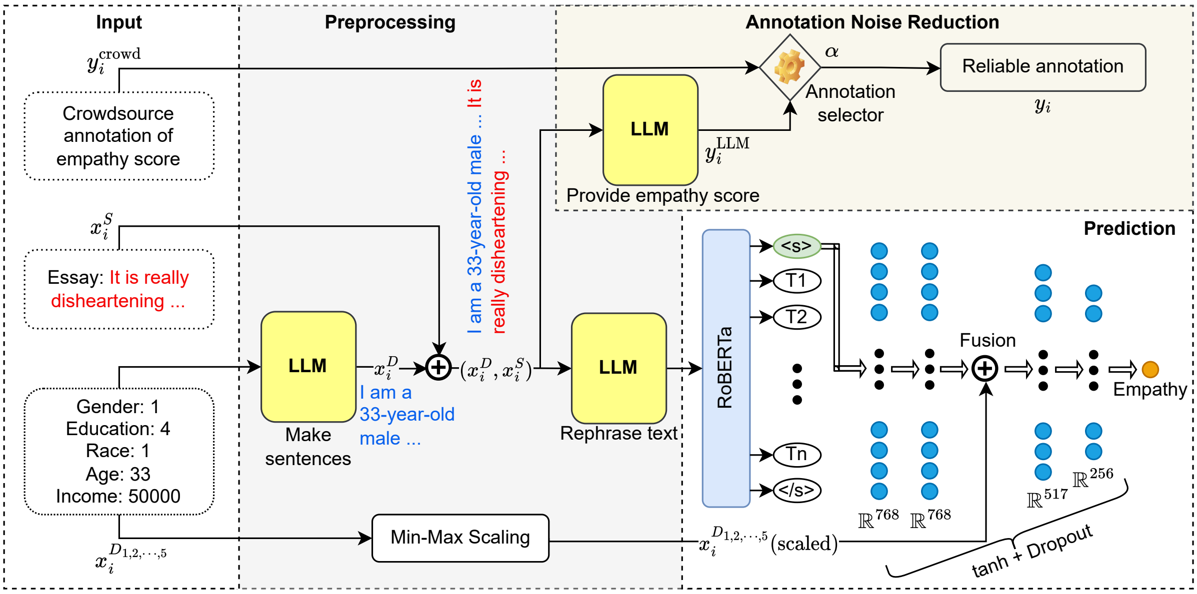

Figure 2: *LLM-GEm* system: we first use the LLM to convert demographic data to meaningful text. Essays and demographic sentences are used to annotate the essays using the LLM, and reliable annotations are then selected for each sample. After rephrasing the texts using the LLM, we train a RoBERTA-MLP model to predict empathy levels.

Essay	**Crowd.**	**LLM**
"After reading the article, you can't help but feel really sad and terrible for the people that were affected by the hurricane. It was a situation that they did not deserve and one that they most likely did not cause but mother nature has other plans for us. I feel bad for all the children as well as animals that are there as well with no shelter or food."	1.00	6.50
"Stories like this always manage to irritate me just a bit. I do not keep up with celebrity news so when some does manage to find it's way in front of me I'm just like "who cares"? I will never see these people in my real life, they will never have an impact on me and will never even cross my mind on their own."	1.33	1.20

Table 1: Two sample essays and their annotations using crowdsource participants and LLM in a continuous range from 1 to 7, where 1 and 7 refer to the lowest and highest empathy, respectively. Although the first essay is empathic, the self-annotation is the lowest, while the LLM annotation seems reasonable and correct. In the second example, both annotations seem correct. Empathic and non-empathic keywords are marked with blue and red colours, respectively.

(x_i^S, x_i^D), where the comma (,) symbol represents string concatenation.

3.2.2 Reducing Annotation Noise Using LLM

To reduce annotation noise, the best practice is to annotate the data with multiple annotators (Geiger et al., 2020). To this end, essay and demographic text sequences are fed together into an LLM to annotate each sample i. Some verified and reliable crowdsource annotations, along with their corresponding text sequences, are employed in a few-shot prompt engineering approach to enhance the consistency of the outputs generated by the LLM.

$$P_i^A = f([x_1, y_1^{\text{crowd}}], [x_2, y_2^{\text{crowd}}], \cdots, [x_n, y_n^{\text{crowd}}], x_i) \tag{3}$$

$$y_i^{\text{LLM}} = \text{LLM}(P_i^A) \tag{4}$$

where $[x_1, y_1^{\text{crowd}}], [x_2, y_2^{\text{crowd}}], \cdots, [x_n, y_n^{\text{crowd}}]$ are n verified and reliable crowdsource annotations and corresponding text sequences.

Two sample annotations, by both LLM and crowdsource, are presented in Table 1. Indeed, the annotation by LLM seems reasonable and accurate compared to the crowdsource annotation (Table 1). Even though the crowdsource annotations are noisy, we do not entirely discard the crowdsource annotations, particularly to predict crowdsource ground truth in the test set. In this regard, the annotation selection threshold guides toward more reliable annotations.

Figure 3 illustrates a histogram of differences between LLM and crowdsource annotations. In most cases, there are 0 to 0.5 differences between

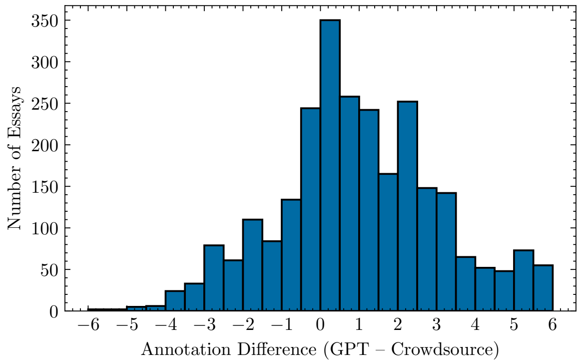

Figure 3: Annotation differences (GPT vs crowdsource).

the annotators. There are, however, cases where LLM and crowdsource annotations differ by larger margins. Refer to Appendix C.2 for more evidence.

An argument could be raised about the necessity of building another AI model guided by LLM, given that LLM, also an AI model, provides reliable empathy scores. LLMs as the prediction model may not be appropriate in some cases. Firstly, dependence solely on LLMs as the prediction model leads to high operating costs and computational demands. Secondly, LLMs may not be appropriate for edge devices such as smartphones and embedded systems. Comparatively smaller models are, therefore, often preferred, which can be optimised to get reasonably good performance compared to LLMs (Wang et al., 2023a). In this paper, we propose a computational empathy model that leverages y_i^{LLM} during training but can infer without needing LLMs.

3.3 LLM-GEm: LLM-Guided Empathy Prediction

Figure 2 depicts the details of the proposed *LLM-GEm* system. Between LLM annotation Y^{LLM} and crowdsource annotations Y^{crowd}, we select reliable annotations y_i for each sample based on annotation selection threshold α:

$$y_i = \begin{cases} y_i^{\text{LLM}} & \text{if } \Delta > \alpha \\ y_i^{\text{crowd}} & \text{otherwise} \end{cases} \tag{5}$$

where $\Delta = |y_i^{\text{crowd}} - y_i^{\text{LLM}}|$, i.e, the absolute difference between two annotations. As an example, if $y_i^{\text{crowd}} = 1$ and $y_i^{\text{LLM}} = 6.5$, the Δ becomes 5.5; therefore, the selected annotation y_i will be 6.5 and 1 for $0 \leq \alpha < 5.5$ and $5.5 \leq \alpha \leq 6.0$, respectively. The thresholds $\alpha = 0$ and $\alpha = \Delta$ mean using all LLM and crowdsource annotation, respectively. In the case of $\alpha \neq \{0, \Delta\}$, the selected annotation pool, concerning the whole training data, will result from both LLM and crowdsource. We, therefore, refer to this case as *mixed* annotation. The threshold α ranges from 0 (both have the same annotations) to the maximum possible annotation difference:

$$\alpha = \begin{cases} 0 & \text{all LLM annotations} \\ > 0 \;\&\; < \max(\Delta) & \text{mixed annotations} \\ \max(\Delta) & \text{all crowdsource annotations} \end{cases} \tag{6}$$

A higher Δ means a higher probability of annotation anomaly in crowdsource annotation. We train the prediction model using the text sequences, demographic information and the ground truth selected through the annotation selection threshold, and we test our system on the crowdsource annotation. The hidden representation corresponding to the first token (<s>) from the last layer of the RoBERTa PLM is extracted and fed into an MLP.

Empathy is subjective and, in fact, heavily dependent on people's demographic information, as proved by earlier studies on computational empathy (Guda et al., 2021; Vasava et al., 2022; Hasan et al., 2023a) and psychology (Borracci et al., 2017). We further leverage numerical demographic data in addition to the textual demographic information. Since the demographic values are in different ranges, we use min-max scaling before fusing the information into the MLP. More details of the architecture are presented in Appendix A.

4 Experiments

4.1 Experimental Setup

4.1.1 Dataset Setup

To evaluate people's empathy towards newspaper articles, we experiment with three datasets, consisting of written essays in English, demographic data and ground truth empathy score, Y^{crowd}. We manually verify that the demographic data are anonymised with no personal identifying information, such as full name or username. The ground truth is annotated by crowdsource participants based on Batson's empathy scale involving six aspects of empathy (Batson et al., 1987). The `NewsEmpathy v2` *training* dataset consists of whole `NewsEmpathy v1` data samples, while the `v2` *validation* and *test* sets consist of new samples. The `v3` dataset (Omitaomu et al., 2022; Barriere et al., 2023), on the other hand, has no overlapping samples with its earlier version (v1 and v2). Details of the datasets are presented in Appendix B.

The task in these datasets is to predict continuous empathy level $Y^{\text{crowd}} \in [1.0, 7.0]$ from input texts $X = \{\text{Essay, Demographic}\}$. The essays (X^{Essay}) are text sequences, while the demographic data (X^{D}) are represented as real numbers. As reported by Omitaomu et al. (2022), $X^{\text{D}_{\text{Gender}}} \in \{1, 2, 5\}$ corresponds to male, female and others; $X^{\text{D}_{\text{Education}}} \in \{1, 2, 3, 4, 5, 6, 7\}$ corresponds to different levels of educations; $X^{\text{D}_{\text{Race}}} \in \{1, 2, 3, 4, 5, 6\}$ corresponds to different races; $X^{\text{D}_{\text{Age}}} \in \mathbb{R}$ corresponds to age in years; and $X^{\text{D}_{\text{Income}}} \in \mathbb{R}$ corresponds to income in USD.

Similar to Barriere et al. (2023), we combine v2 and v3 training datasets and make a single training set, which has 5,268 samples after data augmentation. The model trained on this training set is used for evaluation in v2 and v3 datasets. Evaluation in v1 dataset, however, does not incorporate any external data (no v2, v3 or data augmentation) to maintain consistency with prior work (Buechel et al., 2018). The v1 dataset has 1,670 samples for 10-fold cross-validation.

4.1.2 LLM Setup

To interact with LLM through prompt engineering, we design appropriate prompts based on OpenAI best practices for prompt engineering (Fulford and Ng, 2023). We controlled the degree of randomness of the LLM output by using the *temperature* parameter of OpenAI API. The prompts were mostly sensitive to the presentation of responses, such as responding as '6' or 'six' with additional unnecessary sentences, rather than the contents of the response, such as empathy score. We iteratively tested prompts to get responses in the desired format. For numeric demographic data to text conversion, the prompt includes the mapping between numbers and actual information with a typical example sentence. During annotation, we provide three essays and their empathy scores as examples so that the LLM is likely to output the empathy score in a consistent style. Prompts with sample input and output with numerical to textual conversion, annotations, and rephrasing text are presented in Appendix C.1, Appendix C.2, and Appendix C.3, respectively.

4.1.3 Evaluation

We follow the established evaluation protocols by earlier studies on all three datasets. The v1 dataset comes with no separate validation and test set, and the evaluation protocol reported in Buechel et al. (2018) is 10-fold cross-validation. The v2 and v3 datasets have separate validation and test sets, and prior work (Tafreshi et al., 2021; Barriere et al., 2022, 2023) reported performance on hold-out test sets. The ground truths corresponding to the test sets in the v2 and v3 datasets are not publicly available. Instead, evaluations on test sets are obtainable through the CodaLab (Pavao et al., 2022) challenge websites: v2 dataset at WASSA 2022[1] and v3 dataset at WASSA 2023[2] challenges.

Earlier studies with NewsEmpathy datasets (Hasan et al., 2023a; Mundra et al., 2021) and general fine-tuning of PLMs (Dodge et al., 2020) reported that the initialisation of model parameters and the data orders in training heavily influence the model performance. Thus, we use different initialisation and data ordering in v2 and v3 evaluations through five different seed values (0, 42, 100, 999, 1234). We use Pearson correlation coefficient (r) as the evaluation metric, the official metric of WASSA 2021, 2022, and 2023 challenges using NewsEmpathy datasets.

4.1.4 Implementation Details

We utilise gpt-3.5-turbo-0613[3] version of GPT-3.5 LLM for demographic sentences, rephrasing and annotations. Our manual inspection of the annotations supports the correctness of LLM annotations. To check LLM's consistency in annotation, we annotated 21 samples twice at two different API calls. The annotations are fairly consistent, with a mean variation of 0.3 and a standard deviation of 0.42. On average, the LLM annotation costs us USD 0.94 per 1,000 essays. Of the 5,268 essay samples, GPT-3.5 declined to annotate two samples due to their lack of coherent thoughts or feelings, as they appeared to be a mix of unrelated sentences. Such erroneous samples are indeed challenging to screen out because these samples are textual content in a text dataset; however, GPT-3.5 detects them even without any explicit instructions.

We train and validate the RoBERTa-MLP model, having 125.7M total trainable parameters, utilising Python 3.11 on a single NVIDIA Tesla V100 32GB GPU. The primary software packages in-

[1] https://codalab.lisn.upsaclay.fr/competitions/834

[2] https://codalab.lisn.upsaclay.fr/competitions/11167

[3] GPT 3.5 (version: gpt-3.5-turbo-0613) was the latest version at the time of this research.

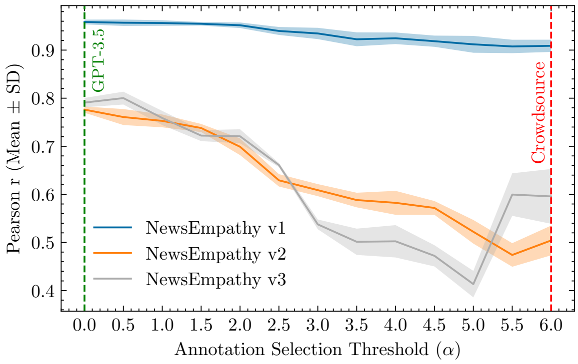

Figure 4: Validation set performance at different annotation selection thresholds, ranging from 0 (all GPT-3.5 LLM annotation) to 6 (all crowdsource annotation).

Data	Annotation	Validation (r) (Mean $\pm$ SD)
v1	Crowdsource	0.909 ± 0.013
	LLM-GEm	$\mathbf{0.958 \pm 0.005}$
v2	Crowdsource	0.504 ± 0.031
	LLM-GEm	$\mathbf{0.776 \pm 0.006}$
v3	Crowdsource	0.596 ± 0.057
	LLM-GEm	$\mathbf{0.791 \pm 0.010}$

Table 2: Validation results with 10-fold cross-validation (`NewsEmpathy v1`) and with five different initialisation and data order (`NewsEmpathy v2` and `NewsEmpathy v3`). The *LLM-GEm* performance is reported at $\alpha = 0$.

clude Transformers 4.31.0, Datasets 2.13.0, Pytorch 2.0.1, CUDA 11.7, scikit-learn 1.2.2 and Pandas 2.0.2. We use off-the-shelf `roberta-base` PLM from Hugging Face (Wolf et al., 2020), which is released under MIT license. To combat overfitting and mitigate catastrophic forgetting, we impose early stopping. Specifically, we stop the training if the validation loss does not significantly decrease (a minimum decrease of 0.01 is considered significant) for three epochs. We train the model for a maximum of 10 epochs with a learning rate of $1e{-}5$ in AdamW (Loshchilov and Hutter, 2019) optimiser, a linear learning rate decay scheduler with 6% warmup steps, a batch size of 16 and weight decay of 0.1. We use a fixed seed value of 0 to ensure reproducibility. To get the text embedding from the RoBERTa PLM, we experimented with concatenating the last four hidden states, which did not provide any benefit compared to using only the last hidden state. As the loss function, we experimented with mean-squared-error, Huber loss, and mean-absolute-error and found mean-squared-error more suitable.

4.1.5 Validation Strategy

Seminal work by Liu et al. (2019) introduced RoBERTa and strategies to fine-tune the RoBERTa PLM on downstream tasks. We adhere to the same hyperparameter settings reported in Liu et al. (2019) across our experiments to ensure the performance improvements are solely based on improvement in data quality rather than the choice of common hyperparameters, such as learning rate and batch size.

Annotation selection threshold (α) is the primary hyperparameter we introduce for minimising annotation noise. The annotation difference ranges from zero (both are the same) to six (one annotation is lowest, i.e., one, and the other annotation is highest, i.e., seven). A value of $\alpha = 0$ and $\alpha = 6$ denote selecting the entire annotations of LLM and crowdsource, respectively. A value of α between 0 and 6 means mixed annotations. In addition to tuning the annotations of train data, we experiment with varying the validation annotation. An elevation in the validation score signifies a corresponding enhancement in the quality of the underlying data, given that all other parts of the workflow remain constant. As seen on Figure 4, the validation performance on all three datasets has a clear pattern as the threshold varies. Data quality is, therefore, improved when we use LLM annotations and gradually degraded as we select more crowdsource annotations. The performance on the `NewsEmpathy v1` dataset appears relatively modest compared to the other datasets. This discrepancy could potentially be attributed to a smaller number of samples: 1,670 in `v1` dataset as opposed to 5,268 in `v2` and `v3` training sets.

Table 2 reports the validation scores in three datasets with annotations by crowdsource and *LLM-GEm*. Importantly, *LLM-GEm* annotations improve the performance of the validation sets by a large margin in all datasets. The performance is best at the `v1` dataset, with a Pearson r of 0.958.

We also experimented with how newspaper article text contributes towards empathy prediction and how our improved data works on a model reported and implemented by others. These results are presented in Appendix D and Appendix E, respectively.

4.2 Benchmarking Results

We compare our system's performance on similar empathy prediction studies on all three datasets (Ta-

Method	Best Model	Test (r)
`NewsEmpathy v1`[a]		
Buechel et al. (2018)	fastText-CNN	0.404
Ours (**LLM-GEm**)	RoBERTa-MLP	**0.924**
`NewsEmpathy v2`		
Vasava et al. (2022)	RoBERTa-MLP	0.470
Ghosh et al. (2022)	BERT-MLP	0.479
Qian et al. (2022)	RoBERTa	0.504
Lahnala et al. (2022)	RoBERTa	0.524
Chen et al. (2022)	RoBERTa	0.537
Plaza-del Arco et al. (2022)	RoBERTa	0.541
Butala et al. (2021)	BERT-MLP	0.358
Mundra et al. (2021)	ELECTRA + RoBERTa	**0.558**
Vettigli and Sorgente (2021)	LR	0.516
Kulkarni et al. (2021)	RoBERTa-MLP	0.517
Ours (**LLM-GEm**)	RoBERTa-MLP	0.505
`NewsEmpathy v3`		
Barriere et al. (2023)	RoBERTa	0.536
Wang et al. (2023b)	RoBERTa	0.331
Hasan et al. (2023a)	BERT	0.187
Srinivas et al. (2023)	RoBERTa-MLP	0.270
Lin et al. (2023)	{RoBERTa, EmoBERTa}-MLP	0.415
Gruschka et al. (2023)	RoBERTa	0.348
Chavan et al. (2023)	RoBERTa-SVM	0.358
Lu et al. (2023)	RoBERTa-MLP	0.329
Ours (**LLM-GEm**)	RoBERTa-MLP	**0.563**

[a] 10-fold cross-validation evaluation as per the prior work on `v1` dataset (Buechel et al., 2018)

Table 3: Comparison with similar empathy prediction works on all three datasets. Note that the test sets' ground truths come from crowdsourcing.

ble 3). Our proposed system, *LLM-GEm*, provides state-of-the-art (SOTA) test results on the `v1` and `v3` datasets. On the `v2` dataset, the performance is 0.053 behind the best result. The major reason behind such suboptimal performance can be the annotation noise in the test set. Given that the test set comes from the same distribution as the training set and we demonstrate how noisy the training set annotation is, it is highly likely that the test set has similar annotation errors. Although prior work (Mundra et al., 2021; Plaza-del Arco et al., 2022; Chen et al., 2022) reported better performance than ours with the same test set, a significant distinction here is the training labels. We train our model with noise-reduced labels, which makes the distribution of training and test labels significantly different. Another reason we anticipate is hyperparameter optimisation. Prior work on `NewsEmpathy` datasets reported significant changes in performance with hyperparameter optimisation (Hasan et al., 2023a; Mundra et al., 2021). As discussed earlier, we adhered to the same hyperparameter settings reported in the original RoBERTa paper to ensure the performance improvements are solely based on improvement in data quality. Therefore, SOTA performance on the `v2` dataset might be achievable through hyperparameter optimisation.

Several observations are explored from Table 3. **(1)** Earlier SOTA result (Mundra et al., 2021) on `v2` dataset and the second best result (Lin et al., 2023) on `v3` dataset leveraged multiple PLMs in ensemble fashion. On the contrary, *LLM-GEm* uses a simple pipeline with a single PLM, followed by some MLP layers and outperforms bulky ensembles.

(2) To use, not to use, or how to use demographic information remains a confounding factor in the literature. For example, Chen et al. (2022) reported decreased performance by using them as fixed sentences, while Hasan et al. (2023a) and Vasava et al. (2022) reported increased performance by using them as fixed sentences and as numbers, respectively. Gruschka et al. (2023), on the other hand, used one-hot encoding, unnecessarily increasing the dimensionality. Our system utilises demographic information both as meaningful varying sentences and as numbers, and the system outperforms earlier work.

(3) There is a decreasing trend of the overall performance of prior work from `v2` to `v3` dataset, which may be attributed to smaller dataset size (2,655 essays in `v2` versus 1,100 essays in `v3`). Our system provides SOTA results and outperforms all studies by a large margin in `v3` dataset.

(4) On the `v1` dataset, our work achieves the best improvement of 0.52 Pearson r as compared to the other two datasets. This notable improvement can be attributed to the reliable annotation and use of demographic sentences – provided by *LLM-GEm* system – utilised on a PLM-based pipeline.

(5) RoBERTa PLM is the most popular in the literature, and several work utilised its fine-tuned versions by emotion-related data (e.g., EmoBERTa and RoBERTa-Twitter (Lin et al., 2023)). We use the RoBERTa base model and achieve SOTA performance.

4.3 Ablation Study

4.3.1 Varying Input

Table 4 presents the ablation experiment in two broad categories: (1) discarding LLM annotations and (2) discarding crowdsource annotations. In each category, we vary training data and features.

Annotation	Training Data	Features	Val. (r)	Test (r)
Without LLM	v2 + v3	E	0.565	0.433
		D_t, E	0.577	0.446
		D_t </s> E	0.560	0.436
		D_t, D_n, E	0.626	0.451
Without Crowd.	v3	D_t, D_n, E	0.656	0.421
	v2 + v3	D_t, D_n, E	0.765	0.468
	v2 + v3 + Augm.	D_t, D_n, E	**0.792**	**0.498**

D_t – Demographic (text), D_n – Demographic (number)
E – Essay, Augm. – Augmentation

Table 4: Ablation study on the most recent v3 dataset by discarding either LLM or crowdsource annotations, varieties in training data samples and features. In the case of features without demographic numbers, no MLP layers are used as they are not required. Experiments are run on the same hyperparameters with a fixed seed value of 0, ensuring the same initialisation and data orders. Note that test set annotations always remain unchanged as crowdsource annotations.

Discarding crowdsource annotation, i.e., including LLM annotation, still improves both validation (0.626 to 0.765) and test (0.451 to 0.468) performance, with the training data and input features remaining unchanged. Verified without LLM annotation, demographic information improves empathy prediction, with an improvement of 0.013 Pearson r in the test set. This aligns with earlier studies by Hasan et al. (2023a) and Vasava et al. (2022). Using demographic information both as text (with essays) and as number (intermediate fusion) in a single experiment further improves the performance by 0.049 Pearson r in the validation set. We also experiment with inputting the demographic sentences and essays with a separator token (</s>), which slightly lowers the performance compared to simply concatenating. Verified with discarding crowdsource annotations, i.e., including LLM annotations, adding v2 training data and data augmentation improves the performance by 0.109 and 0.027 validation Pearson r, respectively.

4.3.2 Varying Annotation Selection Threshold

Table 5 presents test performances on v2 and v3 datasets with varying annotation selection threshold α from zero (all LLM annotations) to six (all crowdsource annotations). On both datasets, the best Pearson r is achieved in a combination of LLM and crowdsource annotations selected using α of 5.5 and 4.5, respectively.

α	NewsEmpathy v2 (r)	NewsEmpathy v3 (r)
0.0 (all LLM)	0.459	0.498
0.5	0.434	0.424
1.0	0.429	0.479
1.5	0.438	0.462
2.0	0.452	0.448
2.5	0.442	0.495
3.0	0.447	0.516
3.5	0.490	0.458
4.0	0.468	0.536
4.5	0.496	**0.563**
5.0	0.495	0.554
5.5	**0.505**	0.495
6.0 (all crowd)	0.458	0.481

Table 5: Test performance on v2 and v3 datasets with different annotation selection thresholds α (defined in Equation (5)) at a fixed seed value of 0.

5 Conclusion and Future Work

Empathy plays a crucial role in social dynamics, such as education, health and business. Evaluating people's empathy levels using computational tools such as AI requires good-quality data. Computational social science often involves collecting data and annotation from crowdsourcing, which often has noise. To this end, our system, *LLM-GEm*, aims to minimise annotation noise and ensure data quality. We experiment with three datasets predicting people's empathy levels towards newspaper articles. We define an annotation selection threshold to systematically select between LLM and crowdsource annotations, which achieves SOTA performance.

Our annotation error mitigation method can be applicable to other self-annotation datasets with necessary adaptations in the prompts (to include/change the details of the problem, range of annotation labels, etc.). For example, Abdul-Mageed et al. (2017) collected self-annotation to detect empathy in social media, where similar error analysis and possible inclusion of LLM may help mitigate annotation noise, if any. Similarly, Hossain and Rahman (2022) used crowdsourcing self-annotated data to detect customers' empathy behaviour, where our LLM-based annotation noise removal can be helpful. Apart from these, it could be applicable to other similar self-annotated datasets across different computational social science and human behaviour studies. Future work can further investigate better loss functions that closely estimate the Pearson r evaluation metric. Finally, experimenting with PLMs that are pre-trained on emotion and empathy-related datasets would be another avenue we leave for future work.

Limitations

The primary limitation is the manual tuning of the annotation selection threshold (α). A more principled approach to determining the optimal threshold represents an interesting avenue for further exploration. Second, our LLM-GEm system is slightly behind SOTA in the `NewsEmpathy v2` dataset. As discussed in Section 4.2, the major reasons we anticipate are annotation noise in the test set and hyperparameter optimisation. Although few prior works reported better performance than ours with the same test set, a significant distinction here is the training labels. We train our model with noise-reduced labels, which makes the distribution of training and test labels significantly different. With such a distribution shift, model performance degrades, which may require other evaluation approaches (Chen et al., 2021). Even so, our model performance is competitive in the NewsEmpathy v2 dataset and beats the SOTA in the `v1` and `v3` datasets.

Another limitation is the reliance on the `NewsEmpathy v1`, `v2` and `v3` datasets, all of which are based on people reading news articles. Evaluating LLM-GEm on more diverse dataset types would strengthen the generalisability of the results. Finally, we could not train or fine-tune LLM (e.g., GPT-3.5) as the primary empathy prediction model. It would be interesting to examine how such a larger language model performs compared to a smaller language model (e.g., RoBERTa). LLM would likely outperform RoBERTa, but training or fine-tuning LLM may be a suboptimal choice at some scenarios due to increased hardware and overall cost requirements.

Ethics Statement

Empathy is subjective, and people's empathy levels depend on demographic factors such as age, gender and ethnicity. This line of research, therefore, should be carefully designed so that the prediction model does not generate biased output by depending more on demographics rather than actual content. Our use of LLM in generating meaningful texts from demographic numbers may not be biased because the LLM here merely constructs sentences according to the pre-defined mapping. Furthermore, rephrasing texts using LLM may not have a significant bias because it is not open-ended text generation (Dhamala et al., 2021). However, LLM outputs may be biased with empathy scores, capturing gender, race or socioeconomic stereotypes, which warrants future experimentation. With the deployment of our proposed empathy detection system, the privacy of people's personal and demographic information can be at risk and, therefore, should be addressed as per appropriate ethical guidelines and protocols that come with the datasets.

Acknowledgements

This research was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government.

References

Muhammad Abdul-Mageed, Anneke Buffone, Hao Peng, Johannes Eichstaedt, and Lyle Ungar. 2017. Recognizing pathogenic empathy in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 448–451.

Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2018. Annotating and modeling empathy in spoken conversations. *Computer Speech & Language*, 50:40–61.

Karen Aldrup, Bastian Carstensen, and Uta Klusmann. 2022. Is empathy the key to effective teaching? a systematic review of its association with teacher-student interactions and student outcomes. *Educational Psychology Review*, 34(3):1177–1216.

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.

C Daniel Batson, Jim Fultz, and Patricia A Schoenrade. 1987. Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of personality*, 55(1):19–39.

Paul S. Bellet and Michael J. Maloney. 1991. The importance of empathy as an interviewing skill in medicine. *JAMA*, 266(13):1831–1832.

Raúl A Borracci, Hernán C Doval, Leonardo Celano, Alejandro Ciancio, Diego Manente, and José GE Calderón. 2017. Patients' perceptions of Argentine physicians' empathy based on the Jefferson scale of patient's perceptions of physician empathy: Psychometric data and demographic differences. *Education for Health*, 30(1):19–25.

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.

Yash Butala, Kanishk Singh, Adarsh Kumar, and Shrey Shrivastava. 2021. Team Phoenix at WASSA 2021: Emotion analysis on news stories with pre-trained language models. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 274–280, Online. Association for Computational Linguistics.

Tanmay Chavan, Kshitij Deshpande, and Sheetal Sonawane. 2023. PICT-CLRL at WASSA 2023 empathy, emotion and personality shared task: Empathy and distress detection using ensembles of transformer models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 564–568, Toronto, Canada. Association for Computational Linguistics.

Mayee Chen, Karan Goel, Nimit S Sohoni, Fait Poms, Kayvon Fatahalian, and Christopher Re. 2021. Mandoline: Model evaluation under distribution shift. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1617–1629. PMLR.

Yue Chen, Yingnan Ju, and Sandra Kübler. 2022. IUCL at WASSA 2022 shared task: A text-only approach to empathy and emotion detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*.

Jean Decety and Philip L Jackson. 2004. The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*, 3(2):71–100.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Isa Fulford and Andrew Ng. 2023. ChatGPT Prompt Engineering for Developers. Short Course by DeepLearning.AI & OpenAI. Accessed 1 June 2023.

R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 325–336.

Soumitra Ghosh, Dhirendra Maurya, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Team IITP-AINLPML at WASSA 2022: Empathy detection, emotion classification and personality detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 255–260.

Fabio Gruschka, Allison Lahnala, Charles Welch, and Lucie Flek. 2023. CAISA at WASSA 2023 shared task: Domain transfer for empathy, distress, and personality prediction. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 553–557, Toronto, Canada. Association for Computational Linguistics.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. EmpathBERT: A BERT-based framework for demographic-aware empathy prediction. *arXiv preprint arXiv:2102.00272*.

Md Rakibul Hasan, Md Zakir Hossain, Tom Gedeon, Susannah Soon, and Shafin Rahman. 2023a. Curtin OCAI at WASSA 2023 empathy, emotion and personality shared task: Demographic-aware prediction using multiple transformers. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 536–541, Toronto, Canada. Association for Computational Linguistics.

Md Rakibul Hasan, Md Zakir Hossain, Shreya Ghosh, Susannah Soon, and Tom Gedeon. 2023b. Empathy detection using machine learning on text, audiovisual, audio or physiological signals. *arXiv preprint arXiv:2311.00721*.

Md Shamim Hossain and Mst Farjana Rahman. 2022. Detection of potential customers' empathy behavior towards customers' reviews. *Journal of retailing and consumer services*, 65:102881.

Jason L Huang, Paul G Curran, Jessica Keeney, Elizabeth M Poposki, and Richard P DeShon. 2012. Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27:99–114.

Bhautesh Dinesh Jani, David N Blane, and Stewart W Mercer. 2012. The role of empathy in therapy and the physician-patient relationship. *Complementary Medicine Research*, 19(5):252–257.

Ronnie Jia, Zachary R Steelman, and Blaize Horner Reich. 2017. Using mechanical turk data in is research: risks, rewards, and recommendations. *Communications of the Association for Information Systems*, 41(1):14.

Arturas Kaklauskas, Ajith Abraham, Ieva Ubarte, Romualdas Kliukas, Vaida Luksaite, Arune Binkyte-Veliene, Ingrida Vetloviene, and Loreta Kaklauskiene. 2022. A review of ai cloud and edge sensors, methods, and applications for the recognition of emotional, affective and physiological states. *Sensors*, 22(20).

Atharva Kulkarni, Sunanda Somwase, Shivam Rajput, and Manisha Marathe. 2021. PVG at WASSA 2021: A multi-input, multi-task, transformer-based architecture for empathy and distress prediction. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 105–111, Online. Association for Computational Linguistics.

Allison Lahnala, Charles Welch, and Lucie Flek. 2022. CAISA at WASSA 2022: Adapter-tuning for empathy prediction. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 280–285.

Emma J Lawrence, Philip Shaw, Dawn Baker, Simon Baron-Cohen, and Anthony S David. 2004. Measuring empathy: reliability and validity of the empathy quotient. *Psychological medicine*, 34(5):911–920.

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.

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 preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34.

Jay Mundra, Rohan Gupta, and Sagnik Mukherjee. 2021. WASSA@IITK at WASSA 2021: Multi-task learning and transformer finetuning for emotion classification and empathy prediction. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 112–116, Online. Association for Computational Linguistics.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*, MIR '10, page 557–566, New York, NY, USA. Association for Computing Machinery.

Sally Olderbak, Claudia Sassenrath, Johannes Keller, and Oliver Wilhelm. 2014. An emotion-differentiated perspective on empathy with the emotion specific empathy questionnaire. *Frontiers in Psychology*, 5.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations. *arXiv preprint arXiv:2205.12698*.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2022. *Codalab competitions: An open source platform to organize scientific challenges*. Ph.D. thesis, Université Paris-Saclay, FRA.

Flor Miriam Plaza-del Arco, Jaime Collado-Montañez, L. Alfonso Ureña, and María-Teresa Martín-Valdivia. 2022. Empathy and distress prediction using transformer multi-output regression and emotion analysis with an ensemble of supervised and zero-shot learning models. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 239–244, Dublin, Ireland. Association for Computational Linguistics.

Shenbin Qian, Constantin Orašan, Diptesh Kanojia, Hadeel Saadany, and Félix Do Carmo. 2022. SURREY-CTS-NLP at WASSA2022: An experiment of discourse and sentiment analysis for the prediction of empathy, distress and emotion. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 271–275.

Kim Bartel Sheehan. 2018. Crowdsourcing research: Data collection with amazon's mechanical turk. *Communication Monographs*, 85(1):140–156.

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.

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.

Himil Vasava, Pramegh Uikey, Gaurav Wasnik, and Raksha Sharma. 2022. Transformer-based architecture for empathy prediction and emotion classification. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 261–264.

Giuseppe Vettigli and Antonio Sorgente. 2021. EmpNa at WASSA 2021: A lightweight model for the prediction of empathy, distress and emotions from reactions to news stories. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 264–268, Online. Association for Computational Linguistics.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yiding Wang, Kai Chen, Haisheng Tan, and Kun Guo. 2023a. Tabi: An efficient multi-level inference system for large language models. In *Proceedings of the Eighteenth European Conference on Computer Systems*, pages 233–248.

Yukun Wang, Jin Wang, and Xuejie Zhang. 2023b. YNU-HPCC at WASSA-2023 shared task 1: Large-scale language model with LoRA fine-tuning for empathy detection and emotion classification. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 526–530, Toronto, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A Architecture Details

The *LLM-GEm* system (presented in Figure 2) comprises LLM (GPT-3.5) to preprocess data and provide annotation. In preprocessing, we concatenate the essay text sequences with the demographic sentences converted by LLM. Using the selected annotation through annotation selection threshold (α), we train a RoBERTa-MLP model. The MLP portion consists of four hidden layers, having `tanh` activation function, followed by a dropout of 0.2 during training. The first hidden layer has a hidden size of 768×768. The second hidden layer has a hidden size of 768×512. Next, we add the five numerical demographic information; therefore, the next hidden layer's input size becomes 517. The last layer's size is 256×1, which provides an empathy score between 1.0 to 7.0. The number of hidden layers, their sizes, activation functions and dropouts are decided through experiments at a fixed seed value of 0.

B Dataset Details

Table 6 provides the statistics of the datasets. We name these datasets as `NewsEmpathy` because they involve people's empathic reactions towards newspaper articles. Buechel et al. (2018) released the first reported dataset of this kind, consisting of 1,860 essays in response to articles involving harm to individuals, organisations or nature. In this `NewsEmpathy v1` dataset, 403 participants read five random newspaper articles from a pool of 418 articles and wrote essays reflecting on each news article they read. The raw article varies in length from 101 to 32,058 characters, with an average number of characters of 4,316.

The `v1` dataset is further extended by Tafreshi et al. (2021), which includes an additional 161 participants. The extended version (named `v2`), with 2,655 essays in total, was utilised in WASSA (Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis) Shared-Task 2022 (Barriere et al., 2022) and 2021 (Tafreshi et al., 2021). The `NewsEmpathy v3` dataset (1,100 essays) – employed in the WASSA 2023 challenge – utilises 100 selected newspaper articles from the total 418 articles and comprises new essay data.

The `v1` dataset is available under the CC BY 4.0 license, and the other two datasets (`v2`, `v3`) are available for scientific or research purposes.

Dataset	Train	Validation	Test	Total
v1 (Buechel et al., 2018)				1,860
v2 (Tafreshi et al., 2021)	1,860	270	525	2,655
v3 (Omitaomu et al., 2022)	792	208	100	1,100

Table 6: Datasets with the corresponding number of essays in train, validation and test sets. The `NewsEmpathy` `v1` dataset comes with no train-validation-test splits, and the standard evaluation protocol is 10-fold cross-validation.

C LLM Prompt and Sample Response

C.1 Numerical Demographics to Text Using LLM

The following prompt template is used for each participant by providing their demographic information.

C.1.1 LLM Prompt

Your task is to format five numerical data (individual's gender, education level, race, age, and income) into meaningful sentences.
The numerical data are delimited by triple backticks.
Write from a first-person point-of-view.
Complete the task with no more than three sentences.

Use the following mapping between the number and the corresponding text:
Gender:
1 = Male
2 = Female
5 = Other

Education level:
1 = Less than a high school diploma
2 = High school diploma
3 = Technical/Vocational school
4 = Some college but no degree
5 = Two-year associate degree
6 = Four-year bachelor's degree
7 = Postgraduate or professional degree

Race:
1 = White
2 = Hispanic or Latino
3 = Black or African American
4 = Native American or American Indian
5 = Asian/Pacific Islander
6 = Other

Age:
<number> = <number> years

Income:
<number> = <number> USD

For example, if the input numbers are: "Gender: 1, Education level: 5, Race: 1, Age: 25, Income: 40000"
The output can be "I am a 25-year-old male of the White race. I completed a two-year associate degree and earn 40000 USD."

Input numbers: ```Gender: `{gender}`, Education level: `{education level}`, Race: `{race}`, Age: `{age}`, Income: `{income}````

C.1.2 Sample Response

Table 7 illustrates four sample input numerical demographic information and corresponding output (text sequence).

C.2 Reducing Annotation Noise Using LLM

To annotate, we provide three sample input-output responses as few-shot prompting. In addition to guiding the LLM in annotation, it further facilitates the responses in a consistent style. The following prompt template is used for each essay.

C.2.1 LLM Prompt

'role': 'system', 'content': 'You are an AI model that annotates written essays to provide an empathy score between 1.0 to 7.0 based on the definition of empathy.\n The essays were written by human participants after reading a newspaper article involving harm to individuals, groups of people, nature, etc. The essay is provided to you within triple backticks. Your response must contain one and only empathy score.',

'role': 'user', 'content': 'Essay: ```I am a 37-year-old male of the African American race. I have a four-year bachelor's degree and earn 40000 USD. It breaks my heart to see people living in those conditions. I hope that all the aid that was sent to the island makes it to the people who need it the most. I do not know what I would do it that was my family and I. I would hope that I would do my best, but I can see how depressing and hopeless you could feel having your whole life changed because of a storm and not knowing where your next meal is coming from.```',

Dataset	Input					LLM **Output**
	`Gender`	`Edu. level`	`Race`	`Age`	`Income`	
v2	1	4	1	33	50000	I am a 33-year-old male of the White race. I have completed some college but have not obtained a degree. Currently, I earn 50000 USD.
	1	6	5	43	50000	I am a 43-year-old male of Asian/Pacific Islander descent. I obtained a four-year bachelor's degree and my income is 50000 USD.
v3	2	6	2	38	24000	I am a 38-year-old female of Hispanic or Latino race. I have a four-year bachelor's degree and earn an income of 24000 USD.
	2	6	1	46	22000	I am a 46-year-old female of the White race. I have a four-year bachelor's degree and earn 22000 USD.

Table 7: Samples of numerical demographic information and corresponding output from LLM.

Dataset	**Input** `essay`	Crowd.	LLM **Output**
v2	"I am a 36-year-old male of Asian/Pacific Islander descent. I have a four-year bachelor's degree and my yearly income is 50000 USD. This doesn't sound too worrisome to me. If you are affected by depression, then you should decide as an individual to find another option. It's just a side effect, I think there are many medicines out there that have depression as a side effect. It is something that individuals should be aware of, but nothing critical."	1	1.2
	"I am a 28-year-old male of the White race. I completed a two-year associate degree and earn 62000 USD. Some people are just disturbed. Everyone is going to point to mental illness in something like this, but I fear that when things like this happens, it gets easier and easier just to point at people and call them mentally ill. I fear that eventually we are going to get to the point where you can classify someone as mentally ill by the crowd. Just finger pointing. I feel for these victims and the victims of crimes in the future because we spend too much time making excuses and pointing fingers rather than dealing with the problems at hand."	1	4.5
v3	"I am a 33-year-old male of other gender identity. I have completed some college but no degree and earn 36000 USD. It is terrible that these kids have to starve but it was kind of funny that they were boiling grass I mean think about these africans they always have some sort of crisis even when they are rich in oil they still manage to ruin everything they are like monkeys they resemble apes they behave like animals"	4.33	1
	"I am a 29-year-old male of the White race. I completed a four-year bachelor's degree and my yearly income is 85000 USD. After reading the article, my reaction is that it is very sad that boys that young have to be put behind bars. I think that children should be able to experience their childhood and have fun at that age. They should not be facing hardships at all. They should be playing with friends and be in school at that age and not locked up behind a cell."	1.5	6.5

Table 8: Samples of essay and its annotations using crowdsourcing and LLM in a continuous range between 1 to 7, where 1 and 7 refer to the lowest and highest empathy, respectively. Empathic and non-empathic keywords are marked with blue and red colours, respectively.

'role': 'assistant', 'content': '6.8',

'role': 'user', 'content': 'Essay: ```I am a 38-year-old female of the White race. I possess a postgraduate or professional degree and earn 42000 USD. It sounds like these boys had a really rough life. I do think we all have personal responsibility for our choices at the end of the day though. Even though you might have it rough, ultimately it is up to you to decide to break the law or use drugs or not. So they had free will too and could have exercised that. Regardless, it is still sad that they went through a rough childhood. Nobody should have to endure that and kids are the saddest victims.```',

'role': 'assistant', 'content': '2.0',

'role': 'user', 'content': 'Essay: ```I am a 29-year-old male of the White race. I have completed some college but do not have a degree. I earn 39,000 USD. I just read an article about how animals in Venezuela are starving to death because of the turmoil that is happening in that country. I feel like there should be some kind of action plan in place for animals in countries that are experiencing hardship so that they can be

Dataset	**Input**		LLM **Output**
	Essay	Emotion	
v2	I am a 33-year-old male of the White race. I have completed some college but have not obtained a degree. Currently, I earn 50000 USD. it is really diheartening to read about these immigrants from this article who drowned. it makes me feel anxious and upset how the whole ordeal happened. it is a terrible occurrence that this had to happen at the mediterranean sea. thankfully there were some survivors. the fact that babies were lost makes it that much more emotional to read all of this	Sadness	I am a 33-year-old White male. I have some college education but no degree. Currently, I earn 50000 USD. It is truly disheartening to read about the immigrants mentioned in this article who drowned. It makes me feel anxious and upset about the entire situation. It is a tragic event that had to occur in the Mediterranean Sea. Fortunately, there were some survivors. The loss of babies in this incident adds to the emotional impact of reading all of this. I feel really sad.
v3	I am a 37-year-old male of the African American race. I have a four-year bachelor's degree and earn 40000 USD. It breaks my heart to see people living in those conditions. I hope that all the aid that was sent to the island makes it to the people who need it the most. I do not know what I would do it that was my family and I. I would hope that I would do my best, but I can see how depressing and hopeless you could feel having your whole life changed because of a storm and not knowing where your next meal is coming from.	Hope / Sadness	I am a 37-year-old African American male. I hold a bachelor's degree and have an annual income of 40000 USD. Witnessing individuals residing in such circumstances deeply saddens me. My fervent wish is that the assistance sent to the island reaches those who require it the most. If I were in their shoes, I cannot fathom the despair and desolation I would feel as my entire life is altered by a storm, uncertain about the source of my next sustenance.

Table 9: Rephrased essays corresponding to input essay text and self-assessed emotion category by participants.

Annotation	**Training Data**	**Model**	**Features**	**Validation** (r)	**Test** (r)
Crowd.	v2 + v3	RoBERTa	Demog (text) + essay </s> article	0.577	0.442
LLM-GEm	v2 + v3 + Augmentation	RoBERTa-MLP	Demog (text, number) + essay </s> article	0.796	0.488
		RoBERTa-similarity	Demog (text, number) + essay </s> article	0.73	0.445

Table 10: Effect of article inclusion with training data samples of `NewsEmpathy v2` and `NewsEmpathy v3` or with their augmentations, evaluated on `NewsEmpathy v3` dataset. All experiments were run on the same hyperparameters with a fixed seed value of 0, ensuring the same initialisation and data orders.

transported to other places in times of crisis. The thought of innocent creatures starving to death in cages really turns my stomach.```',

'role': 'assistant', 'content': '5.7'

'role': 'user', 'content': 'Essay: ```{essay}```

C.2.2 Sample Response

Table 8 reports some sample essays and their annotation by LLM. The self-assessed annotations from crowdsourcing are also presented to compare the annotation between LLM and crowdsourcing.

C.3 Rephrasing Essay for Data Augmentation

We rephrase all essays using LLM prompt engineering as a data augmentation technique. The following prompt template is used for each essay.

C.3.1 LLM Prompt

In a data collection experiment for empathy detection, the study participant writes essay to describe their feeling after reading a newspaper article involving harm to individuals, groups or other entities.

The participant's demographic information are also available within the essay.

As a data augmentation tool for NLP, your task is to paraphrase the demographic and essay information delimited by triple backticks.

Do not add any additional information not contained in the input texts.

Overall, the participant expressed {emotion} emotion. Do not change this overall emotion of the participant's essay.

Your response must not have any backticks or any additional symbols.

Input demographic and essay: ```{essay}```

C.3.2 Sample Response

Table 9 presents some samples of original essays written by participants and corresponding rephrased versions by LLM.

D Inclusion of Newspaper Article Texts

To accommodate long article sequences in a PLM-based pipeline, we summarise these articles using LLM. The `gpt-3.5-turbo-0613` model version could not summarise six articles because this GPT-3.5 model version was limited with a maximum context length of 4,097 tokens. We, therefore, use 16k context length supporting `gpt-3.5-turbo-16k` version to summarise those six articles. The resulting summarised articles vary from 107 to 2,063 characters, with an average length of 776 characters, although we instruct GPT-3.5 to use at most 1,000 characters. We also rephrase the articles as a data augmentation technique.

D.1 LLM Prompt to Summarise Articles

Your task is to summarize given text delimited by triple backticks.
Use at most 1000 characters.
Do not add any additional information not contained in the input text.

Input text: ```{article text}```

D.2 LLM Prompt to Rephrase Articles for Augmentation

As a data augmentation tool for NLP, your task is to paraphrase the newspaper article delimited by triple backticks.
Do not add any additional information not contained in the input texts.
Your response must not have any backticks or any additional symbols.
Input newspaper article: ```{article}```

D.3 Results with Article Texts

To accommodate newspaper articles, we experiment in two different ways: (1) we combine articles and essays (with demographic sentences) with a separator token (</s>) and input them into the empathy prediction pipeline, and (2) we process articles and essays separately on two encoders, calculate their cosine similarity, and input the encoded sequence as well as the similarity score into the prediction pipeline. The idea behind calculating similarity is that for an essay to be empathic, it ideally should have similarities with the articles, with a proportional relationship.

As seen on Table 10, the article texts do not have a meaningful contribution to the overall performance in both crowdsource and *LLM-GEm* annotations. The inclusion of cosine similarity does not benefit either.

α	**Improved data (r)**	**Original data (r)**
0.0 (all LLM)	0.746	-
0.5	0.718	-
1.0	0.726	-
1.5	0.721	-
2.0	0.718	-
2.5	0.695	-
3.0	0.656	-
3.5	0.544	-
4.0	0.496	-
4.5	0.472	-
5.0	0.445	-
5.5	0.392	-
6.0 (all crowd)	0.448	0.458

Table 11: Validation set Pearson r of the model reported by Vasava et al. (2022) on our improved `NewsEmpathy v2` datasets and the original `v2` dataset (performance on original data is taken from Vasava et al. (2022)). The performance on our data is reported on different annotation selection thresholds α (defined in Equation (5)) at a fixed seed value of 0.

E Further Validation of Data Improvement

We use our improved `NewsEmpathy v2` dataset on the model reported by Vasava et al. (2022) to validate our contribution to data improvement further. We chose this specific work because their implementation and hyperparameter are publicly available[4]. Table 11 compares the validation set Pearson r using our improved data versus the original data reported in Vasava et al. (2022). As can be seen, our improved data resulted in a significant boost in performance on most annotation selection thresholds, which proves the enhanced quality of the data.

[4] https://github.com/notprameghuikey0913/WASSA-2022-Empathy-detection-and-Emotion-Classification